https://doi.org/10.1137/1.9781611976700.22

Equitable Allocation of Healthcare Resources with Fair Survival Models[*]

Kamrun Naher Keya[†] Rashidul Islam[†] Shimei Pan[†] Ian Stockwell[‡] James Foulds[†]

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, and does not reinforce harmful systemic bias. We develop multiple fairness definitions and corresponding fair learning algorithms for survival models to ensure equitable allocation of healthcare resources. We demonstrate the utility of our methods in terms of fairness and predictive accuracy on three publicly available survival datasets.

1 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,and limit the enrollment of each program to contain costs. Many waivers are built to serve older adults or individuals with developmental/physical disabilities better by keeping them out of institutional settings, e.g. 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 "first in, first out," 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 [2], 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 substantial research on algorithmic bias [10, 18, 15] which aims to learn non-discriminatory predictive models by enforcing constraints in the training phase [3, 28, 16].

Like any data that involves individuals from different demographics, health data is subject to bias, and the expanding amount and types of data that are accessible today can make it difficult to distinguish where bias can emerge [13]. 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) [7] model is particularly

[*] 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.

[†] Department of Information Systems, UMBC, Baltimore, USA

[‡] The Hilltop Institute, UMBC, Baltimore, USA

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

appropriate, as the multiplicative relationship between covariates and risk aids explainability. Though the AI fairness community has proposed various fairness definitions [10, 18, 15] 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 for linear and nonlinear models. The models' risk scores can then be used to fairly prioritize the Medicaid waitlist. This paper extends our preliminary research, accepted at a non-archival symposium [23].

To the best of our knowledge, this is the first investigation on fairness for survival models to ensure equitable allocation of healthcare resources. The main contributions of this work include:

- We extend three types of fairness definitions to measure bias in the survival analysis problem.
- We develop fair learning algorithms for linear CPH models. We then extend our method to fair deep learning algorithms for nonlinear CPH models.
- We perform extensive experiments validating our models with regard to both fairness and accuracy on three publicly available survival datasets.

2 Background and Related Work

In this section, we describe survival data, the Cox proportional hazards model, and fairness in AI.

2.1 Survival Data Survival data [21, 25] 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. 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, and the individual is said to be *right-censored*. In survival analysis, right-censored data is important and requires special consideration, as it cannot simply be ignored without introducing substantial bias.

2.2 Cox Proportional Hazards Model (CPH) The Cox proportional hazards (CPH) model [7] is the most widely used model for survival analysis. 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 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 continuous time t. The hazard function is defined as

$$h(t) \triangleq \lim_{\Delta t \to 0^+} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \ . \tag{2.1}$$

The CPH model specifies the hazard function via

$$h(t) = h_0(t) \exp(\beta^\intercal \mathbf{x}) \ , \tag{2.2}$$

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 then determined as

$$S(x) = \exp(-H(t)), H(t) = \int_0^t h(u) du \ . \tag{2.3}$$

To perform Cox regression, β can be learned by optimizing the Cox partial likelihood [12, 21]. 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

$$L_c(\beta) = \prod_{i:E_i=1} \frac{exp(\beta^\intercal \mathbf{x}_i)}{\sum_{j \in \Re(T_i)} \exp(\beta^\intercal \mathbf{x}_j)} \ , \tag{2.4}$$

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 [21], recently deep neural networks [21, 25] have been applied to CPH models to solve the problem of nonlinear survival analysis.

2.3 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 [2, 26]. A number of studies have subsequently demonstrated that bias and fairness issues in AI are both harmful and pervasive [1, 3, 4]. The AI community has responded by developing a broad array of mathematical formulations of fairness and learning algorithms which aim to satisfy them [10, 18, 28, 20].

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** [10] definitions aim to ensure that *similar individuals obtain similar outcomes* under the algorithm in question. **Group fairness** [10] definitions aim to preserve fairness at the level of groups of individuals, e.g. women, the elderly, or African Americans. Finally, **intersectional fairness** [15] definitions are those for which fairness is to be ensured for a specified set of subgroups defined by the protected attributes. 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 [27, 5, 15].

The Cox model was previously applied to the problem of detecting racial bias in criminal recidivism prediction [1]. However, there is no prior work that enforces fairness definitions to peform fair survival analysis.

3 Methods

In this section, we describe our methodology to ensure fair risk predictions with survival models. We extend the three main types of AI fairness to survival analysis and develop simple fair learning algorithms for them.

3.1 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 [10] aims to ensure that a model produces similar outcomes to similar individuals. In the context of survival models, we define individual fairness (F_i) as follows:

$$F_i = \sum_{i=1}^{N^{(test)}} \sum_{j=i+1}^{N^{(test)}} \max(0, |\bar{h}_\beta(\mathbf{x}_i) - \bar{h}_\beta(\mathbf{x}_j)| - D(\mathbf{x}_i, \mathbf{x}_j)) \ , \tag{3.5}$$

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 $|\bar{h}_\beta(\mathbf{x}_i) - \bar{h}_\beta(\mathbf{x}_j)|$, which can be defined as $D(x_i, x_j) = C\sqrt{\sum_{k=1}^{n}(x_{i_k} - x_{j_k})^2}$ for a Euclidean n-space with scale factor C. 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 [10], 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

$$F_g = \max_{a \in A} |E[\bar{h}_\beta(a)] - E[\bar{h}_\beta(\mathbf{x})]| \ , \tag{3.6}$$

$$\bar{h}_\beta(a) \triangleq \int_{\mathbf{x}} \exp(\beta^\intercal \mathbf{x}) p(\mathbf{x}|a) \ , \tag{3.7}$$

the worst-case deviation of the *per-group expected hazard function* $E[\bar{h}_\beta(a)]$ 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 [22, 15] 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 [8], 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. 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 [15], intersectional fairness (F_ϵ) for survival models can be extended as

$$F_\epsilon = \max_{s_i \in A, s_j \in A} |\log E[\bar{h}_\beta(s_i)] - \log E[\bar{h}_\beta(s_j)]| \ , \tag{3.8}$$

a worst case of log-ratios of "expected per-group hazard functions" over pairs of intersectional subgroups s_i, s_j (e.g. men over 70, women between 20 - 30). With this formulation, a direct application of Theorem IV.1 of [15] shows that fairness F_ϵ for intersectional subgroups provably guarantees the same degree of fairness F_ϵ for the higher-level groups. E.g., protecting fairness at the intersection of *gender* and *race* (*Black women*, . . .) ensures the same fairness for *gender* (*women, men*).

3.2 Fair Survival Models

We first develop simple and practical Fair linear Cox Proportional Hazards (FCPH) models which balance fairness and accuracy. However, in many applications we cannot assume the survival data satisfies the linear proportional hazards condition as the number of features and interactions increases. Therefore, we extend our approach to develop Fair Deep Cox Proportional Hazards (FDCPH) models based on deep neural networks. The FCPH and FDCPH models enable fair prediction of the time until a particular event occurs, for example, institutionalization into a nursing home. The fair models' risk scores can then be used to prioritize the "waitlist" of patients for fair allocation of healthcare resources.

3.2.1 Learning Algorithms for FCPH

The linear Cox model estimates the hazard function $\hat{h}_\beta(x)$ parameterized by the weight vector β. Following [12, 21], the loss function to learn β can be formulated as the negative log partial likelihood of Equation 2.4:

$$L_{\mathbf{X}}(\beta) = -\sum_{i:E_i=1} (\beta^\intercal \mathbf{x}_i - log \sum_{j \in \Re(T_i)} \exp(\beta^\intercal \mathbf{x}_j)) \ . \tag{3.9}$$

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 $g(\beta)$ is

$$g(\beta) = -(L_{\mathbf{X}}(\beta) + \lambda F_{\mathbf{X}}(\beta)) \ , \tag{3.10}$$

where $L_{\mathbf{X}}(\beta)$ is the log-likelihood for the linear Cox model, $F_{\mathbf{X}}(\beta)$ is a fairness penalty, 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 $F_{\mathbf{X}}(\beta)$ 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 3.10 using Adam via backpropagation (*BP*) and automatic differentiation (*autodif*).

3.2.2 Learning Algorithms for FDCPH

Our FDCPH models are built on the deep neural network-based survival model DeepSurv [21]. The output of the FDCPH model is a single node that estimates the hazard function $\hat{h}_\theta(x)$, where θ denote weights and intercept terms of the deep neural network.

Like FCPH, the loss function of FDCPH models to learn θ can be formulated with the negative log partial likelihood and the corresponding fairness penalty:

$$\begin{gathered} L_{\mathbf{X}}(\theta) = -\sum_{i:E_i=1} (\hat{h}_\theta(\mathbf{x}_i) - log \sum_{j \in \Re(T_i)} \exp(\hat{h}_\theta(\mathbf{x}_j)) \ , \\ g(\theta) = -(L_{\mathbf{X}}(\theta) + \lambda F_{\mathbf{X}}(\theta)) \ , \end{gathered} \tag{3.11}$$

where $g(\theta)$ is the objective function, $L_{\mathbf{X}}(\theta)$ is the negative log-likelihood for the DeepSurv model, and $F_{\mathbf{X}}(\theta)$ is a fairness penality on the model parameters θ. We again set $F_{\mathbf{X}}(\theta)$ to our proposed F_i, F_g, and F_ϵ fairness measures to learn *Individual*, *Group*, and *Intersectional FDCPH* models, respectively.

Following [21], our FDCPH models are designed using a deep architecture, i.e. more than one hidden layer, along with $l2$ regularization, dropout, gradient clipping, and nonlinear activation functions such as rectifier (ReLU), Leaky ReLU, or scaled exponential linear units (SELU), etc. We once again use Adam via *BP* and *autodif* to optimize the FDCP models' objective in Equation 3.11.

4 Experiments

In this section, we validate and compare our fair survival models FCPH and FDCPH with the typical survival models (without any fairness penalty). Our implementation's source code is available online.[2]

4.1 Datasets

Data for the allocation of healthcare resources, e.g. prioritizing the wait list of patients for healthcare, is not publicly available. Therefore, we validate our models on three representative publicly available datasets as proxies for the sensitive data:

- ***COMPAS* Dataset:** The *COMPAS* dataset regarding a system that is used to predict criminal recidivism, and which has been criticized as potentially biased [1]. 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* Dataset:** 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 [9]. 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$\leq$ 65, and age$>$ 65) and *gender* (men, and women) as protected attributes in our fairness analysis.

- ***SUPPORT* Dataset:** Data from a large study to understand prognoses preferences outcomes and risks of treatment (SUPPORT) [24] which analyzed the survival time of seriously ill hospitalized patients. SUPPORT data contains $9,105$ patients and 14 features such as presence of diabetes, presence of dementia, presence of cancer, mean arterial blood pressure, heart rate including protected

[2] Example code implementing fair survival models can be found at `https://github.com/kkeya1/FCPH`

attributes, i.e. *age*, *gender*, and *race*. During the study, 68.10% of patients died with a median death time of 58 days. In this work, we used binary encoded *age* (age ≤ 65 and age > 65), *gender* (men and women), and *race* (white and non-white).

4.2 Experimental Settings We compare our proposed fairness approach with several baseline models which do not incorporate any fairness penalty in the loss function. Typical CPH [7] was used as a baseline for linear survival models, while we also used nonlinear survival models such as DeepSurv [21] and Random Survival Forest (RSF) [19] as baseline models. RSF is a meta estimator that fits a number of survival trees on various sub-samples of the dataset for the analysis of right-censored survival data.

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 [14], we trained all the fair survival models, except *Individual FCPH and FDCPH* models, 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 and FDCPH* models in the batch setting. Furthermore, we found that data sparsity is not a serious issue for individual fairness measures, unlike group and intersectional fairness. To address this problem, we trained the *Individual FCPH and FDCPH* models in the mini-batch setting for 50 epochs with a mini-batch size of 128. We found that the *Individual FDCPH* model suffers from the exploding gradients problem which we addressed using gradient clipping with a clip value of 5.

All the linear models (Typical CPH and FCPH) were trained via Adam optimizer with learning rate 0.01 using PyTorch. We selected the hyper-parameters for DeepSurv via grid search on the development set for each dataset (see Appendix for more details on the hyper-parameters). For FDCPH models, we only selected the fairness specific tuning parameters via grid search (details on the Section 4.4), while other hyper-parameters were set to DeepSurv's hyper-parameters.

We considered *race* for *COMPAS*, and *age* for *FLC* and *SUPPORT* datasets as protected attributes in the *Group FCPH and FDCPH* models, while we considered all the protected attributes (*race*, *gender* for *COMPAS*, *age*, *gender* for *FLC*, and *age*, *gender*, *race* for *SUPPORT* datasets) in the *Intersectional FCPH and FDCPH* models. There is no requirement of protected attributes to learn the *Individual FCPH and FDCPH* models since the F_i metric depends on each individual subject rather than any group/subgroup.

4.3 Evaluation Protocols In addition to the fairness measures we proposed (Equations 3.5, 3.7, and 3.8), in our evaluation we also included traditional accuracy measures for the predictive performance of the survival models: *Concordance Index (C-index)*, *Brier Score*, *AUC*, and *Log Partial Likelihood*.

The *C-index* [17] is a rank order statistic for predictions against true outcomes, thus highly relevant for waitlists. It is based on the assumption that patients who lived longer should have been assigned a lower risk than patients who lived less long. The *Brier Score* [11] 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 time-dependent AUC [6] 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 subjects that experienced an event after this time point (dynamic controls). Finally, *Log Partial Likelihood (LPL)* is used from Equation 3.9 (dropping the minus sign) as a performance measure. The effect of the covariates can be estimated using LPL 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 model parameters.

4.4 Tuning for Fairness In this section, we discuss necessary tuning approaches for fairness specific hyper-parameters to learn fair survival models.

4.4.1 Sensitivity of Individual Fair Models The distance metric in the individual fairness measure (Equation 3.5) needs to be in the same scale as the difference between hazard functions of two individual subjects. We study the sensitivity of the individual fair models with various scale factors C. Figure 1 shows F_i measures for various scale factors *vs.* tuning parameter λ on the development set of *FLC* data, while *Individual FDCPH* model was trained with $F_{\mathbf{X}}(\theta) = F_i$ for $C = 0.1$, $C = 0.01$, and $C = 0.001$. As shown, the tuning parameter λ can smoothly control the F_i metric for *Individual FDCPH* models with most scale factors considered. We get a similar trend in the *Individual*

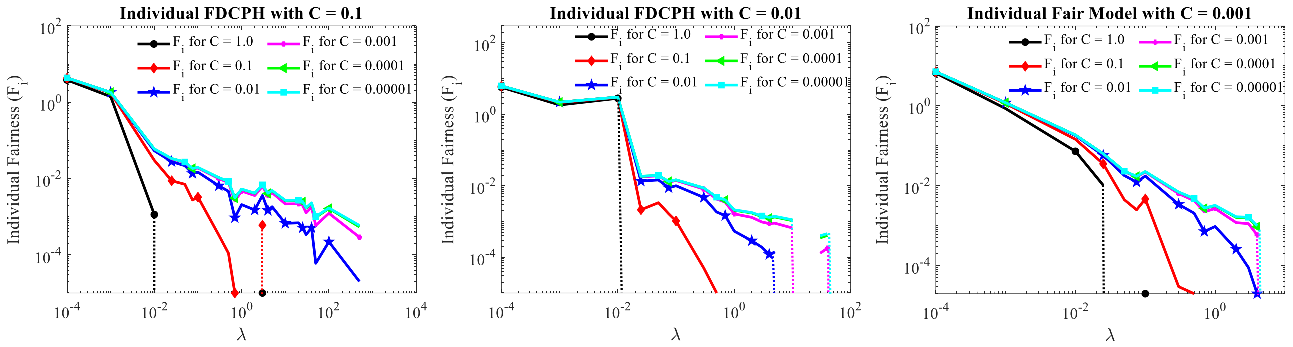

Figure 1: F_i measures for various scale factors *vs.* tuning parameter λ on the development set of *FLC* data for *Individual FDCPH* model. *Individual FDCPH* was trained with $F_{\mathbf{X}}(\theta) = F_i$ for $C = 0.1$, $C = 0.01$, and $C = 0.001$. The models are fairly insensitive to scale factors as long as λ is tuned to compensate. *Dotted lines* represent $F_i = 0$ values in the *log-scale.*

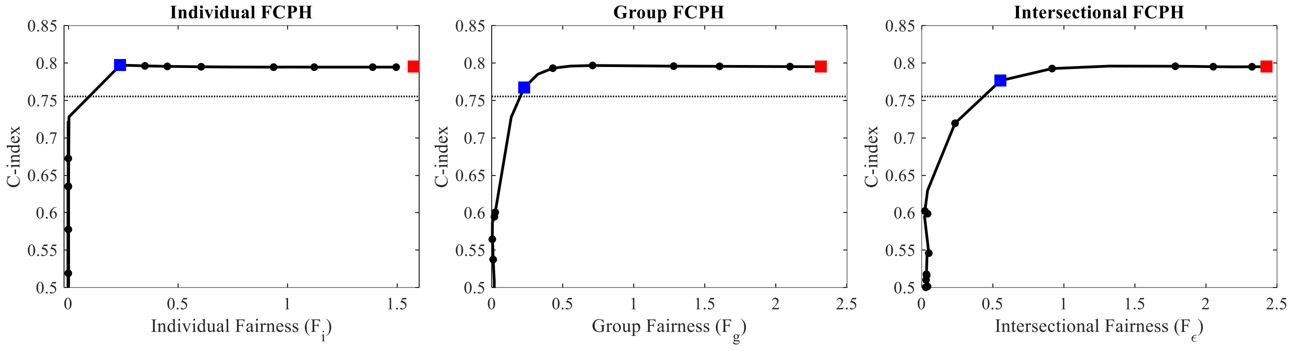

Figure 2: Fairness and accuracy trade-off plots for the development set of the *FLC* dataset. The plot shows the impact of tuning parameter λ on the FCPH models' C-index and corresponding fairness measures. *Black* circles correspond to different λ values (larger to smaller from left to right), while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

FCPH model with various scale factors (see Appendix for additional results). So, the individual fair models are fairly insensitive to scale factors as long as λ is tuned to compensate. We set $C = 0.01$ for all experiments.

4.4.2 Trade-off Between Fairness and Accuracy

AI fairness interventions may hurt accuracy because they divert a system's learning objective from accuracy only to both accuracy and fairness. We assess each proposed model based on this trade-off. Figure 2 shows the fairness and accuracy trade-off plots for the FCPH models on the development set of the *FLC* dataset. 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 fair survival models, while smaller λ values have the opposite impact on the fair 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.* Similarly, we select the λ in FDCPH models that provides the fairest deep Cox model on the development set, allowing up to 5% degradation in *C-index* from the DeepSurv model. In Figure 2, the *red* square represents the typical CPH model without fairness penalty and the *black* circles (corresponding 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

COMPAS Dataset

	Models	Performance Measures				Fairness Measures		
		C-index ↑	Brier Score ↓	AUC ↑	LPL ↑	F_i ↓	F_g ↓	F_ϵ ↓
Linear	Typical CPH	0.6648	0.1877	0.6872	-7.0954	0.4550	0.4198	1.0821
	Individual FCPH	0.5899	**0.1777**	0.6058	-7.2495	**0**	**0.0008**	0.1420
	Group FCPH	0.6577	0.1808	0.6890	-7.1167	0.2375	0.0765	0.2421
	Intersectional FCPH	0.6445	0.1787	0.6745	-7.1542	0.1560	0.0418	0.1067
Nonlinear	RSF	**0.6697**	0.1816	**0.6907**	-7.1368	0.2042	0.1640	0.3006
	DeepSurv	0.6661	0.1869	0.6865	**-7.0779**	0.0128	0.2907	0.8660
	Individual FDCPH	0.6496	0.1946	0.6699	-7.1381	**0**	0.0013	0.8535
	Group FDCPH	0.6607	0.1921	0.6870	-7.1001	**0**	0.0010	1.4588
	Intersectional FDCPH	0.6414	0.1788	0.6722	-7.1567	0.0396	0.0105	**0.1018**

FLC Dataset

	Models	Performance Measures				Fairness Measures		
		C-index ↑	Brier Score ↓	AUC ↑	LPL ↑	F_i ↓	F_g ↓	F_ϵ ↓
Linear	Typical CPH	0.8030	0.2244	0.8015	-6.3737	1.8655	3.0027	2.8334
	Individual FCPH	0.8054	0.1989	0.8099	-6.7081	0.2634	0.4439	0.8226
	Group FCPH	0.7768	**0.1951**	0.8147	-6.7963	0.2282	0.2879	0.7468
	Intersectional FCPH	0.7885	0.1976	0.8155	-6.7299	0.2835	0.3959	**0.6610**
Nonlinear	RSF	0.8031	0.2310	0.8162	-6.4904	1.7651	2.6402	1.7103
	DeepSurv	**0.8102**	0.2245	0.8150	**-6.3459**	0.0119	0.0258	2.5268
	Individual FDCPH	0.8070	0.2550	0.8122	-6.3817	**0**	0.0022	1.3574
	Group FDCPH	0.8068	0.7377	0.8111	-6.3703	**0**	**0.0009**	0.7130
	Intersectional FDCPH	0.7920	0.2060	**0.8180**	-6.6170	0.2628	0.2455	0.7487

SUPPORT Dataset

	Models	Performance Measures				Fairness Measures		
		C-index ↑	Brier Score ↓	AUC ↑	LPL ↑	F_i ↓	F_g ↓	F_ϵ ↓
Linear	Typical CPH	0.7375	0.2468	0.8066	**-6.6474**	1.1769	0.2836	0.8259
	Individual FCPH	0.7129	**0.2208**	0.7798	-6.8965	0.0578	0.0041	**0.0307**
	Group FCPH	0.7081	0.2311	0.7734	-6.7324	0.4555	0.0628	0.3772
	Intersectional FCPH	0.7176	0.2309	0.7832	-6.7307	0.3984	0.0080	0.2140
Nonlinear	RSF	0.7376	0.2928	**0.8083**	-6.6828	8.2810	1.7708	0.5334
	DeepSurv	**0.7379**	0.2345	0.8062	-6.7046	0.3355	0.0373	0.2048
	Individual FDCPH	0.7213	0.2548	0.7867	-6.6811	**0.0013**	**0.0009**	0.9606
	Group FDCPH	0.6913	0.2343	0.7483	-6.7190	0.1293	0.0036	1.0103
	Intersectional FDCPH	0.7081	0.2495	0.7734	-6.7627	0.2118	0.0040	0.0876

Table 1: Comparison of FCPH and FDCPH models with typical survival models (Typical CPH, RSF, and DeepSurv) on the *COMPAS*, *FLC*, and *SUPPORT* datasets. *Higher is better* for measures with ↑, while *lower is better* for measures with ↓. *Bold* models are our proposed approaches. FCPH and FDCPH models outperform typical survival models in terms of all fairness measures.

with the pre-defined rule. The models chosen by our procedure (*blue*) substantially improved their fairness measures with only a slight loss in *C-index*. Additional experimental results on the fairness and accuracy trade-off for the other datasets are provided in the Appendix.

4.5 Performance for Fair Survival Models We evaluated the performance for FCPH and FDCPH models on the test data in terms of accuracy-based performance measures and our proposed fairness measures, and compared our fair models with the typical baseline models: Typical CPH, DeepSurv, and RSF. The goal of our experiments was to demonstrate the practicality of our FCPH and FDCPH models. In Table 1, we show detailed results for *COMPAS*, *FLC*, and *SUPPORT* datasets. All FCPH and FDCPH models outperform typical baseline models in terms of all three fairness measures for all datasets.

In the *COMPAS* dataset, *Individual FCPH*, *Individual FDCPH*, and *Group FDCPH* were the best fair models in terms of F_i metric, while *Individual FCPH* and *Intersectional FDCPH* were the most fair models for F_g and F_ϵ measures, respectively. In the *FLC* dataset, *Individual* and *Group FDCPH* were again the best models for F_i metric, while *Group FDCPH* and *Intersectional FCPH* were the most fair models in terms of F_g and F_ϵ measures, respectively. The *Individual FCPH* and *FDCPH* models show superior performance on the *SUPPORT* dataset outperforming the other fair models, i.e. *Individual FDCPH* was the most fair model in terms of both F_i and F_g, while *Individual FCPH* was the most F_ϵ-fair model.

The *Individual* and *Intersectional* fair models consistently provided better fairness overall on all datasets in terms of all three fairness metrics. This is presumably due to the fact that ensuring fairness for individuals or

COMPAS Dataset						
Models		Train Set			Test Set	
	C-index ↑	Brier Score ↓	AUC ↑	C-index ↑	Brier Score ↓	AUC ↑
DeepSurv	**0.6944**	**0.1516**	**0.7339**	**0.6661**	0.1869	0.6865
Individual FDCPH	0.6633	0.1598	0.6987	0.6496	0.1946	0.6699
Group FDCPH	0.6917	0.1610	0.7310	0.6607	0.1921	**0.6870**
Intersectional FDCPH	0.6496	0.1658	0.6865	0.6414	**0.1788**	0.6722

FLC Dataset						
Models		Train Set			Test Set	
	C-index ↑	Brier Score ↓	AUC ↑	C-index ↑	Brier Score ↓	AUC ↑
DeepSurv	**0.8016**	**0.1261**	**0.8325**	**0.8102**	0.2245	0.8150
Individual FDCPH	0.7955	0.1292	0.8245	0.8070	0.2550	0.8122
Group FDCPH	0.8001	0.7452	0.8302	0.8068	0.7377	0.8111
Intersectional FDCPH	0.7711	0.1418	0.8054	0.7920	**0.2060**	**0.8180**

SUPPORT Dataset						
Models		Train Set			Test Set	
	C-index ↑	Brier Score ↓	AUC ↑	C-index ↑	Brier Score ↓	AUC ↑
DeepSurv	**0.7368**	**0.1638**	**0.8028**	**0.7379**	0.2345	**0.8062**
Individual FDCPH	0.7284	0.1696	0.7939	0.7213	0.2548	0.7867
Group FDCPH	0.7081	0.1713	0.7664	0.6913	**0.2343**	0.7483
Intersectional FDCPH	0.7033	0.1801	0.7716	0.7081	0.2495	0.7734

Table 2: Comparison of the accuracy-based predictive performances for *DeepSurv* and *FDCPH* models on the train and test set of the *COMPAS*, *FLC*, and *SUPPORT* datasets. *Higher is better* for measures with ↑, while *lower is better* for measures with ↓. *FDCPH* models reduce overfitting.

intersectional subgroups imposes a harder constraint to the objective function that automatically ensures fairness for groups. As expected, typical survival models performed best in terms of accuracy-based performance measures. For example, the *C-index* of the *DeepSurv* model is the highest on the *FLC* and *SUPPORT* datasets. *RSF* showed the highest *C-index* and *AUC* on the *COMPAS* data, while the *LPL* of the *Typical CPH* model was the highest on the *SUPPORT* dataset. However, surprisingly, we found that our fair models also outperformed typical survival models in some accuracy-based performance measures, including the *Brier Score* for all three datasets and *AUC* in few cases.

4.6 Do Fair Models Reduce Overfitting? The improved performance of the fair models over typical models is counter-intuitive. We further study this result in this section by comparing the generalization of the models. Table 2 compares the accuracy-based predictive performances for *FDCPH* models with *DeepSurv* model on the train and test set of all datasets.

The *DeepSurv* was the best model for all three datasets in all predictive measures on the training set, but *FDCPH* models performed better than *DeepSurv* on the test set in most of the cases. In the *COMPAS* dataset, *Group* and *Intersectional FDCPH* models were the best models on the held-out data in terms of *AUC* and *Brier Score*, respectively. The *Intersectional FDCPH* model showed the best *Brier Score* and *AUC* measures on the held-out *FLC* data. Finally, the *Group FDCPH* model provided the best *Brier Score* on the test set of the *SUPPORT* data. We see a similar trend in the performances for *FCPH* and *Typical CPH* models on the train and test set of all datasets (see Appendix for details). We also found that our fair models decrease the corresponding gap between accuracy-based predictive measures on the train and test data due to the regularization behavior of the fairness constraints. Thus, fair survival models reduce overfitting of the typical survival models to some extent.

5 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 when deploying fair models, which depends on the stakeholders. To validate our proposed models, we performed experiments on three public proxy datasets, with promising fairness/accuracy results.

The eventual goal of this research is the successful application and deployment of our methods to the fair needs-based ranking of the individuals waiting to receive Medicaid care in the state of Maryland. In future, we plan to study our proposed methods on data from the Maryland Department of Health. 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, and to study how stakeholders can be included in fair AI decision-making processes.

6 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 of resources.

References

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

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

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

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

[5] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292, 2010.

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

[7] D.R. Cox. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*, 34(2):187–202, 1972.

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

[9] A. Dispenzieri, J.A. Katzmann, R.A. Kyle, D.R. Larson, T.M. Therneau, C.L. Colby, R.J. Clark, G.P. Mead, S. Kumar, L.J. Melton III, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clin. Proc.*, volume 87, pages 517–523. Elsevier, 2012.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *ITCS*, pages 214–226, 2012.

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

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

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

[14] J.R. Foulds, R. Islam, K.N. Keya, and S. Pan. Bayesian modeling of intersectional fairness: The variance of bias. In *SDM*, 2020.

[15] J.R. Foulds, R. Islam, K.N. Keya, and S. Pan. An intersectional definition of fairness. In *ICDE*, pages 1918–1921. IEEE, 2020.

[16] N. Goel, M. Yaghini, and B. Faltings. Non-discriminatory machine learning through convex fairness criteria. In *AAAI*, 2018.

[17] M. Pencina R. D'Agostino L. Wei H. Uno, T. Cai. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*, 10(30):1105–1117, 2011.

[18] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in NeurIPS*, pages 3315–3323, 2016.

[19] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, et al. Random survival forests. *Ann. Appl. Stat.*, 2(3):841–860, 2008.

[20] R. Islam, K.N. Keya, S. Pan, and J. Foulds. Mitigating demographic biases in social media-based recommender systems. *KDD (Social Impact Track)*, 2019.

[21] 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 Med. Res. Methodol.*, 18(1):24, 2018.

[22] M. Kearns, S. Neel, A. Roth, and Z.S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*, pages 2564–2572, 2018.

[23] K.N. Keya, R. Islam, S. Pan, I. Stockwell, and J.R. Foulds. Equitable allocation of healthcare resources with fair Cox models. In *AAAI FSS on AI in Government and Public Sector*, 2020.

[24] W.A. Knaus, F.E. Harrell, J. Lynn, L. Goldman, R.S. Phillips, A.F. Connors, N.V. Dawson, W.J. Fulkerson, R.M. Califf, N. Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Ann. Intern. Med.*, 122(3):191–203, 1995.

[25] C. Lee, W.R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI*, 2018.

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

[27] M.B. Zafar, I. Valera, M.G. Rogriguez, and K.P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, pages 962–970. PMLR, 2017.

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