Longitudinal Fairness with Censorship

Wenbin Zhang
Carnegie Mellon University
wenbinzhang@cmu.edu

Jeremy C. Weiss
Carnegie Mellon University
jeremyweiss@cmu.edu

Abstract

Recent works in artificial intelligence fairness attempt to mitigate discrimination by proposing constrained optimization programs that achieve parity for some fairness statistic. Most assume availability of the class label, which is impractical in many real-world applications such as precision medicine, actuarial analysis and recidivism prediction. Here we consider fairness in longitudinal right-censored environments, where the time to event might be unknown, resulting in censorship of the class label and inapplicability of existing fairness studies. We devise applicable fairness measures, propose a debiasing algorithm to bridge fairness with and without censorship, and provide necessary theoretical constructs for these important and socially-sensitive tasks. Our experiments on four censored datasets confirm the utility of our approach.

Introduction

With the rise of big data, artificial intelligence (AI)-based decision making systems are used in a growing number of applications, including many high-impact areas such as healthcare, employment, credit lending and criminal justice (Beutel et al. 2019; Meyer 2018; Liu et al. 2021). There is concern that automated decisions made in this fashion encode and even exacerbate existing real-world disparities, inflicting harm to certain individuals or social groups (Vasudevan and Kenthapadi 2020). This issue has motivated a number of approaches to quantify and mitigate algorithmic unfairness and has given rise to an active field of research deemed AI fairness (Mehrabi et al. 2021). The vast majority of existing studies tackle the problem by taking the existing notions of algorithmic fairness that focus on either individual level fairness, e.g., disparate treatment, to guarantee similar people are treated similarly (Dwork et al. 2012), or group level fairness, e.g., disparate impact, seeking approximate parity of some statistic across different groups (Madhavan and Wadhwa 2020). Moreover, most of the work in this literature studies how to make machine learning algorithms fair in the presence of a class label—where the fairness notions are defined based on the class label, and where the predictive model is trained contingent upon them. Surprisingly, little attention has been given to *censorship* settings in which the time to an event of interest could be inaccessible to the learner, thus making existing fairness notions and approaches inapplicable. In practice, *e.g.*, evaluating the retention rates of each marketing channel in marketing analytics, and judging defendant's criminal recidivism for bail and sentencing in recidivism prediction instruments, the latter is often the case.

In this work, we consider such a censorship setting where the true time to event might be unknown to the learner, while fulfilling the requirements of fair and accurate predictions. Addressing unfairness and censorship simultaneously presents unique challenges, and transferring from respective domains is not straightforward. In contrast to previous works (Verma and Rubin 2018), our goal here is to prevent cases of unequal treatment according to their certain characteristics or sensitive attributes (*e.g.*, gender or race) at the individual and group level. Unlike previous work that was limited to binary sensitive attributes (Zhang and Weiss 2021), our goal also generalizes to an k-way categorical and continuous variables. Our definitions of fairness can therefore be thought of versatile notions focusing on individual and group levels, and inclusive of data with censored individuals.

Our formulation of fairness is motivated by the following observation: in healthcare, the impressive performance of AI systems is balanced against a plethora of observed discriminatory incidents (Rajkomar et al. 2018). As an example, a Propublica report found that state-of-the-art clinical prediction models underperformed on black patients even when a treatment was aimed at a particular type of cancer that disproportionately impacted them (Chen et al. 2020). Such observations extend beyond the medicinal domain with examples in marketing analytics (Chang 2021), actuarial (Frezal and Barry 2019) and recidivism prediction instruments (Angwin et al. 2016), where the common challenge, other than ethical concerns, is the individual's true time to event might be unknown. Thus far the fairness-in-AI community has primarily focused on *no censorship* settings with clearly defined class labels of instances, despite these common censorship scenarios.

Armed with this broader observation of AI unfairness, care must be taken to ensure that an automated decision making system is fair or independent of harmful and sensitive attributes-based stereotypes in the presence of cen-

soring. This motivates the definitions of fairness involving censored individuals and a corresponding algorithm to address discrimination with censorship. Specifically, the novelty of this work comes from four aspects: i) We define a new problem of longitudinal fairness with censorship, which is commonly rooted in socially sensitive applications but remains highly under-explored. ii) Corresponding fairness notions explicitly considering censorship are devised to measure bias in the presence of censorship, as well as a respective fair learner to ensure accurate predictions while also preserving a low discrimination score in longitudinal censorship settings. iii) We theoretically establish the connections of fairness in censored and non-censored settings, offering greater understanding and explanation for AI fairness. iv) Detailed experimental evaluations validating our model with regards to fairness and accuracy on four real-world biased and censored datasets.

Background and Related Work

Longitudinal Biased Data with Censorship

In the typical AI fairness setting, biased data X normally consist of a set of feature instances $x_1, x_2, \cdots, x_n$. Among the features, a special attribute G is referred as the *sensitive attribute* and its attribute values distinguish the discriminated community, *i.e.*, the deprived group, from the privileged community, *i.e.*, the favored group. Instances are also described by their corresponding class labels $y_1, y_2, \cdots, y_n$. However, in the presence of censorship class labels can become inaccessible.

Censored data, in contrast to the typical data representation, contains the survival time T and an event indicator δ in addition to the observed features x, typically represented in the form of a tuple: (x, T, δ). If the event of interest has occurred, T is the duration from the individual entered the study to the time of the event occurring, and δ becomes 1 indicating certainty on the event observation; otherwise T corresponds to the duration between individual entered the study and last follow-up with the individual, and the event indicator $\delta = 0$, i.e., the survival time is censored (Miller Jr 2011).

Compared with AI fairness in supervised settings, addressing discrimination bias in censoring settings leads to censorship on $y_1, y_2, \cdots, y_n$ which limits the applicability of the existing fairness notions. In addition, the uncertainty on $y_1, y_2, \cdots, y_n$ could also further accompany and complicate the biased decision regions. Given the biased and censored data X, the aim of longitudinal AI fairness with censorship is then to model a fair survival function $H(\cdot)$ which makes accurate predictions based on X but also does not discriminate with respect to sensitive attribute G.

AI Fairness

While artificial intelligence is increasingly permeating facets of life, significant concerns on the unfair and discriminatory manner of AI-based systems have been voiced and observed (Beutel et al. 2017). The AI community has responded by proposing a growing body of fairness notions to measure the level of discrimination along with a number of approaches to mitigate bias in order to provide fair decision making systems (Hajian, Bonchi, and Castillo 2016).

The broad set of existing mathematical formulations of fairness can be typically divided into two main families, *individual fairness* and *group fairness*. The former aims to ensure that similarly situated individuals are treated similarly (Dwork et al. 2012) while the latter asks for group level approximate parity of some statistic over class labels (Verma and Rubin 2018). Although a vast of fairness notions exist, most of them formulate fairness depend on class label thus limiting their applicability in censorship settings. Kamrun et al. (Keya et al. 2021) directly extend the existing fairness notions to the application with censoring problems, and it is the only relevant work to the best of our knowledge. However, their definitions exclude the censorship information when measuring discrimination which could introduce substantial bias as censored information can be of importance and cannot simply be ignored (Clark et al. 2003).

The aforementioned fairness definitions can be directly used or adopted as a constraint or a regularizer to enforce fairness, leading to three categories of debiasing mechanisms: *pre-processing approaches* (Kamiran and Calders 2009; Žliobaitė 2017), *in-processing solutions* (Wan et al. 2020; Babaioff, Ezra, and Feige 2021), and *post-processing techniques* (Hardt et al. 2016; Fish, Kun, and Lelkes 2016). The critical limitation of these methods and other fairness works is that they are in need of class label for their unfairness formulations and algorithmic solutions, and fairness when some class labels are unknown has not been well explored (Keya et al. 2021). Our work seeks to alleviate this limitation by jointly addressing bias reduction and censorship management.

Survival Analysis

The critical challenge of the main outcome under assessment could be unknown for a portion of the study group, deemed censorship, hinders the use of many methods of analysis. This motivates the study of *survival analysis* to address the problems of partial survival information access from the study cohort (Clark et al. 2003). The censored data, also known as *survival data*, are generally considered and modeled in terms of two quantitative terms, namely the hazard function and the survival function. The former models the instantaneous rate of event occurs at a specified time t conditioned on surviving to t:

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

The latter is the probability that the event does not occur up to time t and can be determined from the hazard function (and vice versa):

$$S(t|x) = exp(-H(t|x)),\ H(t|x) = \int_0^t h(t|x)dt \quad (2)$$

Given the ubiquity of censored data in real-world applications, survival analysis has gained its popularity in various

applications ranging from medicine to customer and actuarial analytics to predictive maintenance in mechanical operations (De Angelis et al. 1999). Among the various methods proposed for modeling censored data, the Cox proportional hazards model (CPH) (Cox 1972) is the most commonly used in which the multiplicative relation between the risk, as expressed by the baseline hazard function, and covariates is described. More recently, deep neural network structure has also been extended to model the feature interactions of survival data. For example, DeepSurv (Katzman et al. 2018) employs the loss function of CPH with L2 regularization to train the networks. Another line of effort is the tree based methods (Bou-Hamad et al. 2011), particularly random forests due to its superior capabilities in handling nonlinear effect of variables and avoiding restrictive assumptions such as that of proportional hazards (Ishwaran et al. 2008). A comprehensive literature survey covering recent censored data modeling effort is provided in Wang, Li, and Reddy (2019).

With the popularity of survival models, care must be taken to ensure their fairness, the same as other AI approaches. Our in-processing approach incorporates a pairwise-comparison fairness notion in the algorithm design to guide a accuracy-driven and fairness-oriented learning procedure. Relevantly, the survival model is modified to ensure fair risk predictions as in (Keya et al. 2021). Three key differences are that our model: i) does not necessitate a distance metric to be specified, ii) explicitly includes survival information and survival time to mitigate bias in the censoring settings, and iii) is free of hyperparameter tuning to decrease the computational requirements in practice.

Our Approach

Defining Bias with Censorship

Concordance Imparity The presence of censorship in data limits the applicability of commonly used fairness definitions introduced in the existing AI fairness studies. To fill this gap, we introduce *Concordance Imparity (CI)* to specifically account for model unfairness in the presence of censorship. Specifically, CI first considers individual level pairwise comparison based on the consistency between model prediction and true outcomes, then measures, at the group level, whether the discriminative ability of the model is fairly distributed across different groups. Different from the previous definitions (Keya et al. 2021), the survival time and survival information are explicitly involved in CI to avoid important information loss and introducing substantial bias. The sketch of Concordance Imparity is shown in Algorithm 1.

The concordance imparity measurement starts with deciding whether the sensitive attribute G is an continuous attribute (line 1) and discretizes G according to the proposed *fair survival difference* to be discussed in the following section if so (line 2). Next, CI forms all possible pairs of comparison for each individual and omits those incomparable pairs, *i.e.*, the shorter time is censored, and both-censored pairs with identical survival time (line 4-7). The remaining are the permissible pairs across different demographic groups (line 8-9). Among them CI checks three possibilities:

Algorithm 1: Concordance Imparity

Input: Censored and biased dataset D, risk scores r, sensitive attribute G
Output: CI score

1: **if** G is continuous **then**
2: Discretize G according to Equation (6)
3: **end if**
4: **for** each instance d_i in D **do**
5: **for** each instance d_j in D & $d_j \neq d_i$ **do**
6: **if** $t_i < t_j$ & $\delta_i == 0 \mid t_j < t_i$ & $\delta_j == 0 \mid (t_i == t_j$ & ($\delta_i == 0$ & $\delta_j == 0$)) **then**
7: **continue**
8: **else**
9: $P_{G(d_i)==g} = P_{G(d_i)==g} + 1$
10: **end if**
11: **if** $t_i < t_j$ **then**
12: **if** $r(d_i) > r(d_j)$ **then**
13: $C_{G(d_i)==g} = C_{G(d_i)==g} + 1$
14: **else if** $r(d_i) = r(d_j)$ **then**
15: $C_{G(d_i)==g} = C_{G(d_i)==g} + 0.5$
16: **end if**
17: **else if** $t_i > t_j$ **then**
18: **if** $r(d_i) < r(d_j)$ **then**
19: $C_{G(d_i)==g} = C_{G(d_i)==g} + 1$
20: **else if** $r(d_i) = r(d_j)$ **then**
21: $C_{G(d_i)==g} = C_{G(d_i)==g} + 0.5$
22: **end if**
23: **else if** $t_i == t_j$ **then**
24: **if** $\delta_i == 1$ & $\delta_j == 1$ **then**
25: **if** $r(d_i) == r(d_j)$ **then**
26: $C_{G(d_i)==g} = C_{G(d_i)==g} + 1$
27: **else**
28: $C_{G(d_i)==g} = C_{G(d_i)==g} + 0.5$
29: **end if**
30: **else if** $\delta_i == 0$ & $\delta_j == 1$ & $r(d_i) < r(d_j)$ **then**
31: $C_{G(d_i)==g} = C_{G(d_i)==g} + 1$
32: **else if** $\delta_i == 1$ & $\delta_j == 0$ & $r(d_i) > r(d_j)$ **then**
33: $C_{G(d_i)==g} = C_{G(d_i)==g} + 1$
34: **else**
35: $C_{G(d_i)==g} = C_{G(d_i)==g} + 0.5$
36: **end if**
37: **end if**
38: **end for**
39: **end for**
40: CF(G=g) = C_g/P_g
41: **return** CI= $\max_{g,g' \in G \ \& \ g \neq g'} |CF_{(g)} - CF_{(g')}|$

1) if the individual under consideration, d_i, has a shorter survival time, t_i, than the compared individual's survival time t_j, then the concordance count of respective demographic group, $C_{G(d_i)==g}$, that d_i belongs to increments by 1 if the model actually assigns a higher risk score to d_i and by 0.5 if predicted outcomes are tied (line 11-16). 2) Line 17-22 checks the opposite scenario that d_i's survival time is longer than d_j's and counts are incremented similarly. 3) When identical survival times observed (line 23) and neither are censored (line 24), d_i's respective concordance count will be added by 1 on the condition that the predicted outcomes are tied, and by 0.5 otherwise (line 25-29); When the survival times are still the same but not both are censored, $C_{G(d_i)==g}$ increments by 1 if the non-censored individual has a higher predicted risk score and by 0.5 otherwise (line 30-36). Line 40 then evaluates *concordance fraction (CF)*, the group-wise correct pairwise ordering, and the final CI score is measured as the largest deviation of discriminative abilities across different demographic groups of the model (line 41). The lower

the concordance imparity score the fairer the model.

Note that in comparison to existing fairness notions mainly focus on binary protected categorical attributes (Verma and Rubin 2018), other than the explicit inclusion of censorship information, concordance imparity also looks at the generalization of measuring the level of discrimination to k-way categorical attributes. This is done by reformulating the general discrimination measurement to consider the largest difference among sub-community, which is equivalent to the typical fairness definition when $k = 2$. In addition, CI also considers the discrimination in regards to continuous attribute domain. Specifically, the allowing test (Han, Kamber, and Pei 2011) is first used to explore potential binary split candidates, then the allowed split with the largest merit achieved is selected as the threshold for splitting. The merit is gauged according to the proposed *fair survival difference* to be discussed hereafter. The calculation of CI can then proceed the way as the categorical attribute after such a discretization. This gives us a generalized definition extending CI to fair regression tasks, enabling an inclusive discrimination evaluation consisting of censored individuals.

Fair Calibration To exploit the semantic information of survival probabilities produced by the model, which are also labels for individuals (Haider et al. 2020), we propose *fair calibration (FC)* to measure whether the model creates systematic disparity from use of the predicted probabilities.

In summary, fair calibration of survival probabilities is assessed by: i) plotting group-wise (i.e., $\forall g \in G$) observed proportions versus predicted probabilities and ii) by calculating corresponding fair calibration validity. The first step examines the agreement between the predicted probabilities of the model with the observed outcome. To do so, FC sorts and splits the predicted probabilities for a particular time t for each demographic group into deciles, then checks whether these probabilities are sufficiently close to the observed proportions. Note that smaller subgroups will lead to greater statistical uncertainty about within-group predictive and fairness performance, but provide more similarity of instances within subgroups. Fewer subgroups result in the opposite. Convention is to use deciles, which is what we follow. In addition, the observed proportions could be unknown due to the censorship. To this end, the Kaplan-Meier (KM) curve estimate (Ranstam and Cook 2017) is employed and the significance of these group-wise results respect to these bins are defined by the Hosmer-Lemeshow (HL) goodness-of-fit test statistic (Hosmer and Lemesbow 1980):

$$HL_g(S(t|x)) = \sum_{i=1}^{B} \frac{(KM_{ig} - \bar{p_{ig}})^2 n_{ig}}{\bar{p_{ig}}(1 - \bar{p_{ig}})} \tag{3}$$

where B represents the number of bins, KM_{ig} is the KM estimated probability in the ith decile at time t, $\bar{p_{ig}}$ is the predicted probability for individuals in the ith decile and n_{ig} is the number of observations in decile i. Note that all of them are group-wise, i.e., $G = g$.

Based on the group-wise examined level of agreements, the second step of FC evaluates consistency across different demographic groups as shown in Equation (4), involving the first two fair calibrated scenarios and the third biased calibrated scenario otherwise: i) representation consistency: the predicted probabilities are representative of the actual probabilities; the p-value of each group's HL statistic passes the test with a value greater than 0.05, ii) difference consistency: representation inconsistent but the difference between predicted probabilities and actual probabilities (i.e., $\triangle p_g / \triangle p_{g'}$) is accordant across subgroups; the p-value of each difference test among subgroups evaluated by Wilcoxon signed-rank test (Woolson 2007) is not smaller than 0.5, iii) Neither representation nor difference is consistent.

$$FC = \begin{cases} \text{fair calibrated,} & p(HL_g(S(t|x))) \geq 0.05 \\ & \qquad \forall g \in G \\ \text{fair calibrated,} & wilcoxon(\triangle p_g, \triangle p_{g'}) > 0.5 \\ & \qquad \forall g, g' \in G \\ \text{biased calibrated,} & \text{otherwise} \end{cases} \tag{4}$$

Mitigating Bias with Censorship

Armed with the afore established fairness statistics, measuring unfairness in the presence of censorship becomes feasible. This section then serves to fulfill the subsequent bias mitigation amidst censorship.

The proposed approach follows the general idea of random forests (RF) by constructing an array of base learners to improve the predictive ability. In particular to censored data, RF is also nonparametric while enjoying the merits of nonlinear interactions modeling (Wang, Li, and Reddy 2019). However, such ensemble methods aim to optimize for data encoding for predictive performance, and fairness, which we desired to add, is imperceptible (Ishwaran et al. 2008). In this work, to jointly optimizing for censored data encoding and debiasing, we propose *Fair Survival Random Forests (FSRF)* which extends the RF model in two ways: i) by introducing a new splitting criterion that jointly considers the reduction of an attribute split w.r.t. impurity and also w.r.t. discrimination, ii) by illustrating the way to provide fair risk predictions amidst censorship.

The information gain and Gini impurity, when censorship is absent, are commonly used splitting criteria to guide the induction of the tree for classification performance (Han, Kamber, and Pei 2011). However, the presence of censorship leads to inaccessibility of class label thus making their computation impractical. The survival difference can be instead used to measure the impurity reduction for candidate splitting evaluation. In FSRF, such survival difference between different groups are evaluated by the *logrank test* (Bland and Altman 2004):

$$SD = \frac{\sum_{j=1}^{k}(O_j - E_j)}{\sqrt{\sum_{j=1}^{k} V_j}} \sim N(0, 1) \tag{5}$$

where O_j and E_j represent the observed number of events and the expected number of events, respectively with V_j being the variance of O_j. The candidate with a larger logrank

test therefore leads to more similarity within child nodes but also more dissimilarity among child nodes if it is being selected for splitting.

We then combine concordance imparity and survival difference as a conjunctive criterion that takes both predictive performance and fairness into consideration. We define the conjunctive criterion *fair survival difference (FSD)* as:

$$FSD = \begin{cases} \log SD - \log CI & \text{if CI} \neq 0 \\ +\infty & \text{otherwise} \end{cases} \quad (6)$$

Intuitively, FSD closely ties SD and CI. When the candidate attributes that are free of discrimination, i.e., CI equals 0, FSD becomes positive infinite to prioritize fair splitting.

In practice, the calculation of CI depends on the distribution that a potential splitting could lead to and the associated risk predictions based on the distribution. FSRF predicts the risk as the cumulative hazard function $H(t|x)$ to have a direct interpretation of the expected number of events, and it is the intermediate function between hazard and survival functions for direct derivation when needed. Formally, the risk score is estimated by the Nelson-Aalen estimator (Borgan 2014) as:

$$H(t|x) = \sum_{j \leq t} \frac{d_j}{n_j} \quad (7)$$

where d_j and n_j represent the number of individuals experiencing events and have not experienced the event at time j respectively, and t is evaluated as the last event time. In response to cases within the same node sharing identical class label in non-censoring trees, all individuals within the node of FSRF have the same risk score which is used for splitting evaluations but also final risk predictions.

Bridging Fairness in the Presence and Absence of Censorship

The previously proposed fairness definitions and debiasing algorithm explicitly consider the indispensable survival information and lay the groundwork for AI fairness in the presence of censorship. In addition, it is also desire to understand the connections of AI fairness in the presence and absence of censorship to build fundamental theoretical frameworks for AI fairness. So are of practitioners and policy makers' interests to help them have an additional layer of understanding of AI fairness for fair decision making navigation and customization. This section serves to fill this gap.

As previously discussed, the standard AI fairness techniques, such as the most widely used statistical parity, are not suitable in the presence of censoring in the data (Verma and Rubin 2018). Here, we take the devised concordance imparity as the illustrative example to connect AI fairness in the presence and absence of censorship, so as to facilitate the study of fairness with censorship. Recall that CI first measures subgroup-wise fraction of concordant pairs, then gauges concordance difference between different subgroups defined by the sensitive attributes. The first part of CI can therefore be thought of as an extension of the standard concordance index (C-index) (Li et al. 2016) in subgroup-wise and then along with the second part as the weighted subgroup based area under the receiver operating curve (AUC) difference in no censoring settings. We will next elaborate this connection.

Table 1: An overview of the datasets.

	SUPPORT	ROSSI	COMPAS	KKBox
Sample #	8,873	432	10,325	2.8M
Censored%	0.320	0.736	0.732	0.347
Feature #	14	9	14	18
Sensitive Attribute	gender	race	race	gender
Sensitive Value	female	African American	African American	female

Starting from the standard C-index, it is a "global" index for validating the predictive ability of the model. Specifically, it is the fraction of pairs within the whole group, where the observation experienced the event of interest had a higher risk score than an observation who experienced the event later or had not experienced the event, representing the global assessment of the model's discrimination power. By definition, the C-index is a generalization of the Wilcoxon-Mann-Whitney statistics (Austin and Steyerberg 2012) and thus of the AUC with equivalence in binary classification in the absence of censorship. The difference between the concordance part of CI and C-index is that the concordance of CI measures subgroup-wise fraction of concordant pairs in comparison to the whole group-wise concordant probability of C-index. Note that the subgroup-wise concordance calculation of CI compares an observation with both intra group and inter group observations, i.e., with all remaining observations other than itself. The concordance of CI can therefore be regarded as a subgroup based AUC, abbreviated as sAUC, in contract to the standard global AUC of C-index. Next, based on the obtained sAUC from different subgroups, the imparity part of CI gauges concordance difference among them. Finally, CI can be interpreted as the largest deviation among weighted sAUC, and CI's counterpart of the original formulation (e.g., line 41 in Algorithm 1) in the absence of censorship is:

$$CI = \max_{g,g' \in G \ \& \ g \neq g'} |sAUC_{(g)} - sAUC_{(g')}| \quad (8)$$

where $sAUC_{(G=g)}$ is the weighted subgroup based AUC from subgroup that defined by g and is mathematically represented as:

$$sAUC_{(g)} = \frac{n_g sAUC_g + \sum_{j=1}^{|\{G-g\}|} n_{g'} sAUC_{g'}}{\sum_{j=1}^{|G|} n_j} \quad (9)$$

where $sAUC_g$ and $sAUC_{g'}$ are AUC values when comparing with intra- and inter-subgroup observations, respectively, and n_j represents respective number of comparable pairs of each subgroup.

Table 2: Performance comparison of all methods on various datasets. The best results are marked in bold.

Datasets	Metrics / Method	CI%	Fair Calibration	C-index%	Brier Score%	Time-dependent AUC%
SUPPORT	IDCPH	19.12	Not fair calibrated	69.08	31.16	76.17
	GDCPH	13.12	**Fair calibrated**	75.12	24.46	78.21
	CPH	17.45	Not fair calibrated	74.11	21.21	80.02
	RSF	20.11	Not fair calibrated	75.18	16.64	81.01
	DeepSurv	18.65	Not fair calibrated	75.65	16.11	80.68
	FSRF	**9.21**	**Fair calibrated**	**76.17**	**13.23**	**82.86**
ROSSI	IDCPH	15.31	Not fair calibrated	52.28	18.73	77.32
	GDCPH	9.32	**Fair calibrated**	59.34	22.87	78.51
	CPH	11.43	Not fair calibrated	64.24	17.67	77.12
	RSF	16.53	Not fair calibrated	65.56	15.12	79.32
	DeepSurv	12.32	Not fair calibrated	66.67	14.71	77.17
	FSRF	**8.92**	**Fair calibrated**	**69.02**	**12.69**	**79.65**
COMPAS	IDCPH	25.18	Not fair calibrated	62.16	25.03	63.78
	GDCPH	11.77	**Fair calibrated**	72.16	16.32	66.21
	CPH	22.43	Not fair calibrated	69.24	20.35	65.15
	RSF	25.32	Not fair calibrated	72.61	15.62	71.76
	DeepSurv	16.72	Not fair calibrated	75.12	**13.42**	71.83
	FSRF	**9.63**	**Fair calibrated**	**76.24**	13.81	**72.33**
KKBox	IDCPH	17.79	Not fair calibrated	72.61	21.23	69.73
	GDCPH	14.98	**Fair calibrated**	79.45	19.92	73.03
	CPH	18.91	Not fair calibrated	80.02	18.17	72.95
	RSF	21.14	Not fair calibrated	82.32	14.24	78.18
	DeepSurv	20.66	Not fair calibrated	**83.01**	14.33	80.71
	FSRF	**14.42**	**Fair calibrated**	82.43	**13.13**	**82.16**

Experimental Results

Dataset Description

We validate our model on four real-world censored datasets with socially sensitive concerns: i) The *SUPPORT* dataset is from a large study to understand prognoses preferences outcomes and risks of treatment by analyzing the survival time of inpatients (Knaus et al. 1995). ii) The *ROSSI* dataset pertains to predict the reoffending risk score of convicted criminals from Maryland state prisons, who were followed up for one year after release (Fox, Carvalho et al. 2012). iii) The landmark algorithmic unfairness *COMPAS* dataset to predict recidivism from Broward County (Angwin et al. 2016). iv) The *KKBox* dataset from the WSDM-KKBox's Churn Prediction Challenge (Kvamme, Borgan, and Scheel 2019) to study users' risk scores of canceling their subscription from KKBox. Table 1 is a summary description of them. Note that survival time and censoring information are explicitly included in our study to specifically account for censorship.

Comparison Methods

We compare FSRF against five baselines to evaluate its theoretical design: i) two recent proposed fair survival models *IDCPH* and *GDCPH* (Keya et al. 2021), which are the most competitive approaches among several variants proposed therein, and are the only works for fair survival analysis problem to the best of our knowledge, ii) along with the baseline therein, the most commonly used survival analysis tool *CPH* (Cox 1972), iii) state-of-the-art random forests based non-linear survival model *RSF* (Ishwaran et al. 2008), and iv) the most recent deep model on survival analysis *DeepSurv* (Katzman et al. 2018). We do not compare with other fairness baselines due to their inapplicability in the presence of censorship.

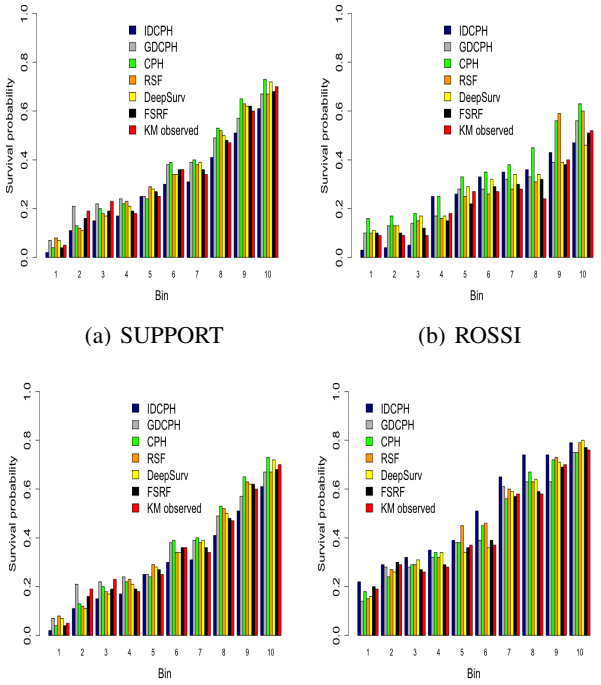
Performance Comparison

Due to the presence of censoring in the data, the standard evaluation metrics of AI fairness such as accuracy and statistical parity are not suitable for measuring the performance in AI fairness with censorship (Verma and Rubin 2018). Instead, in addition to the previously tailored fairness measures considering censorship, the typical survival accuracy metrics, including the C-index, Brier score and Time-dependent AUC, are utilized to evaluate the predictive performance of our model and other baselines. The *C-index* (Harrell et al. 1982) evaluates a model's discrimination power in terms of correct pairwise ordering, and is a generalization of the area under ROC curve (AUC) in the presence of censorship. The *Brier score* (Brier and Allen 1951) is roughly the mean squared difference of the probability estimations assigned to possible outcomes and the actual outcome. Different from C-index, the lower the Brier score the merrier. Finally, the *Time-dependent AUC* (Chambless and Diao 2006) tests how well a model can distinguish individuals who experienced the event of interest from those have not prior to or at time t, and thus the model with a higher Time-dependent AUC score is desired. Table 2 provides detailed 5-fold cross validation results of all methods on various lon-

Table 3: Prediction performance confusion matrix for FSRF.

Datasets	C-index% discriminated		C-index% privileged		Brier Score% discriminated		Brier Score% privileged		Time-dependent% AUC discriminated		Time-dependent % AUC privileged	
	FSRF-	FSRF	FSRF-	FSRF	FSRF-	FSRF	FSRF-	FSRF	FSRF-	FSRF	FSRF-	FSRF
SUPPORT	60.05	69.64	80.16	78.85	25.54	17.65	8.87	10.14	73.45	79.67	87.52	85.92
ROSSI	57.71	63.78	74.24	72.7	21.03	16.36	9.87	10.22	69.98	73.77	84.87	82.21
COMPAS	54.82	68.88	80.14	78.51	18.76	16.71	7.66	11.67	62.81	65.65	77.62	75.31
KKBox	64.53	70.22	85.67	84.64	18.87	14.97	7.16	9.12	72.31	78.03	85.87	84.52

gitudinal biased data with censorship.



(a) SUPPORT (b) ROSSI

(c) COMPAS (d) KKBox

Figure 1: The fair calibration plots of the discriminated community. The color and height of the bar represent different methods and corresponding probabilities. The KM empirical probability is marked in red and FSRF is in black.

The proposed model wins almost every metric across all of the datasets (Table 2). This demonstrates that our proposed method has a superior debiasing capability while maintaining strong predictive performance in the presence of censorship. Specifically, our new FSRF dominates all other methods when diminishing discrimination with censorship. This result is especially important when we contrast to other fair models proposed for censoring settings, which verifies the necessity of including survival information and survival time to mitigate bias with censorship as well as the drawback of using task-specific similarity metrics. In addition, our model, on all datasets tested, does not suffer from performance instability as other methods do, indicating FSRF is a more robust approach to building fair model with censorship. In terms of predictive performance, FSRF outperforms other models on most metrics across all datasets. This reflects that FSRF is able to handle and utilize both censored and uncensored instances when building fair model with censorship. Additionally, FSRF, in contrast to other fair survival baselines, performs no parameter tuning, thus benefiting end user with simplicity while making fair decisions in the presence of censorship.

We also note that our fairness regularizer is able to actually improve predictive performance (Table 2). To have a better understanding of this phenomena, we further analyze the predictive performance confusion matrix of FSRF, according to the sensitive attribute that defines the discriminated community and privileged community as well as with and without our fairness constraints, represented by FSRF and FSRF-. Table 3 summarizes the results. As one can see, improvement on the characterization of discriminated community is indeed achieved by including fairness attention in our method. Also, the improved overall prediction performance demonstrates the potential additional rewards of the debiasing design of FSRF.

The risk prediction of FSRF is fairly calibrated as each demographic group's p-value passes its significance test, suggesting FSRF's predicted probabilities are representative of corresponding community's true probabilities (Table 2). To see the full picture behind the p-values, Figure 1 graphs the predicted probabilities by FSRF in comparison to the KM empirical probabilities (fair calibration plots of privileged community are omitted due to space constraints). In the visualization, the heights of dark bars are always close to the red bars' while bars in other colors do not follow this pattern which conclusively match with the results of fair calibration, suggesting FSRF is indeed an effective fair risk predictor.

Conclusion

Despite the increasing attention on AI fairness, existing studies have mainly focused on no censorship settings. This paper tackles fairness with censorship which is particularly prevalent in many real-world socially sensitive applications. To accomplish this objective, we devised generalized censored-specific fairness notions to quantify unfairness along with a unified debiasing algorithm to mitigate discrimination in the presence of censorship. The results on real biased datasets with censorship show our propose techniques are versatile in censoring settings. This work defines a new research problem and opens possibilities for future work on AI fairness with a broader applicability to practical scenarios concerning fairness.

References

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

Austin, P. C.; and Steyerberg, E. W. 2012. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC medical research methodology*, 12(1): 1–8.

Babaioff, M.; Ezra, T.; and Feige, U. 2021. Fair and Truthful Mechanisms for Dichotomous Valuations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1795–1803.

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

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Bland, J. M.; and Altman, D. G. 2004. The logrank test. *Bmj*, 328(7447): 1073.

Borgan, Ø. 2014. Nelson–Aalen Estimator. *Wiley StatsRef: Statistics Reference Online*.

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

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

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

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

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

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

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

De Angelis, R.; Capocaccia, R.; Hakulinen, T.; Soderman, B.; and Verdecchia, A. 1999. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in medicine*, 18(4): 441–454.

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

Fish, B.; Kun, J.; and Lelkes, Á. D. 2016. A confidence-based approach for balancing fairness and accuracy. In *SDM*, 144–152.

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

Frezal, S.; and Barry, L. 2019. Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 1–10.

Haider, H.; Hoehn, B.; Davis, S.; and Greiner, R. 2020. Effective Ways to Build and Evaluate Individual Survival Distributions. *J. Mach. Learn. Res.*, 21: 85–1.

Hajian, S.; Bonchi, F.; and Castillo, C. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the SIGKDD international conference on knowledge discovery and data mining*, 2125–2126.

Han, J.; Kamber, M.; and Pei, J. 2011. Data Mining: Concepts and Techniques.

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

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

Hosmer, D. W.; and Lemesbow, S. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10): 1043–1069.

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

Kamiran, F.; and Calders, T. 2009. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, 1–6.

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

Keya, K. N.; Islam, R.; Pan, S.; Stockwell, I.; and Foulds, J. 2021. Equitable Allocation of Healthcare Resources with Fair Survival Models. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 190–198. SIAM.

Knaus, W. A.; Harrell, F. E.; Lynn, J.; Goldman, L.; Phillips, R. S.; Connors, A. F.; Dawson, N. V.; Fulkerson, W. J.; Califf, R. M.; Desbiens, N.; et al. 1995. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3): 191–203.

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

Li, Y.; Wang, J.; Ye, J.; and Reddy, C. K. 2016. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1715–1724.

Liu, R.; Jia, C.; Wei, J.; Xu, G.; Wang, L.; and Vosoughi, S. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14857–14866.

Madhavan, R.; and Wadhwa, M. 2020. Fairness-Aware Learning with Prejudice Free Representations. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2137–2140.

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

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

Miller Jr, R. G. 2011. *Survival analysis*, volume 66. John Wiley & Sons.

Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G.; and Chin, M. H. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12): 866–872.

Ranstam, J.; and Cook, J. 2017. Kaplan–Meier curve. *British Journal of Surgery*, 104(4): 442–442.

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

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7. IEEE.

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

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

Woolson, R. F. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, 1–3.

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

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