The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Fair Federated Survival Analysis

Md Mahmudur Rahman, Sanjay Purushotham

University of Maryland, Baltimore County, Baltimore, Maryland, USA
mrahman6@umbc.edu, psanjay@umbc.edu

Abstract

Federated Survival Analysis (FSA) is an emerging Federated Learning (FL) paradigm that enables training survival models on decentralized data while preserving privacy. However, existing FSA approaches largely overlook the potential risk of bias in predictions arising from demographic and censoring disparities across clients' datasets, which impacts the fairness and performance of federated survival models, especially for underrepresented groups. To address this gap, we introduce `FairFSA`, a novel FSA framework that adapts existing fair survival models to the federated setting. FairFSA jointly trains survival models using distributionally robust optimization, penalizing worst-case errors across subpopulations that exceed a specified probability threshold. Partially observed survival outcomes in clients are reconstructed with federated pseudo values (FPV) before model training to address censoring. Furthermore, we design a weight aggregation strategy by enhancing the FedAvg algorithm with a fairness-aware concordance index-based aggregation method to foster equitable performance distribution across clients. To the best of our knowledge, this is the first work to study and integrate fairness into Federated Survival Analysis. Comprehensive experiments on distributed non-IID datasets demonstrate FairFSA's superiority in fairness and accuracy over state-of-the-art FSA methods, establishing it as a robust FSA approach capable of handling censoring while providing equitable and accurate survival predictions for all subjects.

Introduction

Survival analysis plays a critical role in healthcare, enabling the analysis of time-to-event data, such as duration until death or recovery from a disease. Leveraging survival analysis techniques enables healthcare professionals to optimize treatment strategies and resource allocation, thereby enhancing patient care and survival outcomes. A key challenge in survival analysis is handling censored data, where events of interest remain unobserved for some individuals due to the end of the study, loss-to-follow-up, or withdrawal from the study. Additionally, stringent privacy regulations often limit data sharing across multiple institutions. Consequently, a significant amount of data cannot be leveraged to build efficient survival models, leading to suboptimal predictive performance. Federated Survival Analysis (FSA) (Andreux et al. 2020; Zhang, Toni, and Williams 2022; Rahman and Purushotham 2023c) has emerged as a solution, facilitating the analysis of distributed, sensitive survival data without necessitating direct data sharing among participating clients. This approach preserves privacy and data ownership by exchanging model parameters rather than raw data, fostering the development of robust and generalizable models that enhance local survival predictions. Moreover, FSA allows clients to fine-tune global models to their specific data distributions, ensuring that the model is optimally adapted to the specific characteristics of the client's data.

Despite these advancements, a critical challenge in FSA is the potential for bias introduced by sensitive attributes, such as race, gender, age, or ethnicity, which are often present in survival data. Electronic health records (EHRs) from minority groups often exhibit smaller sample sizes, shorter follow-up periods, fewer events, and higher censoring rates (Seyyed-Kalantari et al. 2020; Chen et al. 2023). These disparities can lead to unintentional bias in survival models (Gianfrancesco et al. 2018), and within the federated learning framework, there is a risk that these biases could be propagated and even amplified, resulting in unfair survival predictions (Paulus and Kent 2020; Mhasawade, Zhao, and Chunara 2021). Additionally, nonuniform censoring distributions across clients may introduce additional bias into the predictions made by federated survival models. Although addressing fairness in FSA is a pressing concern, incorporating fairness constraints and debiasing techniques remains a relatively unexplored research area due to the complexity of the challenges involved. This paper introduces `FairFSA`, a novel fair federated survival analysis framework that integrates a novel Federated Pseudo Value (FPV)-based deep learning model with Distributionally Robust Optimization (DRO) (Deng, Kamani, and Mahdavi 2020; Hu and Chen 2022a), called `PseudoDRO`, to address the following identified challenges:

- **Challenge 1: Requiring Access to Entire Training Data.** Traditional fairness approaches in survival analysis typically require full access to the entire training dataset to apply fairness constraints, often relying on pairwise or group comparisons (Keya et al. 2021; Rahman and Purushotham 2022; Do et al. 2023; Sonabend et al. 2022). Many of these methods optimize Cox partial likelihood loss or ranking loss (Keya et al. 2021; Hu and Chen 2022a), both of which involve expensive pairwise comparisons for ranking or similarity score evaluations (Steck et al. 2007; Chen 2020; Wu et al. 2021; Lee et al. 2018). This reliance on pairwise comparison makes these approaches unsuitable for FSA, where data access is decentralized. FairFSA addresses this issue by reformulating the client-specific survival analysis problem as a regression task using independent and identically distributed federated pseudo values (FPV) (Rahman and Purushotham 2023c,b). FPVs effectively handle censoring and ensure the separability of the loss function, satisfying the requirement for independent losses in **Distributionally Robust Optimization (DRO)**. This allows for the incorporation of fairness without the need for complete data access or pairwise comparisons, facilitating the successful integration of fairness into the federated learning framework.

- **Challenge 2: Dependence on Sensitive Demographic Information.** Most of the existing fair survival models require explicit specification and inclusion of sensitive demographic attributes (Keya et al. 2021; Rahman and Purushotham 2023c; Do et al. 2023), which may not always be available or easily identifiable in certain survival datasets. Additionally, removing such attributes can result in information loss or potentially increase bias, as they could be correlated with other non-sensitive features. Our proposed model PseudoDRO overcomes this challenge by using DRO to ensure fairness without requiring to specify sensitive attributes, making the framework applicable even in scenarios where such demographic information is missing or unidentified.
- **Challenge 3: Handling Censoring in Federated Settings.** Censoring can vary significantly across different clients and can be correlated with sensitive attributes, potentially leading to poor convergence and biased predictions in the global model (Rahman and Purushotham 2023c,b). FairFSA addresses this by applying federated pseudo values (FPV) at each institution, improving survival predictions. While the use of covariate-dependent FPV (Binder, Gerds, and Andersen 2014) could address the dependency of censoring on sensitive attributes, it lies beyond the scope of this paper.
- **Challenge 4: Achieving Local and Global Fairness.** Local fairness ensures that individuals or subgroups who are similar in relevant characteristics should receive similar predictions from a local survival model. On the other hand, global fairness ensures that the performance of the global survival model is equitable across different clients in a federated learning setting[1] Achieving fairness at the local level does not automatically guarantee fairness in the aggregated global model (Makhija et al. 2024). The classical Federated Averaging (FedAvg) method (McMahan et al. 2017) often gives more weight to clients with larger datasets, which does not guarantee improvement for all clients. FairFSA introduces a fairness-aware concordance index (CI) aggregation technique, which considers the CI performance on the validation set during the aggregation, ensuring that clients with poorer performance have less influence on the global model, thereby promoting fairness across all participating clients. In summary, FairFSA addresses the above challenges of incorporating fairness into FSA and provides a robust and generalizable survival model.

Our Main Contributions.

- To the best of our knowledge, this is the first effort to integrate fairness into federated survival analysis (FSA).
- We developed FairFSA, a novel fair FSA framework that integrates Federated Pseudo Value (FPV)-based deep learning with Distributionally Robust Optimization (DRO) to address fairness challenges in FSA.
- To ensure global fairness, we introduced a fairness-aware concordance index (CI) model aggregation technique.
- Extensive experiments on public survival datasets demonstrate that our FairFSA approach improves the performance and fairness of local models and achieves comparable performance to centralized survival models.

Problem Formulation

Consider a horizontal federated survival analysis (FSA) framework with K clients and one global server, where all participating clients / silos (e.g. hospitals) have the same set of features but different patients in their local survival dataset D_k, where $D_k = \{\mathbf{X}_{ik}, Y_{ik}, \delta_{ik}\}$. For a patient i in client k, $\mathbf{X}_{ik} \in \mathbb{R}^p$ is a p-dimensional feature vector, $\mathbf{Y}_{ik}$ is the observed event time, and $\delta_{ik} \in 0, 1$ is the event indicator. If $\delta_{ik} = 1$, the event time is the failure time T_{ik} and if $\delta_{ik} = 0$, the event time is the censoring time C_{ik}, i.e., actual failure time $T_{ik} > C_{ik}$. K participating clients train their own fairness-aware local models M_{θ_i} to obtain a fair global model M_θ to predict the conditional survival function $S(t|\mathbf{X}_{ik})$ at time t given the feature vector $\mathbf{X}_{ik}$ where $S(t|\mathbf{X}_{ik}) = P(T_{ik} > t|\mathbf{X}_{ik})$, i.e., the probability that the failure time T_{ik} is greater than a particular time t given $\mathbf{X}_{ik}$.

Related Work

Federated Survival Analysis (FSA) is becoming increasingly prominent in healthcare for its ability to enable privacy-preserving collaborative modeling of time-to-event data. Most FSA studies (Andreux et al. 2020; Zhang, Toni, and Williams 2022; Wang et al. 2022; Masciocchi et al. 2022) have utilized the Federated Averaging (FedAvg) algorithm (McMahan et al. 2017) in combination with Cox-based survival models. However, these models face challenges due to the non-separable loss function and proportional hazards (PH) assumptions within clients, which may not hold across different clients. Rahman et al. (Rahman and Purushotham 2023c) advanced the field by proposing FPV-based deep learning methods, addressing the shortcomings of Cox-based models. To enhance communication efficiency in resource-constrained environments, Archetti et al. (Archetti and Matteucci 2023) and Rahman et al. (Rahman and Purushotham 2023a) introduced federated survival

[1] Local and global fairness definitions are in the supplementary materials at this Github link: https://github.com/umbc-sanjaylab/FairFSA

forests (FedSurF) and a pseudo-value-based random forest model (FedPRF), respectively. Despite these advancements, existing approaches do not ensure fairness in predictions, a critical issue in healthcare that can lead to biased outcomes.

Fairness has been less explored in survival analysis due to the challenges of applying common fairness definitions to survival models. Keya et al. (Keya et al. 2021) introduced hazard-based fairness definitions and used them as constraints into the partial log-likelihood objective function, while Rahman et al. (Rahman and Purushotham 2022) introduced generalized fairness definitions for all survival models and enforced fairness constraints into the pseudo-value-based objective function. Additionally, Rahman et al. propose censoring-based fairness definitions. Zhang et al. (Zhang and Weiss 2022) focused on model accuracy in fairness metrics. Hu et al. (Hu and Chen 2022a) used Distributionally Robust Optimization (DRO) to achieve fairness in CoxPH models, though their approach relied on sample splitting. Our proposed FairFSA framework ensures local fairness by applying DRO to the pseudo-value-based objective function, which does not require sensitive attribute specification, pairwise patient comparisons, or sample splitting. FairFSA can effectively handle censoring by leveraging FPV and ensures global fairness through a fairness-aware aggregation process.

Our Proposed Approach: FairFSA

In this paper, we present a novel framework for fair federated survival analysis, named FairFSA, illustrated in Figure 1. FairFSA enables multi-institution collaboration by jointly training our proposed local fair survival model, PseudoDRO, across their distributed datasets. PseudoDRO is a deep neural network-based survival model which utilizes **Distributionally Robust Optimization (DRO)** to optimize a Federated Pseudo Values (FPV)-based objective function to enforce fairness during the local training process. By employing a sigmoid activation in the output layer, PseudoDRO generates survival probabilities via FPV predictions. Our FairFSA framework integrates three core components that collectively ensure fairness and effectively handle censoring challenge in FSA. First, it utilizes FPV to address non-uniform censoring in FSA. Second, it integrates DRO within the pseudo value-based objective function, which not only guarantees fairness in the local models but also facilitates fair federated training across all clients. Finally, the framework employs a concordance index (CI)-weighted aggregation process to ensure global fairness.

Federated Pseudo Values (FPV): For censored patients, the actual failure time remains unknown, making traditional regression analysis infeasible for time-to-event data. However, we can reformulate the survival analysis problem as a regression problem at the client level by replacing the incompletely observed survival function with consistent Federated Pseudo Values (FPV) (Rahman and Purushotham 2023c,a), which enables patient-specific survival predictions across all clients. Under the assumption that the failure time T_{ik} and the censoring time C_{ik} are independent given covariates $\mathbf{X}_{ik}$ (Graw, Gerds, and Schumacher 2009), the FPVs for the survival function for a subject i in client k at time t are defined as:

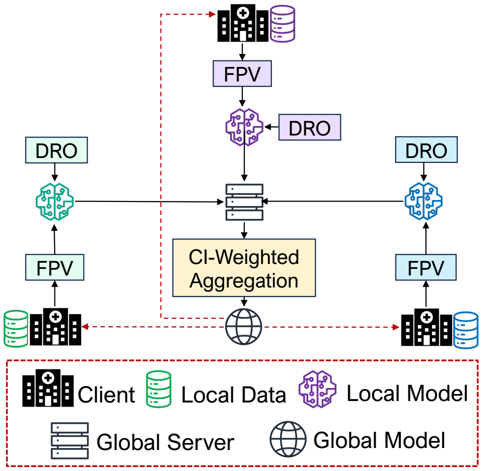

Figure 1: Our proposed FairFSA Framework. **DRO:** Distributionally Robust Optimization, **FPV:** Federated Pseudo Values, **CI:** Concordance Index.

$$J_{ik}(t) = n\hat{S}_G(t) - (n-1)\hat{S}_G^{-ik}(t); i = 1,..,n_k, k = 1,..,K \tag{1}$$

Here, $n = \sum_{k=1}^{K} n_k$. $\hat{S}_G(t)$ is the Kaplan Meier (KM) estimate of the global survival function, and $\hat{S}_G^{-ik}(t)$ is the KM estimate of the leave-one-out global survival function, obtained by omitting i^{th} subject from client k. t are the set of unique time points in the training data.

Using FPVs in federated settings offers several advantages: 1) FPVs address data heterogeneity and non-uniform censoring, 2) FPVs are computed without the need to share raw data, complying with privacy regulations, 3) FPVs make it feasible to apply FL techniques to survival analysis, 4) Due to their IID properties, FPVs can be easily used in separable loss functions, such as mean squared error loss, which relaxes the requirement for access to the entire training dataset, 5) The separable loss with FPVs as output allows applying DRO to ensure fairness without violating the assumption of independent loss terms.

DRO on FPV-based Objective Function: Let $\mathcal{P}_k$ denote the joint distribution of data points $(\mathbf{X}_{ik}, J_{ik}(t))$ in client k, where $\mathbf{X}_{ik}$ are the feature vector and $J_{ik}(t)$ are the FPVs at unique time point t. The joint distribution $\mathcal{P}_k$ can be decomposed into L sub-distributions $\mathcal{P}_{kl}$, each occurring with a probability π_{kl} where $\sum_{l=1}^{L} \pi_{kl} = 1$. In other words, $\mathcal{P}_k$ be a mixture of L distributions corresponding to L groups, $\mathcal{P}_k := \sum_{l=1}^{L} \pi_{kl}\mathcal{P}_{kl}$. We assume that the specific details of these subpopulations, such as the number of subpopulations L, are unknown. The goal of DRO is to minimize the risk $\mathcal{R}_{max}(w_k)$, defined as:

$$\mathcal{R}_{max}(w_k) := \max_{l=1,2,...,L} \mathbb{E}_{(X_k,J_k(t))\sim\mathcal{P}_{kl}}[\ell(w_k; X_k, J_k(t))] \tag{2}$$

where $\ell(.)$ is the FPV-based loss function dependent on model parameters w_k and data point $(\mathbf{X}_k, J_k(t))$. The FPV-based loss function (Rahman and Purushotham 2022) for client k at time t, $\ell_k(t)$, is,

$$\ell_k(t_j) = \frac{1}{n_k}\sum_{i=1}^{n_k}[J_{ik}(t_j)(1-2\hat{S}(t_j|x_{ik}))+\hat{S}^2(t_j|x_{ik})] \tag{3}$$

where $\hat{S}(t_j|x_{ik})$ are the predicted survival probability at time point t_j for i^{th} patient in client k. FPVs are calculated at M unique time points in the training data, and the total loss over M unique time points is $\ell_k(t) = \frac{1}{M}\sum_{j=1}^{M} \ell_k(t_j)$.

Directly minimizing risk $\mathcal{R}_{max}(w_k)$ is infeasible, since $(\mathcal{P}_{kl}, \pi_{kl})_{l=1}^{L}$ and L are unknown. Thus, we solve an optimization problem that minimizes an empirical upper bound on $\mathcal{R}_{max}(w_k)$. We employ the DRO formulation proposed by Hashimoto et al. (Hashimoto et al. 2018) and Hu et al. (Hu and Chen 2022a), which provides a tractable approach to minimizing an upper bound on the worst-case risk $\mathcal{R}_{DRO}(w_k; r_{max})$. We can define the $\mathcal{R}_{DRO}(w_k; r_{max})$ following the proposition below.

Proposition 1 *(Follows from Lemma 1 in (Duchi and Namkoong 2021; Hu and Chen 2022a))*

Let $\ell(w_k, X_k, J_k(t))$ be an upper semi-continuous with respect to w_k. Then

$$\mathcal{R}_{DRO}(w_k; r_{max}) := \underset{\epsilon\in\mathbb{R}}{inf}\{\sqrt{2(\frac{1}{\alpha}-1)^2+1}$$
$$\sqrt{\mathbb{E}_{(X_k,J_k(t))\sim\mathcal{P}_{kl}} max[0,(\ell(w_k; X_k, J_k(t)) - \epsilon)]^2} + \epsilon \tag{4}$$

Here, where ϵ is a regularization parameter. Similar to (Hu and Chen 2022a), we minimize an empirical version of $\mathcal{R}_{DRO}(w_k; r_{max})$ by replacing the expectation in equation 4 by an empirical average, resulting in following optimization problem:

$$\underset{\Theta_k,\epsilon\in\mathbb{R}}{min} \mathcal{L}_{DRO}(w_k; \epsilon) \tag{5}$$

Here, Θ_k is the feasible set of parameters of the model in client k. The empirical loss $\mathcal{L}_{DRO}(w_k; \epsilon)$ is defined as:

$$\mathcal{L}_{DRO}(w_k, \epsilon) := \sqrt{2(\frac{1}{\alpha}-1)^2+1}$$
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n_k} max[0,(\ell_i(w_k; X_k, J_k(t)) - \epsilon)]^2} + \epsilon \tag{6}$$

For an optimal value of ϵ, the loss of a patient that does not exceed ϵ is ignored. To solve the optimization problem, we use an iterative gradient descent approach (Hu et al. 2022; Hu, Wang, and Lyu 2023; Hu and Chen 2022a). The procedure involves the following steps: **Initialization:** Initialize model parameters w_k. **Update ϵ:** Fix w_k and update ϵ by finding the value that minimizes $\mathcal{L}_{DRO}(w_k; \epsilon)$. This is achieved using binary search due to the convexity of $\mathcal{L}_{DRO}(w_k; \epsilon)$ with respect to ϵ. **Update** w_k: Fix ϵ and update w_k by minimizing $\mathcal{L}_{DRO}(w_k; \epsilon)$ using gradient descent. This process iterates until specified stopping criteria are met, such as a maximum number of iterations or early stopping based on a validation metric (Hu and Chen 2022a).

Concordance Index-Weighted Aggregation: While we use the classical Federated Averaging (FedAvg) (McMahan et al. 2017) to aggregate the local models from clients, combining fair local models does not necessarily guarantee fairness in the federated global model. This is because the aggregation process used in FedAvg performs a simple weighted average based on clients' sample sizes to combine local models, which can introduce or amplify biases, even if the individual local models are fair. FedAvg enforces that clients with larger sample sizes have more influence in the global model. However, clients with smaller sample sizes may still show good performance when the data is less noisy and instances are influential. Therefore, the performance of the models should be taken into consideration during the weight aggregation. To address this issue, we propose a fairness-aware weighted aggregation based on the local models' validation C-index performance. For a communication round t, the global weights can be computed as: $w^t = \sum_{k=1}^{K} \frac{CI_k n_k}{\sum_{k=1}^{K} CI_k n_k} w_k^{t-1}$ where CI_k is the validation C-index of local client k. Note that, we can replace the validation C-index with validation fairness metrics, such as individual fairness or the average of individual, group, and intersectional fairness, to ensure global fairness in terms of fairness metrics.

Federated Training: Once the local fair `PseudoDRO` models are trained across clients, the `FairFSA` framework employs a global server to coordinate the communication and aggregation of these models. During each communication round, the server sends the clients a global `PseudoDRO` model, represented by w^v. The local clients then update their respective local models by incorporating the global model parameters and training on their local data. These newly trained local models, denoted as Δw_k^v, are sent back to the global server. The global server then performs a CI-aggregation process to aggregate the local model parameters and update the global model. This CI-aggregation process weighs the contributions of each local model based on its validation C-index performance, as described earlier. The updated global model is subsequently sent back to the local clients by the server. This communication and aggregation process is repeated for a user-specified number of communication rounds, denoted as V. Once the V rounds are completed, the globally updated model is utilized to make subject-specific fair survival probability predictions. By iteratively aggregating the local models through CI-aggregation, `FairFSA` aims to ensure that the global model benefits from the collective knowledge of all clients while maintaining fairness in the final predictions.

Experiments

We conduct extensive experiments on both centralized and decentralized non-IID survival datasets to assess the performance of our proposed `PseudoDRO` and `FairFSA` models. Our study aims to answer the following research questions: **RQ1:** How does our centralized fair PseudoDRO model compare to existing fair survival models in terms of accuracy and fairness in a centralized setting? **RQ2:** In a federated setting, how do the performance and fairness of the `FairFSA` models compare with baseline fair survival models? **RQ3:** How effectively does the proposed `FairFSA` model handle censoring in decentralized settings?

Datasets: We evaluated four publicly available real-world survival datasets namely FLChain (Dispenzieri et al. 2012), SUPPORT (Knaus et al. 1995), SEER (TENG 2019), and MSK-MET (Nguyen et al. 2022) for centralized and federated fair survival analysis. Detailed descriptions of these datasets are provided in the supplementary materials.

Model Comparisons: We conducted a comprehensive evaluation of various survival models in both centralized and federated settings. (a) **Centralized Models:** To evaluate the performance of our proposed `PseudoDRO` model in centralized settings, we compared with the following baseline survival models with different properties. **No Fairness Constraints:** CoxPH (Cox 1972), DeepSurv (Katzman et al. 2018), DeepHit (Lee et al. 2018), DeepPseudo (Rahman et al. 2021; Rahman and Purushotham 2022); **With Fairness Constraints:** FairSurv (Cox) (Keya et al. 2021), FairSurv (DeepSurv) (Keya et al. 2021), FIDP (Rahman and Purushotham 2022); **With DRO:** DRO-Cox (Hu and Chen 2022a), Deep DRO-Cox (Hu and Chen 2022a), DRO-Cox (Split) (Hu and Chen 2022a), Deep DRO-Cox (Split) (Hu and Chen 2022a), DRO-DeepHit (Split) (Hu and Chen 2022b). (b) **Federated Models:** In the federated settings, we compare our `FairFSA` framework with the federated variant of the corresponding baseline models: **No Fairness Constraints:** FedCox (Andreux et al. 2020), FedDP (Zhang, Toni, and Williams 2022), Fed-DeepHit (Rahimian et al. 2022; Rahman and Purushotham 2023c), FedPseudo (Rahman and Purushotham 2023c); **With Fairness Constraints:** Fair-FedCox, Fair-FedDP, FIDP-FSA; **With DRO:** Fed-DRO-Cox, Fed-Deep-DRO-Cox, Fed-DRO-Cox (Split), Fed-Deep-DRO-Cox (Split), Fed-DRO-DeepHit (Split).

Performance Metrics: In both centralized and federated settings, the accuracy and fairness of the models are evaluated using the combined test data shared from clients to the server. **Accuracy metrics:** We employ the time-dependent concordance index (C-Index) (Harrell Jr, Lee, and Mark 1996), integrated Brier Score (IBS) (Graf et al. 1999), MAE-Uncensored (Qi et al. 2023), and MAE-Hinge (Qi et al. 2023) as our accuracy metrics. These metrics are computed using the `SurvivalEVAL` package (Qi, Sun, and Greiner 2023). **Fairness metrics:** The fairness of the models is assessed using individual fairness (FI) (Rahman and Purushotham 2022), group fairness (FG) (Rahman and Purushotham 2022), intersectional fairness (FIN) (Foulds et al. 2020; Hu and Chen 2022a), and a summary fairness metric $F_{avg} = \frac{FI+FG+FIN}{3}$. We assess the global fairness of the federated survival models by analyzing the standard deviation of accuracy and fairness metrics across local clients.

Implementation Details: At each client, local datasets are split into 80% training and 20% test sets. The local test datasets are evaluated in both centralized and federated settings, ensuring that the combined test data remains the same across all experiments. We perform 5-fold cross-validation and compute evaluation metrics on the same combined test set, reporting the mean and standard deviation (shown in the supplementary materials) of each metric. For federated settings, we simulate a non-IID federated environment by assigning subjects to clients in a manner that skews the event time distribution towards specific quantiles of the time horizon. Different clients have different quantiles for the skewness of the event time distribution. We use the Adam optimizer (Kingma and Ba 2014) with early-stopping based on the best validation C-index. In the federated settings, we consider 5 clients, with a total of 10 communication rounds and 20 local epochs per round. We also choose the learning rate from [0.01, 0.001] and set the batch size to 128. We compute the pseudo values for the unique time points in the training data. Detailed hyperparameter settings are provided in the supplementary materials.

Results and Discussion

Performance of our PseudoDRO in centralized settings. Table 1 presents the performance comparison of the proposed PseudoDRO model against the baseline survival models in the centralized settings in terms of accuracy and fairness metrics across three real-world survival datasets: SUPPORT, FLChain, and SEER. The results indicate that PseudoDRO consistently achieves significant performance improvements across all three datasets. It demonstrates competitive accuracy, securing the best or second-best C-Index and MAE scores, which are critical indicators of model precision. In terms of fairness, PseudoDRO outperforms the baselines by obtaining the lowest F_{avg} values across all datasets, highlighting its ability to produce unbiased predictions. The results for individual fairness (FI), group fairness (FG), and intersectional fairness (FIN) are shown in the supplementary materials. PseudoDRO outperforms other survival models significantly in fairness while maintaining high accuracy. PseudoDRO's superior performance is attributed to the utilization of pseudo values for handling censoring and DRO for ensuring fair predictions, which effectively balances the trade-offs between accuracy and fairness, ensuring robust and equitable survival predictions.

Performance of our FairFSA in federated settings. Table 2 provides a comprehensive evaluation of the survival models in federated settings, including the proposed FairFSA, across three non-IID decentralized datasets: SUPPORT, FLChain, and SEER. FairFSA consistently demonstrates top-tier performance across all datasets, securing the best or second-best scores in both accuracy and fairness metrics. It achieves high C-Index values, indicating robust discriminative ability, and maintains competitive scores in MAE-Uncen and MAE-Hinge. Notably, FairFSA excels in fairness metrics, consistently obtaining the lowest F_{avg} values across all datasets. This indicates its strong ability to produce unbiased and equitable predictions. Specifically, in the SUPPORT dataset, FairFSA outperforms other models significantly in fairness while maintaining high accuracy. In the FLChain and SEER datasets, it performs similar or better than the best-performing models in accuracy and leads in fairness metrics.

Handling censoring in federated settings. To demonstrate the efficacy of our pseudo-value-based approach in handling censoring, we compare the performance on uncensored and censored decentralized datasets. The uncensored dataset consists of three clients, each with the same number of uncensored patients and no censored patients. For the censored dataset, we replace two-thirds of the uncensored patients in client 2 with censored patients to investigate the impact of censoring on global model performance. In Figure 3, we illustrate the difference in C-index performance between uncensored and censored datasets. We observe that the performance of the pseudo-value-based models is less affected by censoring and remains consistent across clients. In contrast, non-pseudo-value-based models are significantly

Dataset	Metric	Without Fairness				With Fairness Constraint			DRO					
		CoxPH	DeepSurv	DeepHit	DeepPseudo	FairSurv (Cox)	FairSurv (DeepSurv)	FIDP	DRO-Cox	Deep DRO-Cox	DRO-Cox (Split)	Deep DRO -Cox (Split)	DRO-DeepHit (Split)	PseudoDRO
SUPPORT	C-Index	0.60	0.61	0.57	0.62	0.51	0.57	0.62	0.55	0.52	0.52	0.55	0.53	0.61**
	IBS	0.21	0.20	0.22	0.20	0.23	0.23	0.20	0.23	0.23	0.24	0.37	0.27	0.25
	MAE-Uncen	387	526	479	474	262	263	524	261	301	219	3251	841	364
	MAE-Hinge	442	495	472	467	460	461	503	457	468	464	2116	643	445*
	Favg	0.049	0.075	0.024	0.072	0.003	0.002	0.009	0.004	0.018	0.003	0.003	0.004	0.000*
FLChain	C-Index	0.79	0.79	0.78	0.79	0.74	0.72	0.80	0.78	0.74	0.77	0.71	0.72	0.77
	IBS	0.07	0.07	0.08	0.07	0.10	0.10	0.07	0.09	0.09	0.10	0.11	0.27	0.24
	MAE-Uncen	8885	9369	2674	8165	8736	8733	8921	6868	8291	14312	21973	1158	1986**
	MAE-Hinge	1812	1907	538	1666	1756	1756	1820	1380	1668	2877	4417	1549	426*
	Favg	0.064	0.064	0.022	0.062	0.001	0.000	0.019	0.017	0.029	0.001	0.001	0.002	0.001**
SEER	C-Index	0.73	0.73	0.71	0.75	0.64	0.69	0.74	0.72	0.61	0.72	0.69	0.63	0.73
	IBS	0.10	0.09	0.10	0.10	0.12	0.12	0.10	0.11	0.11	0.12	0.17	0.28	0.22
	MAE-Uncen	153	190	58	140	155	155	145	133	392	175	384	22	53**
	MAE-Hinge	30	38	11	28	30	30	29	26	76	34	80	19	11*
	Favg	0.041	0.050	0.031	0.038	0.001	0.001	0.009	0.012	0.040	0.001	0.007	0.008	0.003**

Table 1: Performance comparison of the models in the **centralized** settings on SUPPORT, FLChain, and SEER datasets. The best performance is marked with (*), and the second-best performance is marked with (**).

Dataset	Metric	Without Fairness				With Fairness Constraint			DRO					
		FedCox	FedDP	Fed-DeepHit	FedPseudo	Fair-FedCox	Fair-FedDP	FIDP-FSA	Fed-DRO -Cox	Fed-Deep -DRO-Cox	Fed-DRO -Cox (Split)	Fed-Deep-DRO -Cox (Split)	Fed-DRO- DeepHit (Split)	FairFSA
SUPPORT	C-Index	0.60	0.60	0.53	0.62	0.58	0.60	0.61	0.55	0.52	0.52	0.53	0.52	0.61**
	IBS	0.21	0.22	0.25	0.20	0.23	0.23	0.23	0.23	0.23	0.25	0.25	0.27	0.25
	MAE-Uncen	387	819	1246	510	262	262	2559	262	329	633	277	801	336
	MAE-Hinge	443	661	834	493	459	460	1830	457	480	652	483	629	422*
	Favg	0.048	0.094	0.021	0.077	0.003	0.002	0.117	0.003	0.024	0.003	0.003	0.011	0.001*
FLChain	C-Index	0.79	0.78	0.78	0.79	0.78	0.78	0.78	0.78	0.77	0.76	0.75	0.67	0.79*
	IBS	0.07	0.07	0.09	0.07	0.09	0.85	0.07	0.09	0.09	0.11	0.11	0.26	0.24
	MAE-Uncen	9275	12303	2729	9243	6588	2238	11905	6786	7330	22329	38938	1175	2139**
	MAE-Hinge	1890	2499	549	1885	1324	3758	2417	1364	1473	4489	7827	1468	447*
	Favg	0.064	0.064	0.024	0.064	0.019	0.001	0.064	0.016	0.023	0.001	0.001	0.009	0.001*
SEER	C-Index	0.71	0.68	0.66	0.72	0.71	0.66	0.72	0.70	0.63	0.68	0.67	0.53	0.70
	IBS	0.10	0.11	0.12	0.10	0.11	0.77	0.10	0.11	0.11	0.12	0.23	0.34	0.23
	MAE-Uncen	167	602	59	209	137	31	269	137	671	283	29	21	64
	MAE-Hinge	33	117	11	41	27	56	53	26	130	55	22	28	12**
	Favg	0.040	0.046	0.028	0.036	0.008	0.001	0.040	0.015	0.040	0.001	0.017	0.026	0.000*

Table 2: Performance comparison of the models in the **federated (decentralized)** settings on SUPPORT, FLChain, and SEER datasets. The best performance is marked with (*), and the second-best performance is marked with (**).

impacted by censoring and exhibit inconsistent performance across clients. In the supplementary materials, we provide a detailed performance comparison of the models between censored and uncensored data, highlighting the superior censoring handling capability of pseudo-value-based models, including FairFSA.

Trade-off Between Fairness and Accuracy. We show the trade-off between fairness and accuracy by varying the trade-off parameter α (ranging from 0.001 to 1.0) in Figure 2 across three survival datasets: FLChain, SEER, and SUPPORT. The accuracy metrics used include the Concordance Index (C-index), Integrated Brier Score (IBS), Mean Absolute Error-Uncensored (MAE-Uncen), and Mean Absolute Error-Hinge (MAE-Hinge). For fairness, we utilize the summary fairness metric F_{avg}. As α increases, both the C-index and F_{avg} show an upward trend, while the IBS decreases. The MAE-based metrics, however, exhibit instability with varying α values. We select the value of α at which F_{avg} and IBS intersect, balancing fairness and accuracy. In table 1 and 2, we present the accuracy levels achieved while optimizing for the best fairness performance. It is important to note that for C-index, higher values indicate better performance, while for IBS, MAE-Uncen, MAE-Hinge, and F_{avg}, lower values suggest better performance.

Global Fairness. Table 3 illustrates the effectiveness of our proposed CI-Weighted aggregation method in enhancing global fairness within the FairFSA model, compared to the traditional FedAvg aggregation. We assess both accuracy and fairness across clients, along with the standard deviation of these performance metrics, on the imbalanced MSK-MET dataset (Nguyen et al. 2022), which contains an unequal distribution of patients across white and non-white groups within all clients. A lower standard deviation reflects more consistent local performance, indicating that the models provide equitable predictions across all clients. The results show that FairFSA when combined with our CI-Weighted aggregation technique, achieves better global fairness (as evidenced by a lower standard deviation) compared to FairFSA using FedAvg. Furthermore, a comparison of global fairness across all models on three survival datasets is presented in the supplementary materials. The results highlight the performance consistency of FairFSA across local clients, with a particular focus on its lowest standard deviation in fairness metrics. This consistency underscores FairFSA's ability to deliver fair and uniform predictions across different clients.

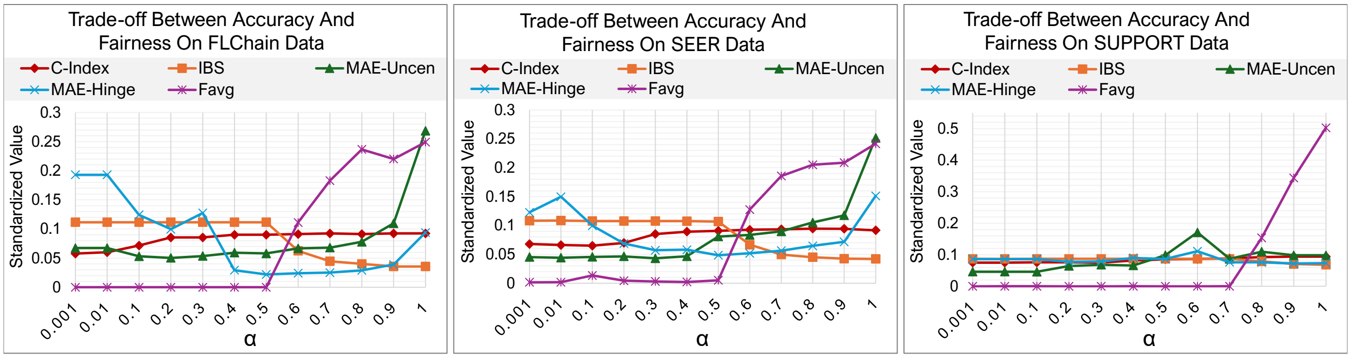

Figure 2: Plot for the trade-off between fairness and accuracy on survival datasets. The figure shows the impact of the trade-off parameter, α, on the accuracy (C-Index, IBS, MAE-Uncen, and MAE-Hinge) and average fairness F_{avg} obtained by Pseudo-DRO. X-axis represents the values of α and Y-axis represents the standardized values of accuracy and fairness measures.

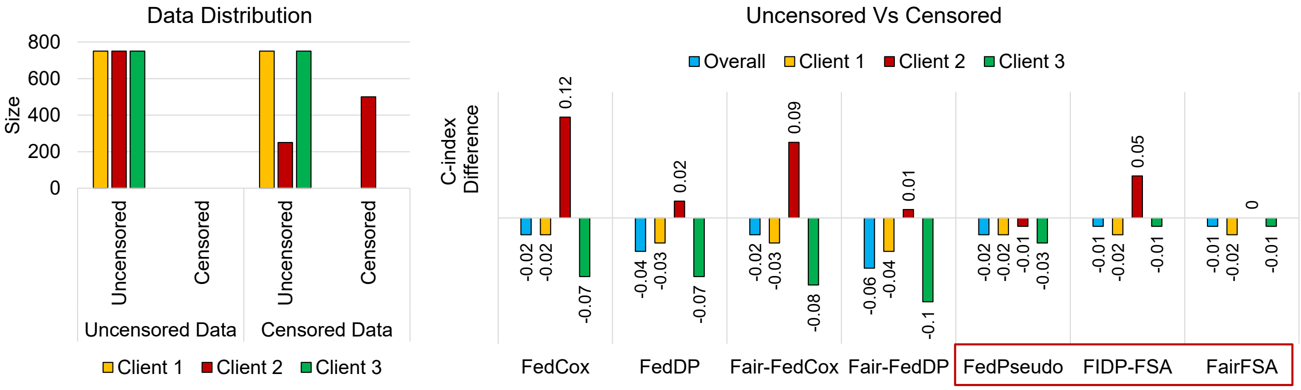

Figure 3: Censoring impact on the C-index performance of global model. **Left:** Data distribution in censored and uncensored groups. Uncensored data contains no censored observations. In the censored data, 500 uncensored subjects in client 2 are replaced by censored subjects. **Right:** The C-index of the global models are compared, where both the entire test data and client-specific test data are utilized. FedPseudo, FIDP-FSA, and FairFSA are pseudo-value-based FSA frameworks.

Metric	Model	Client 1	Client 2	Client 3	Global Fairness (Std) ↓
C-Index	FairFSA-FedAvg	0.67	0.75	0.73	0.04
	FairFSA	0.69	0.73	0.73	**0.03**
IBS	FairFSA-FedAvg	0.17	0.17	0.14	0.02
	FairFSA	0.17	0.17	0.14	0.02
MAE-Uncen	FairFSA-FedAvg	836	537	557	168
	FairFSA	828	560	592	**147**
MAE-Hinge	FairFSA-FedAvg	348	257	246	56
	FairFSA	345	269	255	**49**
FI	FairFSA-FedAvg	0.190	0.191	0.199	0.005
	FairFSA	0.187	0.190	0.193	**0.003**

Table 3: Comparison of global fairness between FairFSA with FedAvg and with our CI-Weighted aggregation. **Std** is the standard deviation of clients' performance.

Conclusion

To the best of our knowledge, this study is the first to integrate fairness into federated survival analysis. We developed FairFSA, a novel framework that jointly trains Federated Pseudo Value (FPV)-based deep learning models, coupled with Distributionally Robust Optimization (DRO) on top of the pseudo-value-based objective function. To ensure global fairness, we introduced a fairness-aware concordance index aggregation process in FedAVG. Extensive experiments on publicly available survival datasets demonstrate that our FairFSA approach balances fairness and accuracy, achieving performance comparable to centralized survival models. Experimental results indicate that FairFSA consistently achieves high accuracy, as evidenced by competitive C-index and MAE metrics while excelling in fairness metrics across various datasets. Future work will investigate the calibration of predicted probabilities to enhance IBS performance while maintaining the balance between fairness and accuracy. Overall, FairFSA sets a new benchmark in federated survival analysis, offering a promising direction for developing fair and accurate predictive models in decentralized settings, with potential applications in healthcare and beyond.

Acknowledgments

This work is supported by CAREER grant #2238743 from the US National Science Foundation (NSF).

References

Andreux, M.; Manoel, A.; Menuet, R.; Saillard, C.; and Simpson, C. 2020. Federated survival analysis with discrete-time cox models. *arXiv preprint arXiv:2006.08997*.

Archetti, A.; and Matteucci, M. 2023. Federated Survival Forests. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–9.

Binder, N.; Gerds, T. A.; and Andersen, P. K. 2014. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis*, 20: 303–315.

Chen, G. H. 2020. Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine Learning for Healthcare Conference*, 537–565. PMLR.

Chen, R. J.; Wang, J. J.; Williamson, D. F.; Chen, T. Y.; Lipkova, J.; Lu, M. Y.; Sahai, S.; and Mahmood, F. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6): 719–742.

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

Deng, Y.; Kamani, M. M.; and Mahdavi, M. 2020. Distributionally robust federated averaging. *Advances in neural information processing systems*, 33: 15111–15122.

Dispenzieri, A.; et al. 2012. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*. Elsevier.

Do, H.; Chang, Y.; Cho, Y. S.; Smyth, P.; and Zhong, J. 2023. Fair Survival Time Prediction via Mutual Information Minimization. In *Machine Learning for Healthcare Conference*, 128–149. PMLR.

Duchi, J. C.; and Namkoong, H. 2021. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3): 1378–1406.

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

Gianfrancesco, M. A.; Tamang, S.; Yazdany, J.; and Schmajuk, G. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11): 1544–1547.

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

Graw, F.; Gerds, T. A.; and Schumacher, M. 2009. On pseudo-values for regression analysis in competing risks models. *Lifetime data analysis*, 15: 241–255.

Harrell Jr, F. E.; Lee, K. L.; and Mark, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4): 361–387.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 1929–1938. PMLR.

Hu, S.; and Chen, G. H. 2022a. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, 62–87. PMLR.

Hu, S.; and Chen, G. H. 2022b. Fairness in Survival Analysis with Distributionally Robust Optimization.

Hu, S.; Wang, X.; and Lyu, S. 2023. Rank-based decomposable losses in machine learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hu, S.; Ying, Y.; Wang, X.; and Lyu, S. 2022. Sum of ranked range loss for supervised learning. *Journal of Machine Learning Research*, 23(112): 1–44.

Katzman, J. L.; et al. 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.

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

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.

Lee, C.; Zame, W.; Yoon, J.; and Van Der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Makhija, D.; Han, X.; Ghosh, J.; and Kim, Y. 2024. Achieving Fairness Across Local and Global Models in Federated Learning. *arXiv preprint arXiv:2406.17102*.

Masciocchi, C.; Gottardelli, B.; Savino, M.; Boldrini, L.; Martino, A.; Mazzarella, C.; Massaccesi, M.; Valentini, V.; and Damiani, A. 2022. Federated Cox Proportional Hazards Model with multicentric privacy-preserving LASSO feature selection for survival analysis from the perspective of personalized medicine. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, 25–31. IEEE.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mhasawade, V.; Zhao, Y.; and Chunara, R. 2021. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8): 659–666.

Nguyen, B.; Fong, C.; Luthra, A.; Smith, S. A.; DiNatale, R. G.; Nandakumar, S.; Walch, H.; Chatila, W. K.; Madupuri, R.; Kundra, R.; et al. 2022. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell*, 185(3): 563–575.

Paulus, J. K.; and Kent, D. M. 2020. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, 3(1): 99.

Qi, S.-a.; Kumar, N.; Farrokh, M.; Sun, W.; Kuan, L.-H.; Ranganath, R.; Henao, R.; and Greiner, R. 2023. An effective meaningful way to evaluate survival models. *arXiv preprint arXiv:2306.01196*.

Qi, S.-a.; Sun, W.; and Greiner, R. 2023. SurvivalEVAL: A Comprehensive Open-Source Python Package for Evaluating Individual Survival Distributions. In *Proceedings of the AAAI Symposium Series*, volume 2, 453–457.

Rahimian, S.; Kerkouche, R.; Kurth, I.; and Fritz, M. 2022. Practical challenges in differentially-private federated survival analysis of medical data. In *Conference on Health, Inference, and Learning*, 411–425. PMLR.

Rahman, M. M.; Matsuo, K.; Matsuzaki, S.; and Purushotham, S. 2021. Deeppseudo: Pseudo value based deep learning models for competing risk analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 479–487.

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

Rahman, M. M.; and Purushotham, S. 2023a. Communication-Efficient Pseudo Value-Based Random Forests for Federated Survival Analysis. In *Proceedings of the AAAI Symposium Series*, volume 2, 458–466.

Rahman, M. M.; and Purushotham, S. 2023b. Federated Competing Risk Analysis. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2106–2115.

Rahman, M. M.; and Purushotham, S. 2023c. FedPseudo: Privacy-Preserving Pseudo Value-Based Deep Learning Models for Federated Survival Analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1999–2009.

Seyyed-Kalantari, L.; Liu, G.; McDermott, M.; Chen, I. Y.; and Ghassemi, M. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, 232–243. World Scientific.

Sonabend, R.; Pfisterer, F.; Mishler, A.; Schauer, M.; Burk, L.; Mukherjee, S.; and Vollmer, S. 2022. Flexible group fairness metrics for survival analysis. *arXiv preprint arXiv:2206.03256*.

Steck, H.; Krishnapuram, B.; Dehing-Oberije, C.; Lambin, P.; and Raykar, V. C. 2007. On ranking in survival analysis: Bounds on the concordance index. *Advances in neural information processing systems*, 20.

TENG, J. 2019. SEER Breast Cancer Data. *IEEE Dataport*.

Wang, X.; Zhang, H. G.; Xiong, X.; Hong, C.; Weber, G. M.; Brat, G. A.; Bonzel, C.-L.; Luo, Y.; Duan, R.; Palmer, N. P.; et al. 2022. SurvMaximin: robust federated approach to transporting survival risk prediction models. *Journal of biomedical informatics*, 134: 104176.

Wu, Z.; Yang, Y.; Fashing, P. A.; and Tresp, V. 2021. Uncertainty-aware time-to-event prediction using deep kernel accelerated failure time models. In *Machine Learning for Healthcare Conference*, 54–79. PMLR.

Zhang, D. K.; Toni, F.; and Williams, M. 2022. A federated cox model with non-proportional hazards. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, 171–185. Springer.

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