Equitable Allocation of Healthcare Resources with Fair Cox Models

Kamrun Naher Keya,[1] Rashidul Islam,[1] Shimei Pan,[1] Ian Stockwell,[2] and James Foulds[1]
[1]Department of Information Systems
[2]The Hilltop Institute
University of Maryland, Baltimore County
{kkeya1, islam.rashidul, shimei, jfoulds}@umbc.edu, istockwell@hilltop.umbc.edu

Abstract

Healthcare programs such as Medicaid provide crucial services to vulnerable populations, but due to limited resources, many of the individuals who need these services the most languish on waiting lists. Survival models, e.g. the Cox proportional hazards model, can potentially improve this situation by predicting individuals' levels of need, which can then be used to prioritize the waiting lists. Providing care to those in need can prevent institutionalization for those individuals, which both improves quality of life and reduces overall costs. While the benefits of such an approach are clear, care must be taken to ensure that the prioritization process is fair or independent of demographic information-based harmful stereotypes. In this work, we develop multiple fairness definitions for survival models and corresponding fair Cox proportional hazards models to ensure equitable allocation of healthcare resources. We demonstrate the utility of our methods in terms of fairness and predictive accuracy on two publicly available survival datasets.

Introduction

Publicly funded healthcare programs such as Medicaid provide crucial services to vulnerable populations. Most states have subprograms within their Medicaid programs meant to serve specific target populations. These programs are known as "waivers," since each state must ask the federal government to waive some portions of the original Medicaid statute in order to better serve their population. With this expanded authority, states can include coverage for services that are not covered under traditional Medicaid programs (such as home and community-based long-term care), expand the financial eligibility requirements for participation, and limit the enrollment of each program for cost containment purposes. Many waivers are built to serve older adults or individuals with developmental/physical disabilities better by keeping them out of institutional settings (such as nursing homes). Participation in these programs with more services, relaxed financial eligibility, and limited enrollment becomes a necessarily scarce resource in need of allocation. The traditional method of allocating spots in these programs is a "first in, first out" model, where the next individual to enter the program is the one who has been waiting the longest.

Artificial intelligence (AI) can potentially improve this situation by predicting individuals' risk of institutionalization, which can then be used to prioritize the list of individuals who would like to participate in the program but for whom a spot on the waiver is not available (also known as the "waitlist"). On October 1, 2019, the Maryland Department of Health deployed an AI system which performs a needs-based prioritization of the Medicaid waitlist as a function of predicted time to institutionalization, i.e. admission to a nursing home.[1] While the benefits of such an approach are clear, care must be taken to *ensure that the prioritization process is fair*. AI models can have impacts with lawful, moral or ethical consequences when utilized to predict outcomes in societal, governmental, and public sector applications. Structural and systemic processes, often unfair and/or biased against certain groups of people, impact individuals' lives and and hence their data (Barocas and Selbst 2016), for example based on age, race, gender, nationality, class or sexual orientation. Since systemic bias is inherent in data, machine learning models must account for this to avoid creating discriminatory decisions. In recent years, the machine learning (ML) community has conducted a notable amount of research on algorithmic bias (Dwork et al. 2012; Hardt, Price, and Srebro 2016; Kusner et al. 2017; Foulds et al. 2020b) which aims to learn non-discriminatory predictive models by enforcing constraints in the training phase (Goel, Yaghini, and Faltings 2018; Chouldechova 2017; Kilbertus et al. 2017; Bolukbasi et al. 2016; Zhao et al. 2017).

Like any data that involves individuals from different demographics, health data is subject to bias in various manners, and the expanding amount and types of data that are accessible today can make it difficult to distinguish where bias can emerge (Ferryman and Pitcan 2018). The goal of this work is therefore to develop AI techniques for attenuating harmful bias in the allocation of healthcare resources.

To predict individuals' risk of institutionalization, a natural approach is to use survival models. The Cox proportional hazards (CPH) (Cox 1972) model is particularly appropriate in this research, as the multiplicative relationship between covariates and risk aids explainability. It is crucial to ensure fairness in the risk prediction task to obtain fair healthcare resource allocation with survival models such as

[1]https://tinyurl.com/yy3odnmq

CPH. Though the fairness community has proposed various fairness definitions (Feldman et al. 2015; Dwork et al. 2012; Hardt, Price, and Srebro 2016; Kearns et al. 2018; Foulds et al. 2020b) to measure different aspects of societal or demographic biases in AI systems, to the best of our knowledge there are currently no fairness definitions specific to survival models. In this paper, we propose multiple fairness definitions for survival models and develop corresponding fair learning algorithms. The models' risk scores can then be used to fairly prioritize the Medicaid waitlist.

The main contributions of this work include:

- To the best of our knowledge, this is the first investigation on fairness for survival models to ensure equitable allocation of healthcare resources.
- We extend three categories of fairness definitions to measures bias in the survival analysis problem.
- We develop corresponding fair learning algorithms for CPH models to ensure fair risk predictions.
- We perform extensive experiments validating our models with regards to both fairness and accuracy on two publicly available survival datasets.

Background and Related Work

In this section, we define survival data and the Cox proportional hazards model, a popular method for survival analysis. In addition, we discuss related work on fairness in AI.

Survival Data

Survival data (Katzman et al. 2018; Lee et al. 2018) contains three pieces of information for each individual: 1) observed covariates/features x, 2) actual time of the event T, and 3) event indicator E. If an event, e.g. death, has occurred, T corresponds to the elapsed time between when the covariates were first collected and the time of the event occurring, and the event indicator is $E = 1$. If an event is not observed, T corresponds to the elapsed time between the collection of the covariates and the last contact with the individual subject, e.g. end of the study, and E becomes 0. In this scenario, the individual subject is said to be *right-censored*. In standard regression models missing data such as right-censored data may typically be discarded. In survival analysis, however, right-censored data (e.g. data on individuals who survive to the end of the study) is important and cannot simply be ignored without introducing substantial bias. Right-censored data therefore requires special consideration in this context.

Cox Proportional Hazards Model (CPH)

The Cox proportional hazards (CPH) model (Cox 1972) is the most widely used model for survival analysis (Lee et al. 2018). It is a semiparametric model often used in clinical (and many other) settings for modeling and predicting the time until a particular event occurs, e.g. death of a patient.

Let $S(t)$ be the probability that the event does not occur before time t (the *survival function*). The key concept to define these models is the *hazard function*, defined to be the instantaneous rate that the event, e.g. death or institutionalization, occurs at time t, here supposing that time is continuous. The hazard function is defined as

$$h(t) \triangleq \lim_{\Delta t \to 0^+} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} . \quad (1)$$

The CPH model specifies the hazard function via

$$h(t) = h_0(t) \exp(\beta^\intercal \mathbf{x}) , \quad (2)$$

where h_0, called the *baseline hazard*, is the hazard value regardless of features $\mathbf{x}$, and β is a parameter vector. The survival function is determined from the hazard function via

$$S(x) = \exp(-H(t)), H(t) = \int_0^t h(u)du . \quad (3)$$

To perform Cox regression, the parameter β can be learned by optimizing the Cox partial likelihood (Faraggi and Simon 1995; Katzman et al. 2018). The partial likelihood is the product of the probability at each event time T_i that the event E_i has occurred to individual i, given the set of individuals still at risk at time T_i and can be calculated as

$$L_c(\beta) = \prod_{i:E_i=1} \frac{exp(\beta^\intercal \mathbf{x}_i)}{\sum_{j \in \Re(T_i)} \exp(\beta^\intercal \mathbf{x}_j)} , \quad (4)$$

where the product is defined over the set of patients with an observable event $E_i = 1$ and the risk set $\Re(t) = \{i : T_i \geq t\}$ is the set of patients still at risk of failure at time t.

The CPH model assumes that an individual's risk of an event occurring is a linear combination of the patient's covariates, referred to as the linear proportional hazards condition. Since this assumption may be too simplistic in many applications such as personalized treatment recommendations (Katzman et al. 2018), recently deep neural networks (Faraggi and Simon 1995; Katzman et al. 2018; Lee et al. 2018; Huang et al. 2019) have been applied to CPH models to solve the problem of nonlinear survival analysis.

Fairness in AI

The increasing impact of artificial intelligence (AI) and machine learning technologies on many facets of life, from commonplace movie recommendations to consequential criminal justice sentencing decisions, has prompted concerns that these systems may behave in an unfair or discriminatory manner (Barocas and Selbst 2016; Munoz, Smith, and Patil 2016; Noble 2018). A number of studies have subsequently demonstrated that bias and fairness issues in AI are both harmful and pervasive (Angwin et al. 2016; Buolamwini and Gebru 2018; Bolukbasi et al. 2016). The AI community has responded by developing a broad array of mathematical formulations of fairness and learning algorithms which aim to satisfy them (Dwork et al. 2012; Hardt, Price, and Srebro 2016; Berk et al. 2017; Zhao et al. 2017). An overview of fair AI is given by (Berk et al. 2018).

While a number of fairness definitions have been proposed in the literature, at the highest level there are three broad categories of fairness measures. **Individual fairness** definitions aim to ensure that *similar individuals obtain similar outcomes* under the algorithm in question. **Group**

fairness definitions aim to preserve fairness at the level of groups of individuals, e.g. women, the elderly, or African Americans. Finally, **intersectional fairness** definitions are those for which fairness is to be ensured for a specified set of subgroups defined by the protected attributes (Kearns et al. 2018; Hebert-Johnson et al. 2018; Foulds et al. 2020b). These fairness definitions can be slightly modified to form a fairness penalty that can be added as a constraint or a regularization term to the existing optimization objective to enforce fairness in the algorithm (Calders and Verwer 2010; Zafar et al. 2017b; Zafar et al. 2017a; Foulds et al. 2020b).

(Angwin et al. 2016) applied Cox models to the problem of criminal recidivism prediction, in order to detect racial bias in the offender's risk score prediction. However, there is no prior work that introduces and enforces formal fairness definitions for the fair survival analysis problem.

Methods

In this section, we describe our methodology to ensure fair risk predictions with survival models. First, we extend the three broad categories of fairness measures to precise application of fairness in survival analysis problems. We then develop a simple learning algorithm for fair survival models.

Fairness Definitions for Survival Models

Fairness in healthcare is a multi-stakeholder issue, and so we cannot simply settle it with a single solution. We instead provide stakeholders with three different proposed implementations of fairness for survival models.

Individual fairness: Individual fairness (Dwork et al. 2012) aims to ensure that a system or model produces similar outcomes to similar individuals. In the context of survival models, we define individual fairness (F_i) inspired by (Dwork et al. 2012) as follows:

$$F_i = \sum_{i=1}^{N^{(test)}} \sum_{j=i+1}^{N^{(test)}} \max(0, |\bar{h}_\beta(\mathbf{x}_i) - \bar{h}_\beta(\mathbf{x}_j)| - D(\mathbf{x}_i, \mathbf{x}_j)) \,, \tag{5}$$

where $\bar{h}_\beta(\mathbf{x}) = \exp(\beta^\intercal \mathbf{x})$, the hazard function where the base hazard $h_0(t)$, which is not individual-specific, is dropped, and $D(x_i, x_j)$ is a distance metric (e.g. Euclidean distance) between x_i and x_j encoding fair similarity, on the same scale as $|\bar{h}_\beta(\mathbf{x}_i) - \bar{h}_\beta(\mathbf{x}_j)|$. Note that this penalizes differences in predicted hazard scores that exceed the distance between the data points. Here, we can make use of knowledge of the individuals who are to be in the test set such as the individuals on the waitlist for care (a *transductive* approach to fairness).

Group fairness: In group fairness definitions, e.g. demographic parity (Dwork et al. 2012), a system is fair if outcomes are distributed fairly across different demographic groups, e.g. different genders or races. We define the group fairness (F_g) measures for survival models as

$$F_g = \max_{a \in A} |\bar{h}_\beta(a) - E[\bar{h}_\beta(\mathbf{x})]| \,, \tag{6}$$

$$\bar{h}_\beta(a) \triangleq \int_{\mathbf{x}} \exp(\beta^\intercal \mathbf{x}) p(\mathbf{x}|a) \,, \tag{7}$$

the worst-case deviation of the *per-group expected hazard function* $\bar{h}_\beta(a)$ from the population average hazard where A is the set of values in the protected attribute. We estimate the above integral via an average over the empirical data.

Intersectional fairness: Intersectional fairness (Kearns et al. 2018; Hebert-Johnson et al. 2018; Foulds et al. 2020b) definitions consider subgroups of protected groups, usually defined to be their intersecting subgroups. This can be used to enforce fairness metrics that encode the principle of intersectionality (Crenshaw 1989), namely that individuals at the intersections of protected groups, e.g. along lines of race and gender, are vulnerable to additional harms and should be protected (Foulds et al. 2020b). In this case, $A = S_1 \times S_2 \times \ldots S_K$ is a space of multi-dimensional protected attributes. Building on our earlier work on the *differential fairness* metric (Foulds et al. 2020b), intersectional fairness (F_ϵ) for survival models can be extended as

$$F_\epsilon = \max_{s_i \in A, s_j \in A} |\log \bar{h}_\beta(s_i) - \log \bar{h}_\beta(s_j)| \,, \tag{8}$$

a worst case of log-ratios of "per-group hazard functions" over pairs of intersectional subgroups s_i, s_j (e.g. men over 70, women between 20 - 30). In this formulation, fairness for intersectional subgroups provably guarantees fairness for the higher-level groups (Foulds et al. 2020b).

Fair Cox Proportional Hazards Models (FCPH)

We develop simple and practical Fair Cox Proportional Hazards (FCPH) models which balance fairness and accuracy. FCPH models allow us to fair prediction of the time until a particular event occurs, for example, institutionalization into a nursing home. Therefore, the FCPH models' risk scores can be used to prioritize the "waitlist" of patients for fair allocation of healthcare resources, independent of the patient's demographics, e.g. *age*, *gender*, *race* etc.

The linear Cox model estimates the hazard function $\hat{h}(t)$ parameterized by the weight vector β. Following (Faraggi and Simon 1995; Katzman et al. 2018), the loss function to learn β can be formulated with the negative log partial likelihood of Equation 4:

$$-L_{\mathbf{X}}(\beta) = -\sum_{i:E_i=1} (\beta^\intercal \mathbf{x}_i - log \sum_{j \in \Re(T_i)} \exp(\beta^\intercal \mathbf{x}_j)) \,. \tag{9}$$

Our FCPH models are developed upon a general framework for solving fairness in linear Cox models using a penalized maximum likelihood estimation approach. The general learning objective function for our FCPH models is

$$-L_{\mathbf{X}}(\beta) + \lambda R(\beta) \,, \tag{10}$$

where $L_{\mathbf{X}}(\beta)$ is the log-likelihood for the linear Cox model, $R_{\mathbf{X}}(\beta)$ is a fairness penality, which also doubles as a regularizer, and $\lambda > 0$ is a trade-off parameter which strikes a balance between predictive accuracy and fairness. We set $R_{\mathbf{X}}(\beta)$ to F_i, F_g, and F_ϵ fairness measures to learn *individual*, *group*, and *intersectional FCPH* models, respectively. We optimize the objective function in Equation 10 via gradient descent algorithm (Ruder 2016) to learn models' parameter vector β. This approach is applicable to training deep Cox models as well, which we will study in future work.

Experiments

In this section, we validate and compare our fair Cox (FCPH) models with the typical Cox (CPH) model on the survival datasets for fair risk predictions which can be used to perform sensitive tasks such as prioritizing the waiting list for institutionalization in the healthcare programs.

Datasets

Data for the allocation of healthcare resources, e.g. prioritizing the wait list of patients for institutionalization, is not publicly available. Therefore, we validate our models on representative proxy datasets. We performed all experiments on two publicly available survival datasets:

- ***COMPAS* Data:** The *COMPAS* dataset regarding a system that is used to predict criminal recidivism, and which has been criticized as potentially biased (Angwin et al. 2016). Angwin et al. applied a Cox model to test the performance of the *COMPAS* system on offender's risk prediction and demonstrated that the system overpredicts African-American defendant's future recidivism. The company that developed COMPAS system and markets it to Law Enforcement, also used a Cox model in their validation study (Brennan, Dieterich, and Ehret 2009). Although the COMPAS system is used for bail and sentencing, it could potentially be used to allocate social work resources. Therefore, it is a useful example dataset in our study to show the effectiveness of our methods to fair risk prediction. The *COMPAS* dataset consists of $10,314$ offenders and 6 features including demographic attributes, while the task is to predict risk scores of a convicted criminal to reoffend. A total of 26.75% of subjects reoffended during the survey for data collection with a median event time of 173 days. We used binary-coded *race* (white, and African-American) and *gender* (men, and women) as protected attributes in our study.
- ***FLC* Data:** This dataset is taken from a study that investigated to which extent the serum immunoglobulin free light chain (FLC) assay can be used predict overall survival (Dispenzieri et al. 2012). Dispenzieri et al. assayed FLC levels on the patients with permission from a previous study for the prevalence of monoclonal gammopathy of undetermined significance (Kyle et al. 2006) and found that elevated FLC levels were indeed associated with higher death rates. The *FLC* Dataset consists of $7,874$ patients with 6 features such as age, gender, serum creatinine, FLC group for the patients, kappa and lambda portion for serum free light chain, while the task is to predict the risk score for death. A total of 27.55% of patients died during the survey with a median death time of $2,165$ days. We used binary-coded *age* (age≤ 65, and age> 65) and *gender* (men, and women) as protected attributes in our fairness analysis.

Experimental Settings

All the models were trained via the adaptive gradient descent optimization (Adam) algorithm (Kingma and Ba 2014) with learning rate 0.01 using PyTorch. There is no requirement of protected attributes to learn the *Individual FCPH* model since the individual fairness depends on each individual subject rather than any group/subgroup of peoples. We considered *race* for *COMPAS*, and *age* for *FLC* datasets as protected attributes in the *Group FCPH* model, while we considered all the pre-selected protected attributes (*race*, *gender* for *COMPAS*, and *age*, *gender* for *FLC* datasets) in the *Intersectional FCPH* model.

We held out 20% of each dataset as the test set, using the remainder for training. We further held out 20% from each training dataset as the development set for each dataset. Since it is challenging to estimate group and intersectional fairness reliably on mini-batches due to data sparsity (Foulds et al. 2020a), we trained all the models, except the *Individual FCPH* model, in a batch setting for 500 iterations. It becomes very expensive to measure individual fairness on the whole training set in each iteration when training *Individual FCPH* models in a batch setting. Furthermore, we found that data sparsity is not a serious issue for individual fairness measures, unlike group and intersectional fairness. Therefore, to address the bottleneck we trained the *Individual FCPH* models in a mini-batch setting for 50 epochs with a mini-batch size of 128.

Evaluation Protocols

In addition to the fairness measures we proposed (Equations 5, 7, and 8), in our evaluation we also included traditional accuracy measures for the predictive performance of the survival models: *Concordance Index (C-index)*, *Brier Score*, *Time-dependent AUC*, and *Log Partial Likelihood*. The *C-index* (Raykar et al. 2007; H. Uno 2011; Brentnall and Cuzick 2018) is a rank order statistic for predictions against true outcomes, thus highly relevant for waitlists. It is a generalization of the Area Under the ROC Curve (AUC) for continuous response and censored data that reflects a measure of how well a model predicts the ordering of individual's event time. The *C-index* is based on the assumption that patients who lived longer should have been assigned a lower risk than patients who lived less long (Harrell et al. 1982). A high *C-index* means that there is a high likelihood that for two random samples, the order of their predicted response matches the order of their observed response.

The *Brier Score* (E. Graf and Schumacher 1999) measures the accuracy of probabilistic predictions. Given a set of N predictions, the empirical Brier Score measures the weighted mean squared difference between the predicted probability assigned to possible outcomes for sample i and the actual outcome. The weight for sample i can be estimated by considering the Kaplan-Meier estimator (Kaplan and Meier 1958) of the censoring distribution on the dataset.

The *Time-dependent AUC* (Chambless and Diao 2006) is a function of time that extends the ROC curve to continuous outcomes, in particular survival time, assuming a subject's event status is typically not fixed and changes over time, e.g. patients who are disease-free earlier may develop the disease later due to longer study follow-up. It reflects the area under the cumulative/dynamic ROC at time t to determine how well a model can distinguish subjects that experienced an event prior to or at time t (cumulative cases) from sub-

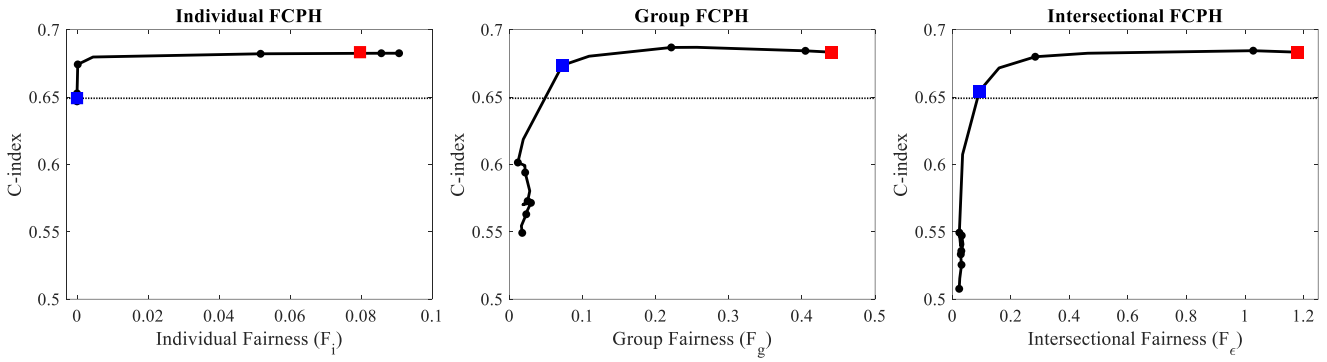

Figure 1: Fairness and accuracy trade-off plots for the development set of the *COMPAS* dataset. Shows the impact of tuning parameter λ on the FCPH models' C-index and corresponding fairness measures. *Black* circles correspond to different λ values (larger to smaller from left to right), while *blue* square indicates the selected FCPH model for a specific λ value. *Red* square: typical CPH model without fairness penalty. *Dotted* line represents 5% degraded *C-index* from the typical CPH model. C-index: higher is better. Fairness measures: lower is better.

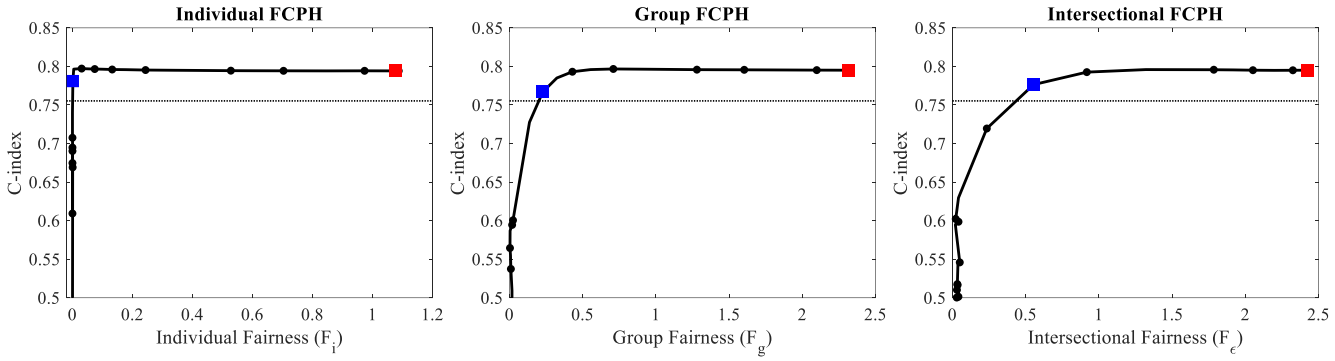

Figure 2: Fairness and accuracy trade-off plots for the development set of the *FLC* dataset. Shows the impact of tuning parameter λ on the FCPH models' C-index and corresponding fairness measures. *Black* circles correspond to different λ values (larger to smaller from left to right), while *blue* square indicates the selected FCPH model for a specific λ value. *Red* square: typical CPH model without fairness penalty. *Dotted* line represents 5% degraded *C-index* from the typical CPH model. C-index: higher is better. Fairness measures: lower is better.

jects that experienced an event after this time point (dynamic controls).

Finally, we used the *Log Partial Likelihood* from Equation 9 (dropping the minus sign) as a performance measure. The effect of the covariates can be estimated using the *Log Partial Likelihood* without the need to model the change of the hazard over time and it measures the goodness of fit of models to a sample of data for learned parameter β.

Trade-off Between Fairness and Accuracy

Since fairness may hurt accuracy because it diverts a system's learning objective from accuracy only to both accuracy and fairness, we will assess each proposed model based on this trade-off. Figure 1 and Figure 2 show the fairness and accuracy trade-off plots for the FCPH models on the development set of the *COMPAS* and *FLC* datasets, respectively. The *C-index* is selected as the accuracy-based performance measure in this experiment. The impact of the tuning parameter λ on the C-index and corresponding fairness measures for the proposed FCPH models are demonstrated in these figures. Larger λ values allow us to learn more fair, but less accurate FCPH models, while smaller λ values have the opposite impact on the FCPH models.

The tuning parameter λ needs to be chosen as a trade-off between the *C-index* and fairness. We chose λ for all FCPH models via grid search on the development set based on a pre-defined rule: *select the λ that provides the fairest (under the corresponding fairness metric, e.g. F_i, F_g, and F_ϵ for* individual, group, *and* intersectional FCPH *models, respectively) Cox model on the development set, allowing up to* 5% *degradation in* C-index *from the typical CPH model*. In Figure 1 and Figure 2, the *red* square represents the typical CPH model without fairness penalty and the *black* circles (correspond to different λ values) above the *dotted* line are FCPH models that degrade the *C-index* but not over 5%. Finally, *blue* squares indicate the selected fairest FCPH model for a specific λ value that complies with the pre-defined rule.

Performance for FCPH Models

We evaluated the performance for FCPH models on the test data in terms of accuracy-based performance measures and proposed fairness measures, and compare our fair models with the typical CPH model. The goals of our experiments were to demonstrate the practicality of our FCPH models.

	Models	Typical CPH	Individual FCPH	Group FCPH	Intersectional FCPH
Tuning	λ	X	25	1	1
Performance	C-index	**0.6648**	0.6341	0.6577	0.6445
	Brier Score	0.1877	0.1823	0.1808	**0.1787**
	Time-dependent AUC	0.6872	0.6584	**0.6890**	0.6745
	Log Partial Likelihood	**-7.0954**	-7.1664	-7.1167	-7.1542
Fairness	F_i-Fairness	0.0968	**0**	0.0090	0.0004
	F_g-Fairness	0.4198	0.1999	0.0765	**0.0418**
	F_ϵ-Fairness	1.0821	0.7291	0.2421	**0.1067**

Table 1: Comparison of FCPH models with typical CPH model on the *COMPAS* dataset. *Higher is better for accuracy-based performance measures*; *lower is better for fairness measures*. FCPH models outperform typical CPH model in terms of all fairness measures.

	Models	Typical CPH	Individual FCPH	Group FCPH	Intersectional FCPH
Tuning	λ	X	5	0.7	0.4
Performance	C-index	**0.8030**	0.7898	0.7768	0.7885
	Brier Score	0.2244	**0.1911**	0.1951	0.1976
	Time-dependent AUC	0.8015	**0.8173**	0.8147	0.8155
	Log Partial Likelihood	**-6.3737**	-6.9207	-6.7963	-6.7299
Fairness	F_i-Fairness	1.4073	**0.0009**	0.0243	0.0343
	F_g-Fairness	3.0027	**0.1466**	0.2879	0.3959
	F_ϵ-Fairness	2.8334	**0.3761**	0.7468	0.6610

Table 2: Comparison of FCPH models with typical CPH model on the *FLC* dataset. *Higher is better for accuracy-based performance measures*; *lower is better for fairness measures*. FCPH models outperform typical CPH model in terms of all fairness measures.

In Table 1 and Table 2, we show detailed results for *COMPAS* and *FLC* datasets, respectively. All FCPH models outperform typical CPH models in terms of all three fairness measures for both datasets. Note that the λ values for FCPH models in Table 1 and Table 2 represent the *blue* square from Figure 1 and Figure 2, respectively. In the *COMPAS* dataset, the *Individual FCPH* model was the most fair model in terms of F_i measures, while *intersectional FCPH* model was the most fair model in terms of F_g and F_ϵ measures. The *Individual FCPH* model shows superior performance on the *FLC* dataset that outperforms all models in terms of all three fairness measures. The *Group FCPH* model gives middling performance, and cannot win over the *Individual* and *intersectional FCPH* models in terms of fairness. This is presumably due to the fact that ensuring fairness for individuals or intersectional subgroups imposes a harder constraint to the objective function that automatically ensures fairness for groups. As expected, typical CPH performed best in terms of *C-index* and *Log Partial Likelihood*. However, surprisingly, we found that FCPH models also outperform typical CPH models in accuracy-based performance measures such as *Brier Score* and *AUC* (see *group* and *intersectional FCPH* models in Table 1, and *individual FCPH* model in Table 2).

Do Fair Models Reduce Overfitting?

The improved performance of FCPH models over typical CPH models is counter-intuitive. We further study this result in this section by comparing the generalization of the models. Table 3 and Table 4 compare the accuracy-based predictive performances for FCPH models with typical CPH models on the train and test set of both datasets.

The typical CPH model was the best model for both datasets in all predictive measures on the training set, but FCPH models performed better than typical CPH models on the test set for both datasets in most of the cases. In the *COMPAS* dataset, *group* and *intersectional FCPH* models were the best models on the held-out data in terms of *AUC* and *Brier Score*, respectively. Similarly, *individual FCPH* models showed the best *Brier Score* and *AUC* measures on the held-out *FLC* data. We also found that FCPH models decrease the corresponding gap between accuracy-based predictive measures on train and test data due to the regularization behavior of the fairness constraints. Thus, FCPH models reduce overfitting of the typical Cox model to some extent.

Discussion and Future Work

In this work, we investigated fairness for survival models and developed methods to ensure fair risk scores. Balance between accuracy and fairness is an important decision to make prior to learning fair models, and this depends on the stakeholders. To validate our proposed fair models, we performed all the experiments on two public proxy datasets.

In future, we plan to apply our proposed methods on the data from healthcare programs such as Medicaid to ensure fair risk prediction in ranking the waitlist for individuals to receive care. We further plan to study the impact of an intervention to the prioritization process on each individual which determines the waiting time to receive home and community-based healthcare services. The successful application and deployment of our methods to the healthcare program is the eventual goal of this research.

Models	Train Set			Test Set		
	C-index	Brier Score	AUC	C-index	Brier Score	AUC
Typical CPH	**0.6859**	**0.1530**	**0.7178**	**0.6648**	0.1877	0.6872
Individual FCPH	0.6609	0.1646	0.6962	0.6341	0.1823	0.6584
Group FCPH	0.6654	0.1610	0.7031	0.6577	0.1808	**0.6890**
Intersectional FCPH	0.6593	0.1648	0.6964	0.6445	**0.1787**	0.6745

Table 3: Comparison of the accuracy-based predictive performances for typical CPH and FCPH models on the train and test set of the *COMPAS* dataset. FCPH models reduce overfitting.

Models	Train Set			Test Set		
	C-index	Brier Score	AUC	C-index	Brier Score	AUC
Typical CPH	**0.7944**	**0.1298**	**0.8149**	**0.8030**	0.2244	0.8015
Individual FCPH	0.7740	0.1719	0.8089	0.7898	**0.1911**	**0.8173**
Group FCPH	0.7581	0.1613	0.7961	0.7768	0.1951	0.8147
Intersectional FCPH	0.7680	0.1555	0.8012	0.7885	0.1976	0.8155

Table 4: Comparison of the accuracy-based predictive performances for typical CPH and FCPH models on the train and test set of the *FLC* dataset. FCPH models reduce overfitting.

Conclusion

We developed three fairness definitions for survival models and corresponding learning algorithms to ensure equitable allocation of healthcare resources. In extensive experiments on publicly available datasets, we demonstrated that our methods are practical and effective. The proposed methods for fair prioritization of healthcare have the potential to prevent avoidable institutionalization of elderly and disabled individuals, thereby improving quality of life and saving taxpayer dollars, while ensuring fair and equitable allocation.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No.'s IIS1927486; IIS1850023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology.

References

[Angwin et al. 2016] Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* 23.

[Barocas and Selbst 2016] Barocas, S., and Selbst, A. 2016. Big data's disparate impact. *Cal. L. Rev.* 104:671.

[Berk et al. 2017] Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *FAT/ML Workshop*.

[Berk et al. 2018] Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2018. Fairness in criminal justice risk assessments: The state of the art. *In Sociological Methods and Research* 1050:28.

[Bolukbasi et al. 2016] Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in NeurIPS*.

[Brennan, Dieterich, and Ehret 2009] Brennan, T.; Dieterich, W.; and Ehret, B. 2009. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior* 36(1):21–40.

[Brentnall and Cuzick 2018] Brentnall, A., and Cuzick, J. 2018. Use of the concordance index for predictors of censored survival data. *Statistical Methods in Medical Research* 27(8):2359–2373.

[Buolamwini and Gebru 2018] Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**, 77–91.

[Calders and Verwer 2010] Calders, T., and Verwer, S. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

[Chambless and Diao 2006] 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.

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

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

[Crenshaw 1989] Crenshaw, K. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.* 139–167.

[Dispenzieri et al. 2012] Dispenzieri, A.; Katzmann, J. A.; Kyle, R. A.; Larson, D. R.; Therneau, T. M.; Colby, C. L.; Clark, R. J.; Mead, G. P.; Kumar, S.; Melton III, L. J.; et al.

2012. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, 517–523. Elsevier.

[Dwork et al. 2012] 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.

[E. Graf and Schumacher 1999] E. Graf, C. Schmoor, W. S., and Schumacher, M. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18(17-18):2529–2545.

[Faraggi and Simon 1995] Faraggi, D., and Simon, R. 1995. A neural network model for survival data. *Statistics in medicine* 14(1):73–82.

[Feldman et al. 2015] Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

[Ferryman and Pitcan 2018] Ferryman, K., and Pitcan, M. 2018. Fairness in precision medicine. *Data & Society*.

[Foulds et al. 2020a] Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020a. Bayesian modeling of intersectional fairness: The variance of bias. In *Proceedings of the SIAM International Conference on Data Mining*.

[Foulds et al. 2020b] Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020b. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1918–1921. IEEE.

[Goel, Yaghini, and Faltings 2018] Goel, N.; Yaghini, M.; and Faltings, B. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA*.

[H. Uno 2011] H. Uno, T. Cai, M. P. R. D. L. W. 2011. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* 10(30):1105–1117.

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

[Harrell et al. 1982] 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.

[Hebert-Johnson et al. 2018] Hebert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Dy, J., and Krause, A., eds., *Proceedings of the 35th ICML, PMLR 80*, 1944–1953.

[Huang et al. 2019] Huang, Z.; Zhan, X.; Xiang, S.; Johnson, T. S.; Helm, B.; Yu, C. Y.; Zhang, J.; Salama, P.; Rizkalla, M.; Han, Z.; et al. 2019. Salmon: Survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in genetics* 10:166.

[Kaplan and Meier 1958] Kaplan, E., and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Statistics in Medicine* 53(282):457–481.

[Katzman et al. 2018] 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):24.

[Kearns et al. 2018] Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2564–2572.

[Kilbertus et al. 2017] Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

[Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Kusner et al. 2017] Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *NeurIPS*.

[Kyle et al. 2006] Kyle, R. A.; Therneau, T. M.; Rajkumar, S. V.; Larson, D. R.; Plevak, M. F.; Offord, J. R.; Dispenzieri, A.; Katzmann, J. A.; and Melton III, L. J. 2006. Prevalence of monoclonal gammopathy of undetermined significance. *New England Journal of Medicine* 354(13):1362–1369.

[Lee et al. 2018] Lee, C.; Zame, W. R.; Yoon, J.; and van der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Munoz, Smith, and Patil 2016] Munoz, C.; Smith, M.; and Patil, D. 2016. *Big data: A report on algorithmic systems, opportunity, and civil rights*. Exec. Office of the President.

[Noble 2018] Noble, S. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[Raykar et al. 2007] Raykar, V. C.; Steck, H.; Krishnapuram, B.; Dehing-Oberije, C.; and Lambin, P. 2007. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 1209–1216.

[Ruder 2016] Ruder, S. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[Zafar et al. 2017a] Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 1171–1180.

[Zafar et al. 2017b] Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. 962–970.

[Zhao et al. 2017] Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of EMNLP*.