ECAI 2024
U. Endriss et al. (Eds.)
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA240578

Group Fairness with Individual and Censorship Constraints

Zichong Wang[a] **and Wenbin Zhang**[a,*]

[a]Florida International University, Miami, FL, 33199

Abstract. The widespread use of Artificial Intelligence (AI) based decision-making systems has raised a lot of concerns regarding potential discrimination, particularly in domains with high societal impact. Most existing fairness research focused on tackling bias relies heavily on the presence of class labels, an assumption that often mismatches real-world scenarios, which ignores the ubiquity of censored data. Further, existing works regard group fairness and individual fairness as two disparate goals, overlooking their inherent interconnection, *i.e.,* addressing one can degrade the other. This paper proposes a novel unified method that aims to mitigate group unfairness under censorship while curbing the amplification of individual unfairness when enforcing group fairness constraints. Specifically, our introduced ranking algorithm optimizes individual fairness within the bounds of group fairness, uniquely accounting for censored information. Evaluation across four benchmark tasks confirms the effectiveness of our method in quantifying and mitigating both fairness dimensions in the face of censored data.

1 Introduction

AI-based decision-making systems are becoming increasingly prevalent across various sectors of society [11]. However, the implementation of these systems in real-world scenarios has, at times, resulted in bias and discrimination against marginalized groups or populations. This trend has sparked growing concerns about the potential ethical and fairness issues arising from AI-driven automated decision-making. These concerns are particularly pronounced in high-stakes scenarios such as loan approve [50], criminal justice [10], and healthcare [9], where the consequences of unfair decisions can have far-reaching impacts and serious ethical implications. For example, it would become a serious ethically problematic if a bank's loan decision were influenced by the race information of the applicant or their close contacts.

To address the aforementioned problem, literature has explored methods to improve the fairness of AI algorithms by mitigating biases from training data [48, 49] or training algorithms with fairness-aware frameworks [44, 47, 53]. Existing fairness works are typically categorized into two main areas: group fairness and individual fairness [15]. The principle behind group fairness is to prevent a socially salient group from being collectively assigned more favorable outcomes than another group. Existing work [7, 27, 31] achieves group fairness through statistical fairness by first identifying protected group (*e.g.,* females) and unprotected group (*e.g.,* males) groups by identifying sensitive attributes, which is potential sources of bias, and then ensuring that the predictor yields similar outcome statistics (*e.g.,* prediction accuracy and true-positive rates) across different subgroups. In contrast, the principle behind individual fairness is to prevent two similar individuals from receiving differential treatment. Unlike group fairness, which requires pre-defined sensitive attributes, individual fairness [32, 28, 17] aims to evaluate fairness at a more granular level by requiring similar individuals to receive comparable probability distributions on class labels, thereby preventing unfair treatment.

A major obstacle to the practical application of existing work on fairness is the assumption that the precondition that class labels are fully usable holds, which fails in the presence of uncertainty in class labels due to censoring [55, 56, 25]. Therefore, existing fairness notions cannot be directly applied to censorship settings. For example, as illustrated in Figure 1 (a), for censored individuals A_3 and A_7, the true loan status is unknown because A_3 dropped out of the application process and for A_7, the time of loan decision-making exceeded the study period among other reasons, leading to uncertainty in the class labels, *i.e.,* load application status. Due to the inability to handle censorship information, existing fairness works quantify and mitigate bias by focusing on the proportion of data with certain class labels. Consequently, these studies either drop observations with uncertain class labels [10, 16, 52] or omit censoring information [36, 35, 53]. However, removing this information would bias the results even towards individuals with known class labels [33, 34, 59, 57].

In addition, existing fairness works often treat individual fairness and group fairness as distinct tasks, failing to consider potential implications among them [17, 56, 58, 55]. However, this separation of objectives could introduce additional bias into each other. Typically, an outcome ranking that satisfies statistical constraints on group fairness is considered to achieve group fairness, implicitly suggesting that individual fairness does not affect group fairness. However, this dichotomy can sacrifice individual fairness, especially for individuals on the decision boundaries of unprotected groups. For example, as illustrated in Figure 1 (b), we have two types of results: the Performance-Driven Result at the top and the Group Fairness-Driven Result at the bottom. In the Group Fairness-Driven Result, where the sensitive attribute is race, it ensures that the loan approval rates for different racial groups are consistent with achieving group fairness. Additionally, the existing Group Fairness-Driven Result aims to achieve group fairness while minimizing performance loss. Thus, although applicants A_3 and A_4 are less risky than A_5 and A_6, they are more risky than A_1 and A_2 of the same sensitivity group, and thus A_3 and A_4 are denied loans to maintain racial balance in loan approvals and minimize performance loss. This causes individuals A_3 and A_4, who are at the margins of decision-making, to always bear the loss of group fairness, thus introducing individual bias (*i.e.,*

* Corresponding author. Email: wenbin.zhang@fiu.edu

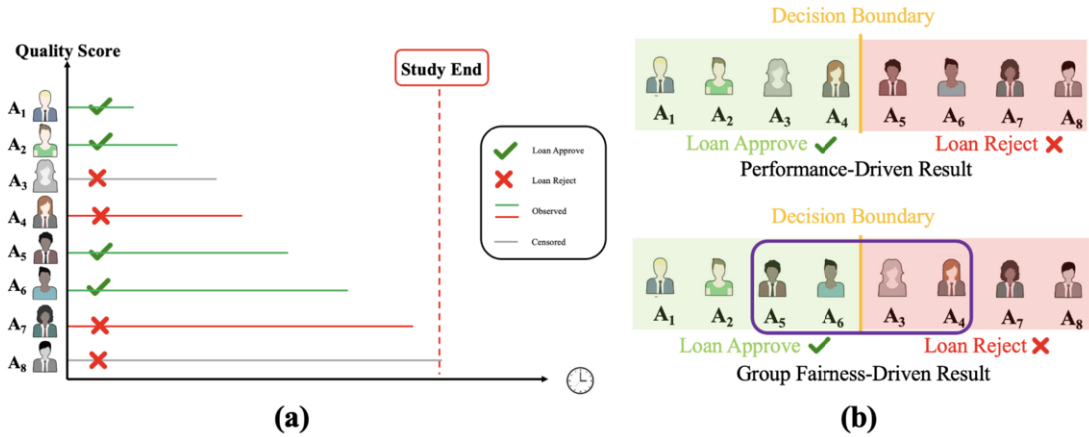

Figure 1: An illustration of the challenge of achieving fairness when censorship occurs. Applicants A_3 and A_4 have a lower risk than applicants A_5 and A_6; Applicants A_3 and A_4 are not approved for a loan to maintain a balanced loan approval ratio across racial groups. Moreover, applicants A_3 and A_7 are censored, while others are non-censored.

Purple Rectangle).

Therefore, there is an urgent need to address fairness in the presence of censorship while simultaneously balancing the impact of group and individual fairness, which remains largely unexplored and presents unique challenges: **i) Quantifying and Mitigating Bias in Censored Settings:** Most existing concepts in fairness studies rely on the availability of class labels. However, these assumptions become inapplicable in a censored setting. Thus, measuring and mitigating algorithmic bias in cases where class labels are unavailable poses a significant challenge. **ii) Balancing Group and Individual Fairness:** Existing fairness works often treat group fairness and individual fairness as distinct objectives. However, achieving group fairness may inadvertently lead to differential treatment of instances within unprotected groups, especially those near the decision boundary. This introduces additional biases, thereby potentially undermining individual fairness. **iii) Model Agnosticism:** Most fairness approaches are tailored to specific models. This specialization limits their broad applicability, often hindering their ability to enhance the fairness of different models.

In response to these challenges, this paper presents a preliminary study on *censored group fairness under individual and censorship constraints*, aiming to achieve fairness guarantees that align better with realistic scenarios. Corresponding debiasing algorithms are developed. *To the best of our knowledge, this is the first work to reconcile both group and individual fairness in the context of censorship settings.* Specifically, our strategy integrates the Rawlsian difference principle [30], ensuring optimized well-being for individuals and subgroups in the resultant outcome order. To counter the individual-level unfairness potentially stemming from enforcing group fairness [17], we employ a probabilistic distribution of possible valid rankings in line with distributive justice theory. This ensures uniform individual fairness loss, culminating in consistent treatment at the individual level. Further, by using group and individual fairness metrics attuned to the censorship setting, we embed vital censorship information into our fairness evaluations, making our methodology fitting for the censorship setting. The paper's principal contributions are as follows:

- We propose a new fairness notion-Censorship Group individual aware fairness, which considers the fairness of the outcome in terms of both group and individual fairness.
- We propose a novel framework GIFC designed to achieve both group and individual fairness. Specifically, we measure the fairness loss of the sorted distribution based on both group fairness loss and individual fairness loss. Also, we demonstrate the advantage of probabilistic ordering over optimal deterministic ordering.
- Extensive experiments on four benchmark datasets show that we propose GIFC acquires superior performance on both group fairness and individual fairness and achieves comparable prediction performance in downstream tasks.

2 Related Work and Background

2.1 Censored Data

In numerous real-world scenarios, the primary outcome, or the class label, can become inaccessible in the presence of censorship. Censored data is prevalent in various fields, such as clinical prediction (Support) [24], marketing analytics (KKBox) [25], and recidivism prediction instrument datasets (COMPAS [2] and ROSSI [18]). Censored data is typically characterized by three pieces of information [38]: i) the observed covariates/features x, which provide certain information that is observed for each individual, ii) The survival time, T, represents the elapsed time from when an individual entered the study until their last follow-up, indicating when the event of interest was either observed or censored, and iii) the event indicator δ, denoting whether or not the event has been observed. If $\delta = 1$, it signifies that the event is observed, confirming the event time T or class label, and vice versa. Uncertainty primarily stems from the last two pieces of information. In contrast to supervised settings, AI fairness in censored environments presents unique difficulties. Censorship not only restricts the existing fairness concepts but also amplifies the uncertainty and complexity within decision-making regions. This paper centers on the concept of group fairness with individuals within the context of censorship, a dimension not yet explored by existing fairness studies that have primarily concentrated on either group or individual fairness.

2.2 AI Fairness

Extensive research [12, 39, 40, 41, 42, 46] has been conducted to quantify and mitigate the bias of underlying learning algorithms. Existing fairness measures can be mainly divided into two main categories, *i.e.,* group fairness and individual fairness [27]. Specifically, group fairness aims to achieve statistical parity between different groups defined

by sensitive attributes [19]. On the other hand, individual fairness scrutinizes potential bias and discrimination at a much finer granularity, ensuring that similar individuals receive similar probability distributions over class labels, thereby mitigating unfair treatment [17]. Despite their great success, they typically presume the availability of class labels and are thus inapplicable in censorship settings where the class labels are uncertain [61]. In addition, most existing methods tend to focus on addressing either individual or group biases but rarely both [43]. This singular focus can overlook the complex interplay between the two types of fairness, which can be crucial for fully understanding and mitigating biases in AI systems [4]. While there are some studies [3, 26, 45] that attempt to bridge this gap by considering both individual and group fairness, they often strive for globally optimal solutions. These solutions, while theoretically appealing, may not effectively address the nuances of individual treatment in practical applications, potentially leading to oversights in fairness at the personal level.

To jointly address these challenges, our method proposes a holistic approach that aims for group fairness in censored settings while concurrently ensuring equitable treatment across individuals.

2.3 Survival Analysis

Censored data, also known as survival data, widely appears in numerous real-world scenarios and underscores the importance of survival analysis. Survival analysis aims to address the issues associated with accessing partial survival information from study cohorts [13]. Among the various survival analysis methods proposed, the Cox Proportional Hazards (CPH) model [14], a semiparametric model, has gained recognition as the most extensively used. It describes the multiplicative relation between risk, as expressed by the hazard function, and covariates. This model is primarily defined by two key concepts: the hazard function, which calculates the instantaneous rate of an event occurring at a specified time t, conditional on survival up to that point. Mathematically, it is expressed as follows:

$$h(t|x) = \lim_{\triangle t \to 0} \frac{\Pr\left(\mathrm{t} < \mathrm{T} < \mathrm{t} + \triangle \mathrm{t} | \mathrm{T} \geq \mathrm{t}, \mathrm{x}\right)}{\triangle t} \tag{1}$$

The CPH model specifies the hazard function as follows:

$$h(t|x) = h_0(t) \exp(\beta^T x) \tag{2}$$

where $h_0(t)$ denotes the baseline hazard, which represents the hazard value independent of features x, and β is a parameter vector. This can be estimated by applying the partial likelihood estimation, as Equation 3 shows:

$$L(\beta) = \prod_{T_i \text{ observable}} \frac{\exp(\beta^T x_i)}{\sum_{T_j \geq T_i} \exp(\beta^T x_j)} \tag{3}$$

The survival function, which is the probability that the event does not occur up to time t and can be determined from the hazard function, and vice versa. Mathematically, it is expressed as follows:

$$S(t|x) = \exp\left(-\int_0^t h(t|x)dt\right) \tag{4}$$

The CPH model assumes that an individual's risk of an event occurring is a linear combination of the individual's covariates, referred to as the linear proportional hazards condition. Developing the CPH model, in order to solve nonlinear problems, deep neural network structures have been extended to feature interaction models for survival data [22]. An alternative research approach involves the use of tree-based methodology [5, 21]. In this methodology, the splitting rule is adjusted to accommodate censored data, freeing it from the proportional assumption inherent in the CPH model.

To evaluate survival models, the *concordance index, or C-index*, is commonly used [37]. Specifically, the concordance index is evaluated by constructing every comparable pair of comparisons for a given individual. Critically, the consistency score filters out non-comparable pairs, *i.e.,* the shorter time is censored, and both censored pairs have identical survival time. For any two comparable samples x_i and x_j, concordance is realized under three circumstances: i) If individual x_i's survival time t_i is briefer than that of t_j for x_j, the model ought to designate a more elevated risk score to x_i. ii) Conversely, if t_i surpasses t_j, x_j should be assigned a diminished risk score. iii) When both individuals share equivalent survival times, and both are uncensored, they should bear matching risk scores. However, in instances where one individual is censored, the model should attribute a heightened risk score to the one not censored. Mathematically, this can be represented as:

$$\begin{aligned} C_{x_i} &= \frac{1}{\sum_{j \neq i} \mathbb{1}[\delta_< = 1]} \sum_{j \neq i} \mathbb{1}[h(t|x_>) < h(t|x_<), \delta_< = 1] \\ &= \frac{1}{\sum_{j \neq i} \mathbb{1}[\delta_< = 1]} \\ &\quad \times \sum_{j \neq i} \mathbb{1}[\exp(\beta^\top x_>) < \exp(\beta^\top x_<), \delta_< = 1] \end{aligned} \tag{5}$$

where $x_>$ and $x_<$ are the individuals with a longer, *i.e.* $t_> = \max(t_i, t_j)$, and a shorter, *i.e.* $t_< = \min(t_i, t_j)$, survival time, and $\delta_<$ is the event indicator of shorter survival time. C_{x_i} can be interpreted as the fraction of all other individuals whose predicted survival times are correctly ordered with x_i considering their actual survival times.

Like other AI methods, survival models can suffer from fairness issues and may make biased decisions about deprived subgroups. Beginning with [55], several lines of research have investigated fairness under censorship. Specifically, most fairness works [55, 29] under censorship only consider group fairness, while recent work [55, 51] has started to address individual fairness in censored settings. However, they often treat group fairness and individual fairness as separate goals and lack consideration of the interplay between them. Unlike these works, our approach explicitly considers the effects of both group fairness and individual fairness, and modifies the survival model to balance the additional individual fairness loss from group fairness constraints. In addition, we explicitly incorporate survival information to address discrimination in the presence of censorship.

3 Methodology

In this section, we first discuss how to quantify the amount of group fairness in a censorship setting. Following that, we delve into understanding the individual fairness losses that arise due to group fairness constraints in a censorship setting. Finally, we introduce our proposed methodology, GIFC.

3.1 Quantifying Bias Under Censorship

3.1.1 Censored Group Fairness

The commonly used notions of group fairness in AI systems typically require clear and definable class labels to measure fairness. However, in censorship settings, where data may be incomplete or labels

uncertain, these traditional metrics cannot be applied directly. This limitation has prompted the exploration of alternative metrics that can adapt to the constraints of censored data while accurately quantifying group bias. Inspired by prior work [54], we employ the Statistical Concordance Parity Difference (SCPD) to quantify the model's performance in different subgroups amidst censorship. Fundamentally, the SCPD aims to determine whether a model consistently underperforms for subgroups identified by sensitive attributes. It achieves this by gauging the discrepancies in the model's pairwise comparisons between predicted and actual outcomes for individuals both within and between these subgroups. The mathematical representation is as follows:

$$SCPD = \max\{\forall s_i, s_j \in S | CF(s_i) - CF(s_j), i \neq j\} \quad (6)$$

where $CF(\cdot)$ represents the concordance fraction (CF) evaluating the subgroup-wise correct pairwise ordering based on its respective group members, and the CF for a subgroup s_i is defined as:

$$CF(s_i) = \frac{\sum_{\forall x_i \in s_i} C_{x_i}}{CNum(s_i)} \quad (7)$$

where C_{x_i} denotes the concordance index of sample x_i and $CNum(s_i)$ denotes the total number of pairwise comparisons possible within the subgroup s_i.

Overall, a fair model should demonstrate consistent performance across all population groups. Any significant variation in model outcomes across different subgroups might suggest inherent bias. Therefore, a lower SCPD score implies a closer agreement between concordance within the subgroups, and thus, a fairer model. Notably, the SCPD approach goes beyond previous definitions of group fairness by explicitly incorporating elements of survival time and censorship information into its computations. This inclusion is vital as it ensures that crucial censorship data, which could significantly influence the model's performance assessment, is not overlooked. By integrating this data, the SCPD prevents the potential loss of critical information that could otherwise introduce substantial biases into the fairness evaluation.

3.1.2 Censored Individual Fairness amidst Group Fairness

Existing metrics for group fairness typically assess the treatment of subgroups from a global perspective, focusing on statistical parity among groups rather than the experiences of individuals within those groups. This approach, while effective for detecting and correcting systemic biases at the group level, may inadvertently overlook how individual experiences differ within these constraints. As a result, achieving group fairness might not necessarily translate into individual satisfaction, particularly for those who prioritize personal utility over collective well-being. To bridge this gap between group fairness and individual fairness, we introduce a novel metric, the *constraint concordance difference (CCD)*, which specifically measures how group fairness constraints affect individual fairness. The CCD evaluates the impact of these constraints by comparing the concordance index of each individual sample before and after the application of group fairness measures. This comparison involves pairwise evaluations of predictions, assessing whether the relative order of outcomes for individuals is preserved when group fairness constraints are applied. Mathematically, the CCD is defined as follows:

$$CCD = C'_{x_i} - C_{x_i} \quad (8)$$

where C_{x_i} is the concordance index of sample x_i prior to applying group fairness constraints, and C'_{x_i} is the concordance index of the same sample after the constraints have been enforced.

A key advantage of the CCD is its ability to identify individuals who are disproportionately affected by group fairness constraints. A lower CCD value suggests that the loss of individual fairness is minimal, indicating that the fairness measures are effectively balanced between the needs of the individual and the group. This is particularly important in scenarios where the personal impact of AI decisions is significant, such as in employment, lending, and healthcare settings. Furthermore, similar to the SCPD, the CCD incorporates critical elements, *i.e.,* survival time and censorship information into its computations. This inclusion is vital for handling censored data effectively, ensuring that the CCD can be applied in a wide range of contexts where data may be incomplete or partially observed. By integrating these elements, the CCD enhances the robustness of fairness assessments, making it a versatile tool for ensuring fairness in complex datasets.

3.2 GIFC framework

In this section, we present ranking algorithms specifically designed to guarantee both group and individual fairness in censored contexts. A significant hurdle we encounter is the optimality dilemma, wherein the individual with the most minimal loss in individual fairness, due to the enforcement of group fairness constraints, tends to disproportionately bear the majority of the group fairness loss. To address this challenge, our approach focuses on bolstering individual fairness. We achieve this by randomizing the probability distribution over possible rankings that satisfied group fairness. This strategy ensures that every individual, on average, bears a consistent level of group fairness loss.

The core rationale of our approach is to distribute the group fairness loss uniformly across individuals. To achieve this, we construct a probability distribution over outcomes that adhere to group fairness and then independently sample from this distribution to determine our final outcome. As a result, the sample's empirical characteristics converge to those of the fair distribution. A primary step in this methodology is to pinpoint the efficient outcome. To this end, we incorporate the Rawlsian difference principle [30], rooted in John Rawls' theory of distributive justice. This principle aims to promote equity by maximizing the welfare of the most disadvantaged groups. When in equilibrium, the Rawlsian difference principle ensures that all groups retain their status quo, as the welfare of the least advantaged group cannot be further enhanced, ensuring a balanced performance across groups. In our framework, welfare is synonymous with individual fairness in the model, quantified by its individual fairness loss. The mathematical representation of the Rawlsian difference principle is given by:

$$min_\theta \sum_{s=0}^{n} Var(U(r_i, \theta)) \quad (9)$$

where r_i is a valid group fairness ranking result and θ represents the individual fairness loss function, and U(r_i,θ) represents the utility function that assesses the loss across a outcome of samples, utilizing the model defined by parameters θ.

Building on the aforementioned principle, we formally characterize an effective ranking as an alteration in ranking that meets the group fairness constraint while simultaneously optimizing the util-

ity for the most vulnerable individual. Specifically, we take into account a set of outputs from the survival analysis model, denoted as $R_{\text{ori}} = \{x_1, x_2, \ldots, x_n\}$, ordered in descending sequence based on risk scores, from the highest to the lowest. For the sake of clarity, we presume that each individual possesses a distinct risk score. Moreover, we establish constraints grounded on the group fairness metric delineated in Section 3.1.1, represented as $SCPD \leq \omega$. Here, ω, lying in the interval [0, 1], acts as a modulating parameter dictating our group fairness tolerance threshold. We represent $R = \{r_1, r_2, \ldots, r_n\}$ as the entire set of rankings for R_{ori} that satisfy group fairness, and $S \subseteq R$ stands for the ensemble of all permissible rankings. The mathematical representation is presented below:

$$S = argmin\{\forall\, r_i \in R, SCPD \leq \omega | \sum_{i=0}^{n} CCD(x_i)\} \qquad (10)$$

For each valid ranking $s_i \in S$, we associate it with an individual $x_i \in X$, yielding a real-valued satisfaction measure $f(S, x_i)$. We define a randomized algorithm, denoted as $\mathcal{F}$, that for a given problem instance Q, deterministically selects a solution $\mathcal{F}(Q) \in S$. This algorithm $\mathcal{F}$ induces a probability distribution $\mathcal{D}$ over S such that $P(S) = P(\mathcal{F}(Q) = S)$. The expected satisfaction for each $x_i \in X$ under D is represented as $D[u] = \mathbb{E}_{S\sim D}[f(S, x_i)]$. A distribution $\overline{D}$ over S is deemed group-individual aware fairness for $(X, \mathcal{F})$ if it is infeasible to enhance the expected satisfaction of any individual without diminishing it for another individual in a comparable or worse-off position, *i.e.,* for all distributions D over S and all $x_i \in X$. Consider two distributions: a group-individually just distribution, denoted as $\overline{D}$, and another arbitrary distribution, denoted as D. At any given instance x_i, if the level of satisfaction under $\mathcal{D}$ exceeds that under $\overline{D}$, it implies that individual x_i achieves a superior level of satisfaction with the $\mathcal{D}$ distribution compared to the $\overline{D}$ distribution. However, the core principle behind a collectively-individual fair distribution is the equitable distribution of satisfaction across all individuals. Consequently, if x_i enjoys an elevated satisfaction level under $\mathcal{D}$, it becomes imperative, for the sake of overall fairness, that there exists another instance x_j wherein the satisfaction under $\mathcal{D}$ is diminished compared to $\overline{D}$. To sustain this fairness, the satisfaction of x_j under $\overline{D}$ should not surpass the satisfaction of x_i under $\overline{D}$. This arrangement is pivotal in ensuring that a mere alteration in distribution cannot indiscriminately elevate everyone's satisfaction. To amplify the satisfaction of one individual, another individual, with an equal or lesser satisfaction level, inevitably experiences a reduction.

A significant obstacle we encounter is the potentially exponential nature of viable solutions and distributions. Consequently, guaranteeing that outcomes are computed within exponential time becomes an intricate task. In light of this, we operate under the assumption that a weighted optimization oracle is present when the formulated distribution D does not epitomize utmost fairness. By augmenting the weights of individuals that D fails to satisfy, we can harness the weighted optimization oracle to derive a new, more equitable solution, denoted as $\overline{D}$. Subsequently, we enhance the weights corresponding to these individuals, effectively "pushing D towards optimal fairness." Consequently, the anticipated satisfaction of the collective-individual fairness distribution can be ascertained within polynomial time.

To derive the group-individual fair distribution, denoted as $\overline{D}$, we begin by initializing it as an empty set. We progressively prove all of its expected satisfactions, CCP_{x_i}, by continually expanding the subset $\overline{D}$. The heart of the proof involves expanding $\overline{D}$ incrementally until $\overline{D}$ encompasses the entirety of X. A pivotal tool in our approach is the separating oracle. Assisted by a weighted optimization oracle, we construct a separating oracle tailored for the dual problem. This particular oracle ensures the minimal loss of both group and individual fairness. Its function isn't merely to verify the feasibility of a proposed solution; when confronted with an infeasible solution, it returns the constraints that have been violated. To efficiently tackle the dual problem, we utilize the ellipsoid method, ensuring that a solution is ascertainable in polynomial time. The output from linear programming then guides the update for the set $\overline{D}$. We iterate through this procedure until the subset $\overline{D}$ spans the entirety of the set X. It's vital to note that our reliance on the ellipsoid algorithm necessitates only a finite number of calls to the Separation Oracle. As a result, we only need to account for the constraints returned by the Separation Oracle, enabling the crafting of a more compact linear programming challenge. Owing to the diminished constraint count, we can tackle this problem within polynomial time, culminating in the desired distribution. Thus, although the size of the effective solution set may be exponential, we can still obtain it in polynomial time.

Overall, GIFC ensures consistent expectations of individual fairness loss under group fairness constraints. This is achieved by introducing randomization to create probability distributions for group-individual fairness rankings. As a result, each individual can be assured that they are not the most disadvantaged when adhering to the group fairness requirement.

4 Experiment

4.1 Datasets

We evaluate our approach on four real-world datasets that include socially sensitive attributes coupled with censorship. Detailed characteristics of these datasets are provided in Table 1. The **ROSSI** dataset features data on individuals who were convicted and released from Maryland state prisons, and subsequently monitored for a year post-release [18]. The **COMPAS** dataset, known for its significance in algorithmic unfairness research, includes data utilized for predicting recidivism rates in Broward County [2]. The **KKBox** dataset, sourced from WSDM-KKBox's Churn Prediction Challenge 2017 [25]. Finally, the **Support** dataset incorporates data on patients admitted to five tertiary care academic medical centers, allowing for an exploration into healthcare scenarios [24]. Notably, these datasets contain explicit survival information, enabling us to specifically account for censoring during our analysis.

Table 1: Summary of the datasets used in the evaluations.

Dataset	ROSSI	COMPAS	KKBox	Support
Sample#	432	10,325	2,814,735	8,873
Feature#	9	14	18	14
Sensitive Attribute	Race	Race	Gender	Gender
Sensitive Value	African American	African American	Female	Female
Censored#	318	7,558	975,834	2,840
Censored Rate%	73.6%	73.2%	34.7%	32.0%

4.2 Comparison Methods

To evaluate our proposed method, we benchmarked it against six state-of-the-art methods: GFCPH [23], FSRF [55], CPH [14], RSF [21], DeepSurv [22] and IFS [60]. GFCPH and FSRF are the methods considered for group fairness in survival analysis. CPH represents a classical approach to survival analysis. In contrast, RSF is a modern

model employing random forests for censored data analysis, DeepSurv integrates deep learning into survival analysis, and IFS aims to achieve individual fairness in a censorship setting. We excluded other fairness methods from our comparison since they don't cater to censoring. Uniquely, our work is the pioneering effort to harmoniously address both individual and group fairness under a censorship setting.

4.3 Evaluation Metrics

We evaluate our proposed method, GIFC, and existing methods in terms of 5 different metrics. In this section, we first introduce the fairness metrics and then describe performance metrics. For evaluating model fairness, the proposed group fairness metric SPCD and individual fairness metric CCD are employed. It's important to note that existing widely-used fairness metrics could not be applied as they are not adaptable to censorship settings. For evaluating model performance, we utilized a suite of performance metrics: i) C-index [20]: This metric evaluates the concordance between predicted and actual event times. It gauges the probability that for any chosen pair of individuals, their predicted event times align with their actual event times in terms of relative order, where a high value denotes better performance. ii) Brier score [6]: This measures the mean squared difference between predicted probabilities of outcome assignments and the actual outcomes. A superior prediction is denoted by a lower Brier score. iii) Time-dependent AUC [8]: This evaluates the likelihood that, given a random pair of individuals at time t, where one has experienced the event and the other has not, they are correctly ranked in terms of risk, with a high value indicating better performance.

4.4 Experiment Results

4.4.1 Effective Evaluation of GIFC.

In this section, we first evaluate the performance and fairness of our proposed model GIFC. Note that CPH, RSF, DeepSurv, and IFS are not group fairness-away by design. Consequently, we cannot compute CCD for them. we experiment on four datasets with the comparison to the baselines. Each experiment is conducted 10 times. The best results are highlighted in bold. As Table 2 shows, the results indicate that the GIFC model consistently outperforms the baseline models in terms of the SCPD and CCD metrics. A lower SCPD value indicates better group fairness, as it shows smaller disparities in treatment outcomes across different subgroups. In all four datasets, GIFC demonstrates superior group fairness with lower SCPD percentages compared to the baselines. Additionally, the reduced CCD values highlight that the GIFC model achieves this enhanced fairness with fewer sacrifices in individual fairness. This improvement is attributed to GIFC's explicit consideration and balancing of individual losses resulting from group fairness constraints, rather than solely optimizing for minimal performance loss. While performance metrics such as C-index, Brier Score, and Time-dependent AUC exhibit variance, GIFC maintains competitive scores, either surpassing or closely trailing the best-performing models. Overall, the GIFC model demonstrates a remarkable capability to balance group fairness and individual losses effectively across diverse scenarios. This balance is crucial in high-stakes environments where decisions influenced by AI can have significant impacts on individuals' lives. The ability of GIFC to maintain competitive performance on traditional metrics while significantly improving fairness metrics positions it as a potent solution for enhancing fairness in AI-driven decision-making processes.

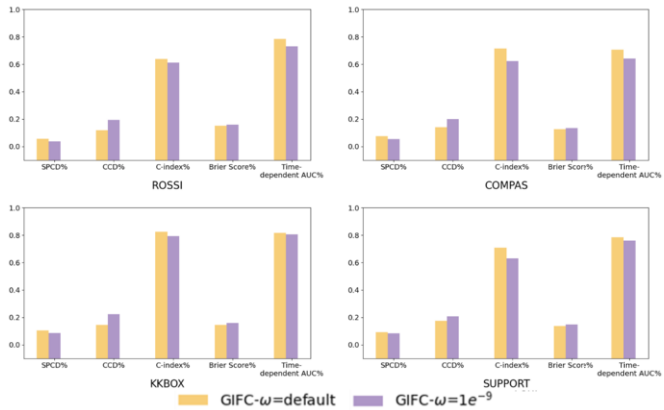

Figure 2: Study on the group and individual fairness and accuracy trade-off on fairness tolerance ω.

4.4.2 The Effect of ω on Model Utility and Fairness.

In this section, we evaluate the effect of fairness tolerance ω on the model's performance. We consider the following settings: i) Default Setting: This setting is inspired by the four-fifths rule [1], a legal criterion from employment law which suggests that the selection rate for any race, sex, or ethnic group should not be less than four-fifths (or 80%) of the rate for the group with the highest rate. Applying this principle, we set ω such that the performance disparity between any two sensitive subpopulations does not exceed 20%. ii) Enforced Fairness: In this more stringent setting, ω is adjusted to enforce nearly identical performance across all sensitive subpopulations, essentially aiming for perfect group fairness. This setting tests the boundaries of fairness by minimizing performance discrepancies to an almost negligible level, ensuring that no group is disproportionately favored or disadvantaged by the model. The results shown in Figure 2 indicate that while a tighter group fairness constraint can improve group fairness performance, it may substantially degrade overall model performance and adversely impact individual fairness. This suggests that a larger proportion of individuals might bear the costs associated with stricter group fairness constraints.

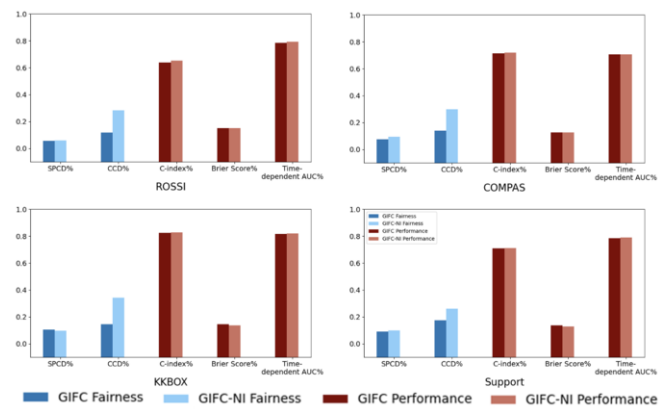

Figure 3: Ablation study results for GIFC and GIFC-NI.

4.4.3 Ablation Study.

We conducted an ablation study to assess the contributions of the two optimizations in our proposed framework. For benchmarking, we introduced the GIFC-NI variant, emphasizing only the group fairness optimization objective. The results show in Figure 3, the GIFC-NI variant, with its sole optimization focus, did not display significant

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

Dataset	Method \ Metrics	SPCD% (↓)	CCD% (↓)	C-index% (↑)	Brier Score% (↓)	Time-dependent AUC% (↑)
ROSSI	GFCPH	9.32	32.77	52.28	14.68	63.92
	FSRF	6.71	27.53	61.44	**14.66**	65.12
	CPH	13.41	-	64.24	19.45	65.46
	RSF	17.17	-	65.47	15.05	79.54
	DeepSurv	13.43	-	**66.67**	14.52	**80.12**
	IFS	23.61	-	65.78	14.79	77.63
	GIFC	**5.63 (16.10%)**	**11.94 (56.63%)**	63.95 (-4.08%)	15.01 (-2.56%)	78.53 (-1.98%)
COMPAS	GFCPH	13.27	35.18	62.16	**12.37**	60.30
	FSRF	10.41	31.38	52.28	13.78	63.92
	CPH	24.51	-	69.24	18.89	67.72
	RSF	27.64	-	72.61	13.02	71.33
	DeepSurv	18.18	-	**75.21**	12.54	**73.68**
	IFS	31.87	-	73.83	12.98	71.67
	GIFC	**7.65 (26.51%)**	**13.91 (55.67%)**	71.47 (-4.97%)	12.73 (-2.91%)	70.67 (-4.09%)
KKBox	GFCPH	16.61	36.72	72.61	**13.55**	73.31
	FSRF	13.75	32.85	78,53	13.57	79.72
	CPH	18.32	-	80.02	17.42	78,47
	RSF	21.41	-	82.32	13.84	80.22
	DeepSurv	20.45	-	**83.01**	14.32	80.69
	IFS	25.12	-	81.97	14.41	80.04
	GIFC	**10.47 (23.85%)**	**14.65 (55.40%)**	82.56 (-0.54%)	14.42 (-7.53%)	**81.75** (1.31%)
Support	GFCPH	13.53	27.14	62.58	13.04	72.72
	FSRF	11.15	22.47	59.28	**12.98**	73.92
	CPH	26.92	-	69.31	20.31	77.64
	RSF	29.17	-	71.73	15.50	80.77
	DeepSurv	21.44	-	**72.32**	14.89	81.13
	IFS	34.67	-	70.03	15.37	**81.38**
	GIFC	**9.17 (17.76%)**	**17.40 (22.56%)**	70.81 (-2.09%)	13.67 (-5.37%)	78.53 (-3.50%)

variations in group fairness or overall performance relative to the full-fledged GIFC. Nevertheless, a marked deterioration in individual fairness was observed. This highlights that overlooking individual fairness while imposing group fairness constraints can inadvertently introduce notable biases in population treatment. While such biases might not directly correlate with specific sensitivity attributes, they might cause a subset of the population to perceive discrimination, thus making the model of individual fairness further degraded.

5 Conclusion

In this paper, we explore a novel research question regarding the achievement of both group and individual fairness in the context of censorship. We introduce a strategy where each individual's experience of group fairness loss is made consistent through randomization. This equalization of the group fairness overhead ensures that no member of the group feels discriminated against. Furthermore, our proposed group-individual fairness ordering distribution offers robust fairness guarantees, ensuring in-group meritocracy and preserving access to outcomes within exponential timeframes. The methodologies and concepts presented here hold promise for quantifying and mitigating biases in various socially sensitive real-world applications. Beyond this, our work delineates a fresh avenue of inquiry, paving the way for future research endeavors aiming to holistically address fairness in AI.

Acknowledgement

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2245895.

References

[1] A. V. Adams. *Toward Fair Employment and the EEOC: A Study of Compliance Procedures Under Title VII of the Civil Rights Act of 1964; Final Report Submitted to Research Division, US Equal Employment Opportunity Commission [by] Avril V. Adams. August 31, 1972*. US Government Printing Office, 1973.

[2] 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.

[3] P. Awasthi, C. Cortes, Y. Mansour, and M. Mohri. Beyond individual and group fairness. *arXiv preprint arXiv:2008.09490*, 2020.

[4] R. Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 514–524, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367.

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

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

[7] J. Chakraborty, S. Majumder, and T. Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

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

[9] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4:123–144, 2021.

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

[11] Z. Chu, S. Ni, Z. Wang, M. Yang, and W. Zhang. History, development, and principles of large language models-an introductory survey. *arXiv preprint arXiv:2402.06853*, 2024.

[12] Z. Chu, Z. Wang, and W. Zhang. Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter, 2024*, pages 34–48, 2024.

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

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

[15] T. V. Doan, Z. Chu, Z. Wang, and W. Zhang. Fairness definitions in language models explained. *arXiv preprint arXiv:2407.18454*, 2024.

[16] M. Du, N. Liu, F. Yang, and X. Hu. Learning credible dnns via incorporating prior knowledge and model local explanation. *Knowledge and Information Systems*, 63(2):305–332, 2021.

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

[18] J. Fox, M. 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.

[19] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

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

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

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

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

[24] W. A. Knaus, F. E. Harrell, J. 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.

[25] H. Kvamme, Ø. Borgan, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

[26] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[28] F. Petersen, D. Mukherjee, Y. Sun, and M. Yurochkin. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955, 2021.

[29] M. M. Rahman and S. Purushotham. Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1452–1462, 2022.

[30] A. Rawls. Theories of social justice, 1971.

[31] N. A. Saxena, W. Zhang, and C. Shahabi. Missed opportunities in fair ai. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 961–964. SIAM, 2023.

[32] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth. Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, 32, 2019.

[33] A. J. Turkson, F. Ayiah-Mensah, and V. 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] K. Turner, N. C. Brownstein, Z. Thompson, I. El Naqa, Y. Luo, H. S. Jim, D. E. Rollison, R. Howard, D. Zeng, S. A. Rosenberg, et al. 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, 2022.

[35] S. Vasudevan and K. Kenthapadi. Lift: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2773–2780, 2020.

[36] C. Wan, W. Chang, T. Zhao, S. Cao, and C. Zhang. Denoising individual bias for fairer binary submatrix detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 2245–2248, 2020.

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

[38] X. Wang, W. Zhang, A. Jadhav, and J. Weiss. Harmonic-mean cox models: A ruler for equal attention to risk. In *survival prediction-algorithms, challenges and applications*, pages 171–183. PMLR, 2021.

[39] X. Wang, T. Gu, X. Bao, L. Chang, and L. Li. Individual fairness for local private graph neural network. *Knowledge-Based Systems*, 268: 110490, 2023.

[40] Z. Wang, G. Narasimhan, X. Yao, and W. Zhang. Mitigating multisource biases in graph neural networks via real counterfactual samples. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 638–647. IEEE, 2023.

[41] Z. Wang, N. Saxena, T. Yu, S. Karki, T. Zetty, I. Haque, S. Zhou, D. Kc, I. Stockwell, A. Bifet, et al. Preventing discriminatory decision-making in evolving data streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[42] Z. Wang, C. Wallace, A. Bifet, X. Yao, and W. Zhang. Fg^2an: Fairness-aware graph generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 259–275. Springer Nature Switzerland, 2023.

[43] Z. Wang, Y. Zhou, M. Qiu, I. Haque, L. Brown, Y. He, J. Wang, D. Lo, and W. Zhang. Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking. *arXiv preprint arXiv:2302.08018*, 2023.

[44] Z. Wang, Z. Chu, R. Blanco, Z. Chen, S.-C. Chen, and W. Zhang. Advancing graph counterfactual fairness through fair representation learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2024.

[45] Z. Wang, J. Dzuong, X. Yuan, Z. Chen, Y. Wu, X. Yao, and W. Zhang. Individual fairness with group awareness under uncertainty. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2024.

[46] Z. Wang, M. Qiu, M. Chen, M. B. Salem, X. Yao, and W. Zhang. Toward fair graph neural networks via real counterfactual samples. *Knowledge and Information Systems*, pages 1–25, 2024.

[47] Z. Wang, D. Ulloa, T. Yu, R. Rangaswami, R. Yap, and W. Zhang. Individual fairness with group constraints in graph neural networks. In *27th European Conference on Artificial Intelligence*. 2024.

[48] D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, pages 570–575. IEEE, 2018.

[49] S. Yan, H.-t. Kao, and E. Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1715–1724, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3340531.3411980.

[50] S. Zhang, D. Zhou, M. Y. Yildirim, S. Alcorn, J. He, H. Davulcu, and H. Tong. Hidden: hierarchical dense subgraph detection with application to financial fraud detection. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 570–578. SIAM, 2017.

[51] W. Zhang. Fairness with censorship: Bridging the gap between fairness research and real-world deployment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22685–22685, 2024.

[52] W. Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 2024.

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

[54] W. Zhang and J. C. Weiss. Fair decision-making under uncertainty. In *2021 IEEE international conference on data mining (ICDM)*, pages 886–895. IEEE, 2021.

[55] W. Zhang and J. C. Weiss. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12235–12243, 2022.

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

[57] W. Zhang, J. Tang, and N. Wang. Using the machine learning approach to predict patient survival from high-dimensional survival data. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016.

[58] W. Zhang, A. Bifet, X. Zhang, J. C. Weiss, and W. Nejdl. Farf: A fair and adaptive random forests classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 245–256. Springer, 2021.

[59] W. Zhang, L. Zhang, D. Pfoser, and L. Zhao. Disentangled dynamic graph deep generation. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 738–746. SIAM, 2021.

[60] W. Zhang, T. Hernandez-Boussard, and J. Weiss. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 14611–14619, 2023.

[61] W. Zhang, Z. Wang, J. Kim, C. Cheng, T. Oommen, P. Ravikumar, and J. Weiss. Individual fairness under uncertainty. In *26th European Conference on Artificial Intelligence*, pages 3042–3049, 2023.