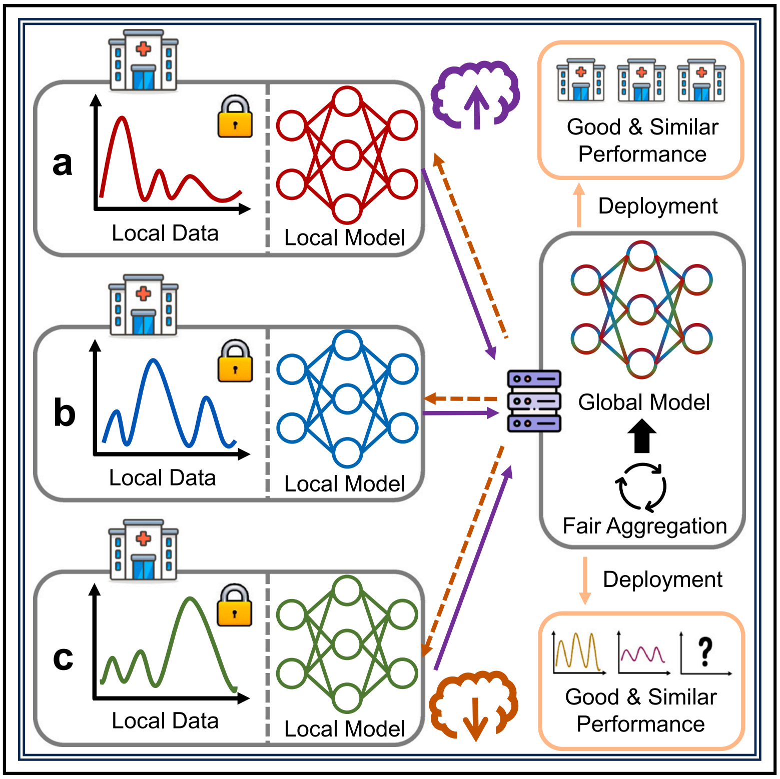
Patterns

Unified fair federated learning for digital healthcare

Graphical abstract



Highlights

- We enabled fair federated model training among healthcare institutions
- Our framework unifies diverse levels of fairness in federated learning
- Our efficient solution flexibly balances utility and different levels of fairness
- We applied our solution to the four digital healthcare tasks in federated settings

Authors

Fengda Zhang, Zitao Shuai, Kun Kuang, Fei Wu, Yueting Zhuang, Jun Xiao

Correspondence

kunkuang@zju.edu.cn

In brief

This study employed robust optimization to achieve consistent and high-performing federated models for diverse demographics. To ensure fairness in federated learning, a unified framework for healthcare institutions is presented that includes various fairness metrics and is implemented using an efficient and collaborative machine learning procedure. The proposed solution was tested in four digital medical scenarios in a federated setting. The experimental results show that the proposed approach could maintain model performance while improving the desired levels of fairness.

Zhang et al., 2024, Patterns *5*, 100907
January 12, 2024 © 2023 The Author(s).
https://doi.org/10.1016/j.patter.2023.100907

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Unified fair federated learning for digital healthcare

Fengda Zhang,[1,2] Zitao Shuai,[1,2] Kun Kuang,[1,3,*] Fei Wu,[1] Yueting Zhuang,[1] and Jun Xiao[1]
[1]Zhejiang University, 38 Zheda Road, Hangzhou 310058, Zhejiang, China
[2]These authors contributed equally
[3]Lead contact
*Correspondence: kunkuang@zju.edu.cn

https://doi.org/10.1016/j.patter.2023.100907

THE BIGGER PICTURE Federated learning (FL) enables healthcare institutions to collaboratively train high-quality medical machine learning models without accessing private medical data. Traditional FL may yield an unfair model with a good overall performance but still perform poorly with underrepresented subgroups. Because of the varied performance of these models among patients from different racial or ethnic groups, or those with specific physical conditions, differences in the quality of treatment recommendations may consequently cause serious social inequality problems. Frameworks that improve the fairness of FL will be essential in the implementation of digital healthcare.

1 2 3 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Federated learning (FL) is a promising approach for healthcare institutions to train high-quality medical models collaboratively while protecting sensitive data privacy. However, FL models encounter fairness issues at diverse levels, leading to performance disparities across different subpopulations. To address this, we propose Federated Learning with Unified Fairness Objective (FedUFO), a unified framework consolidating diverse fairness levels within FL. By leveraging distributionally robust optimization and a unified uncertainty set, it ensures consistent performance across all subpopulations and enhances the overall efficacy of FL in healthcare and other domains while maintaining accuracy levels comparable with those of existing methods. Our model was validated by applying it to four digital healthcare tasks using real-world datasets in federated settings. Our collaborative machine learning paradigm not only promotes artificial intelligence in digital healthcare but also fosters social equity by embodying fairness.

INTRODUCTION

Artificial intelligence (AI) holds tremendous potential for revolutionizing the medical field and advancing digital health applications.[1–6] However, its widespread implementation faces challenges, particularly concerning the handling of medical data and privacy concerns. Stringent regulations, such as the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), mandate the protection of sensitive patient information, making it impractical to gather all the necessary data for comprehensive AI training. Therefore, striking a delicate balance between harnessing the potential of AI and adhering to legal and ethical data privacy principles is important. To address this, researchers and developers must explore innovative approaches that enable AI models to be trained without compromising individual data privacy.[7–9]

Federated learning (FL) represents a crucial machine learning paradigm in which distributed clients (e.g., several medical institutions) collaboratively train a shared global model while retaining their private data.[10–13] However, inherent biases may arise in the federated model because of spurious correlations and distribution shifts across data subpopulations.[14–18] Consequently, the model's performance may significantly degrade for certain data subpopulations, leading to concerns regarding unfairness, particularly in critical domains such as healthcare. Recently, addressing this issue and achieving an unbiased federated model with equitable performance have emerged as paramount objectives and pivotal research themes.

In this study, we present a systematic exploration of the multifaceted nature of the fairness of model performance in FL. To enhance its clarity, as shown in Figure 1, we categorize these levels into four distinct dimensions: client-level fairness, attribute-level

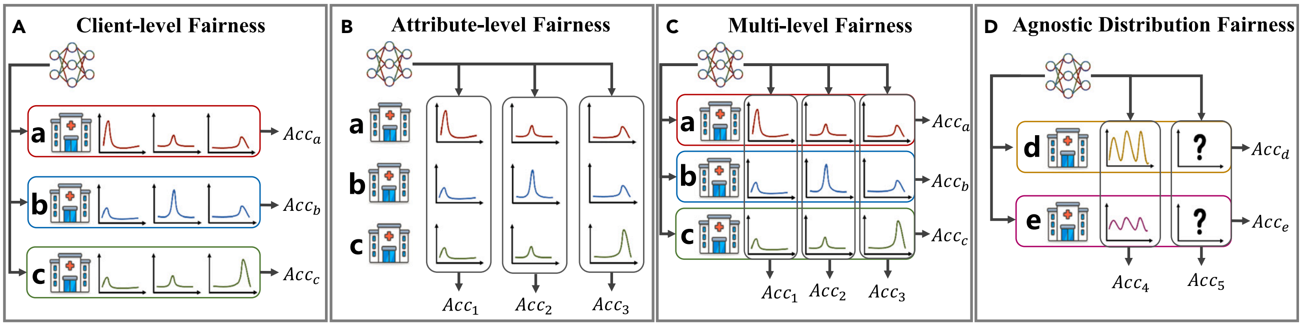


Figure 1. Illustration of diverse levels of fairness in federated learning scenario

(A) Client-level fairness (horizontal fairness) requires that the federated model have consistent accuracy across different clients (hospitals), i.e., $Acc_a = Acc_b = Acc_c$.

(B) Attribute-level fairness (vertical fairness) requires that the federated model have consistent accuracy across different attributes (e.g., physical conditions and/or sensitive demographics), i.e., $Acc_1 = Acc_2 = Acc_3$.

(C) Multilevel fairness requires that the federated model achieve horizontal and vertical fairness simultaneously.

(D) Agnostic distribution fairness requires that the federated model achieve fairness on subpopulations with unknown distributions (e.g., other hospitals that do not participate in FL), i.e., $Acc_d = Acc_e$ and $Acc_4 = Acc_5$.

fairness (also referred to as horizontal and vertical fairness, respectively), multilevel fairness, and agnostic distribution fairness, each with distinct practical implications for distribution fairness.

The primary objective for hospitals (clients) participating in FL is to obtain a model with optimal performance.[19–21] The federated model must ensure that it does not disproportionately favor or disadvantage specific hospitals. Based on experience, a FL algorithm that neglects fairness considerations may yield a model of inferior quality compared with a model solely trained on local data. Consequently, hospitals experiencing poor performance in the federated model with their data may be discouraged from participating in FL initiatives. This reluctance could significantly hinder the development of a robust FL ecosystem and impede the broader application of AI in digital health.

Moreover, fairness at the attribute level, encompassing explicit covariates (e.g., gender and race), implicit groups (e.g., domain), the target variable, or their combinations, must also be ensured. For instance, in the context of a predictive model projecting a patient's remaining lifespan, if the accuracy of the model significantly varies among patients of different races, it may lead to higher rates of incorrect treatments prescribed to certain racial groups, exacerbating social inequality.

A model violating any of the above fairness measures may lead to serious consequences; therefore, it is important to implement FL model performance fairness at multiple levels simultaneously, which is a more difficult problem than single-level fairness.

Finally, the generalizability of fairness in our medical machine learning model based on FL must be considered. This entails ensuring consistent performance not only within the existing participating hospitals but also in other hospitals with unseen distributions while maintaining fairness. By prioritizing this aspect, we can maximize social welfare in the medical field and ensure that our research findings have a meaningful impact on a broader scale.

Recently, progress in FL has led to increased research on achieving fair model performance.[22] However, most of these studies have focused on single-level fairness, particularly at the client level.[23–30] These methods can be broadly classified into two categories: personalization[31–33] and fair aggregation.[34] Personalization allows individual clients to maintain distinct local models, leveraging the diversity of other hospital data while emphasizing local data distribution. However, this approach fails to create a global model with robust generalization capabilities suitable for broad deployment. To overcome this limitation, the fair aggregation method assigns varying weights to local models of different clients during the aggregation process, ensuring consistent performance of aggregated models across diverse clients. To the best of our knowledge, while the aforementioned methods have contributed to the advancement of specific fairness in FL, limited attention has been given to unified and multilevel FL performance fairness.[35–37]

In this study, we attempted to find a solution for unified fairness in FL, with the requirement that federated models meet some or all of the above four levels of fairness. We also present a comprehensive framework called Federated Learning with Unified Fairness Objective (FedUFO). Our approach unifies the diverse levels of fairness considerations by leveraging a unified uncertainty set. Our unified framework considers existing optimization objectives as special cases, thereby providing a cohesive and encompassing perspective on fairness in FL. The model allows customizable uncertainty sets that offer users greater flexibility in managing the trade-offs between accuracy and fairness at various levels. Moreover, we address the challenging federated optimization problem by introducing a highly efficient algorithm called federated mirror descent ascent, which provides theoretical guarantees. Researchers and practitioners can gain a comprehensive understanding of performance fairness considerations in FL by adopting the proposed framework. This insight enables them to effectively assess and address fairness concerns, both coherently and meaningfully.

To validate the effectiveness of our approach, we conducted rigorous testing on four healthcare-related datasets and meticulously evaluated fairness at four distinct levels: client-level, attribute-level, multilevel, and agnostic distribution fairness. Our experimental findings clearly demonstrate that, in comparison with existing methods, our unified solution can improve fairness at the desired level(s) without any significant loss in overall accuracy. Moreover, our approach allows a flexible balance between accuracy and fairness, as well as between different fairness levels. These compelling results underscore our commitment to advancing fair and impactful FL practices with a dedicated focus on the crucial domain of medicine. Considering this, we aim to preserve the ecological integrity of FL in the medical field and actively promote social equality.

RESULTS

Unified fair FL

Optimization objective for fair FL

We aim to encourage the federated model to achieve uniform performance over subpopulations. The unified optimization objective can be written as

$$\mathcal{R}_{\text{fair}}(\theta) \; = \; \overline{\mathcal{R}}(\theta) + C\sqrt{\frac{\mathrm{Var}(\mathcal{R}^{g}(\theta))}{|\mathcal{G}|}},$$

where $\overline{\mathcal{R}}$ is the average risk, $g \in \mathcal{G}$ is the group index, $\mathcal{R}^g$ is the risk of group g, $\mathrm{Var}(\mathcal{R}^g(\theta))$ is the variance of risk across groups, and the constant C balances the utility and fairness. The first and second terms guarantee the utility of the federated model and the performance fairness for group $\mathcal{G}$, respectively. $\mathcal{R}_{\text{fair}}(\theta)$ is unified because $\mathcal{G}$ can be defined as any group as needed, and any single-level fairness in FL will be the special case of our framework.

Unified framework for fair FL

Unfortunately, $\mathcal{R}_{\text{fair}}(\theta)$ is computationally intractable because of the variance term, particularly in a federated setting.[10,38,39] To solve this problem, inspired by the techniques used for distributionally robust optimization (DRO),[40] we introduced an approximate surrogate for $\mathcal{R}_{\text{fair}}$:

$$\mathcal{R}_{\text{dro}}(\theta) := \sup_{Q \in \mathcal{Q}^g} \left\{ E_{(x,y)\sim Q}[\ell(\theta,(x,y))] \right\},$$

and

$$\mathcal{Q}^g := \left\{ \text{distribution } Q \text{ such that } D_f\left(Q \| P^{\mathcal{G}}\right) \; \le \; \rho \right\},$$

where ℓ: $\Theta \times (\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R}_+$ is a loss function, $P^{\mathcal{G}}$ is the distribution when the entire data distribution is grouped based on $\mathcal{G}$, $D_f(\,\cdot \| \cdot\,)$ is the f-divergence between distributions, ρ is the radius of the uncertainty set, and the uncertainty set $\mathcal{Q}^g$ contains the distribution shifts near distribution $P^{\mathcal{G}}$. The constant C and radius ρ are positively correlated; therefore, we can balance the model performance and fairness by varying the radius ρ. If the radius $\rho \; = \; 0$, then $\mathcal{R}_{\text{dro}}$ will degrade to empirical risk minimization (ERM) (FedAvg). Conversely, if we allow the radius to be infinite, then $\mathcal{R}_{\text{dro}}$ will degrade to minimize the risk of the worst-performing group.

Next, we show that the optimization objectives designed for diverse levels of fairness are special cases of our proposed unified risk $\mathcal{R}_{\text{dro}}$. First, if we select a client-level uncertainty set (i.e., $\mathcal{G}$ specified as client index set C), the unified objective will degrade to a client-level fairness method with a risk given by

$$\mathcal{R}_{\text{client}}(\theta) := \sup_{Q \in \mathcal{Q}^c} \left\{ E_{(x,y)\sim Q}[\ell(\theta,(x,y))] \right\},$$

and

$$\mathcal{Q}^c := \left\{ \text{distribution } Q \text{ such that } D_f\left(Q \| P^{\mathcal{C}}\right) \; \le \; \rho \right\},$$

Similarly, if we select an attribute-level uncertainty set $\mathcal{Q}^a$, the framework degrades to an attribute-level fairness FL method with a risk given by

$$\mathcal{R}_{\text{attribute}}(\theta) := \sup_{Q \in \mathcal{Q}^a} \left\{ E_{(x,y)\sim Q}[\ell(\theta,(x,y))] \right\},$$

and

$$\mathcal{Q}^a := \left\{ \text{distribution } Q \text{ such that } D_f\left(Q \| P^{\mathcal{A}}\right) \; \le \; \rho \right\},$$

where $P^{\mathcal{A}}$ is the distribution when all the data are grouped according to attribute index set $\mathcal{A}$. Moreover, the corresponding objective that simultaneously constrains both the variance of client- and attribute-level risks is given below.

$$\begin{aligned}
\mathcal{R}_{\text{fair}}(\theta) \; &= \mathcal{R}(\theta) \, + \, C_1\sqrt{\frac{\mathrm{Var}(\mathcal{R}^{c}(\theta))}{|\mathcal{C}|}} \, + \, C_2\sqrt{\frac{\mathrm{Var}(\mathcal{R}^{a}(\theta))}{|\mathcal{A}|}} \\
&= \frac{C_1}{C_1 + C_2}\left(\mathcal{R}(\theta) \, + \, (C_1 \, + \, C_2)\sqrt{\frac{\mathrm{Var}(\mathcal{R}^{c}(\theta))}{|\mathcal{C}|}}\right) \\
&+ \frac{C_2}{C_1 + C_2}\left(\mathcal{R}(\theta) \, + \, (C_1 \, + \, C_2)\sqrt{\frac{\mathrm{Var}(\mathcal{R}^{a}(\theta))}{|\mathcal{A}|}}\right)
\end{aligned}$$

This can be approximated by the following DRO-based objective:

$$\mathcal{R}_{\text{multi}}(\theta) := \sup_{Q \in \mathcal{Q}^m} \left\{ E_{(x,y)\sim Q}[\ell(\theta,(x,y))] \right\},$$

and

$$\mathcal{Q}^m := \left\{ \text{distribution } Q \text{ such that } D_f\left(Q \| P^{\mathcal{C}}\right) \; \le \; \beta\rho \textit{ or } D_f\left(Q \| P^{\mathcal{A}}\right) \; \le \; (1-\beta)\rho \right\},$$

where the uncertainty set of multilevel fairness is the union of the client-level and attribute-level uncertainty set; radius ρ balances the accuracy and the fairness; and the coefficient β balances different levels of fairness. Finally, we consider agnostic distribution fairness, which requires the federated model to be fair for subpopulations with unknown distributions. To achieve this goal, we defined a sufficiently wide uncertainty set that covers the possible distribution shifts. A natural implementation is to assign $\mathcal{G}$ as the combination of the client index and all the attributes, and then degrade the uncertainty set to the individual level, which is too wide by taking into account too much unnecessary distributional drift. Consequently, this may lead to an overly pessimistic problem in practice.[41–43] Therefore, structural constraints must be introduced on the uncertainty set to overcome this pessimism. Moreover, we specified $\mathcal{G}$ as the combination of the client index $\mathcal{C}$ and some of the sensitive attribute(s) $\{\mathcal{A}_1, \mathcal{A}_2, \cdots\}$, rather than using all the attributes. The objective of the agnostic distribution fairness can be written as

$$\mathcal{R}_{\text{unknown}}(\theta) := \sup_{Q \in \mathcal{Q}^u} \left\{ E_{(x,y)\sim Q}[\ell(\theta,(x,y))] \right\},$$

Algorithm 1. FedUFO algorithm for multilevel fairness

```
Input: The number of local iterations $E$, total number of iterations $T$, number of rounds $R = T/E$, model update stepsize $\eta$, weight update stepsize $\gamma$, initialized model parameters $\theta^{(0)}$, client weight $\{\lambda_i^{c(0)}\}$, attribute weight $\{\lambda_k^{a(0)}\}$, uncertainty set radius $\rho$, and coefficient $\beta \in [0,1]$ for balancing client and attribute-level fairness.
for $r = 0, 1, \dots, R-1$ do
  The server broadcasts $\theta^{(rE)}$, $\{\lambda_i^{c(r)}\}$, and $\{\lambda_k^{a(r)}\}$ to the corresponding clients
  for client $i = 1, 2, \dots, N$ do
    Set the local model parameters $\theta_i^{(rE)} = \theta^{(rE)}$
    for $t = rE, rE+1, \dots, (r+1)E-1$ do
      Sample data $\xi_{i,k}^{(t)}$ uniformly
      $\lambda_{i,k}^{(t)} = \lambda_i^{c(r)}$ $w.p.$ $\beta$, and $\lambda_{i,k}^{(t)} = \lambda_i^{a(r)}$ $w.p.$ $1-\beta$
      Update the model with weights: $\theta_i^{(t+1)} = \theta_i^{(t)} - \eta{\lambda_{i,k}}^{(t)}\nabla l(\theta_i^{(t)}; \xi_{i,k}^{(t)})$
      Compute the loss $v_i^{c(r)}$ of model $\theta^{(rE)}$ on the local dataset $D_i^c$
      Compute the loss ${v_{i,k}^a}^{(r)}$ of model $\theta^{(rE)}$ on each subgroup $D_{i,k}^a$
    end for
  end for
  The client $i$ sends $\theta_i^{((r+1)E)}$, $v_i^{c(r)}$, and ${v_{i,k}^a}^{(r)}$ to the server
  The server computes: $\theta^{(r+1)E} = \frac{1}{N}\sum_{i=1}^N \theta_i^{((r+1)E)}$
  The server computes: $\{(\lambda_i^c)^{(r+1)}\} = \mathrm{Proj}\left(\left\{\frac{(\lambda_i^c)^{(r)} e^{\gamma E v_i^{c(r)}}}{\sum_{i=1}^N (\lambda_i^c)^{(r)} e^{\gamma E v_i^{c(r)}}}\right\}, \rho\right)$
  The server computes: $\{(\lambda_k^a)^{(r+1)}\} = \mathrm{Proj}\left(\left\{\frac{(\lambda_k^a)^{(r)} e^{\gamma E \sum_{i=1}^N {v_{i,k}^a}^{(r)}}}{\sum_{k=1}^M (\lambda_k^a)^{(r)} e^{\gamma E \sum_{i=1}^N {v_{i,k}^a}^{(r)}}}\right\}, \rho\right)$
end for
return $\theta^{(T)}$
```

and

$$\mathcal{Q}^u := \left\{\text{distribution } Q \text{ such that } D_f\left(Q \| P^{\{\mathcal{C}, \mathcal{A}_1, \mathcal{A}_2, \cdots\}}\right) \leq \rho\right\},$$

Notably, a trade-off exists between in-distribution fairness and unknown out-of-distribution fairness because an overly conservative risk with a very wide uncertainty set usually leads to an upper bound that is too loose for in-distribution fairness. In practice, appropriate combinations can be used to balance the two. Please refer to Appendix B for more discussion on the size of uncertainty sets.

Tractable centralized optimization algorithm

Next, we developed an efficient algorithm to solve the above optimization objective $\mathcal{R}_{\text{dro}}$ in federated setting. The objective $\mathcal{R}_{\text{dro}}$ is rewritten as

$$\sup_{\lambda^g \in \Delta_{|\mathcal{G}|-1}} \left\{F(\theta, \lambda^g) := \sum_i \lambda_i^g f_i^g(\theta) \text{ s.t. } D_f(|\mathcal{G}| \cdot \lambda^g \| (1, 1, \cdots, 1)) \leq \rho\right\},$$

where $f_i^g(\theta) : = \mathbb{E}_{(x,y) \sim \widehat{P}_i^g}[\ell(\theta; (x,y))]$ is the empirical risk on i th group and $\widehat{P}_i^g$ is the empirical distribution over samples of data subset D_i^g. We can alternately optimize model parameters θ and weights λ^g to minimize the above optimization objective. We updated model $\theta_i^{(t)}$ using the stochastic gradient descent method with corresponding weight $\lambda_i^{g(t)}$ at each iteration t.

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \lambda_i^{g(t)} \nabla_\theta l\left(\theta_i^{(t)}; \left(x_i^{(t)}, y_i^{(t)}\right)\right).$$

We adopted the mirror gradient ascent method to update weight λ^g, which is expressed as

$$(\tilde{\lambda}^g)^{(t+1)} = \text{argmax}_{\lambda \in \Delta_{|\mathcal{G}|-1}} \left\{F\left(\theta^{(t+1)}, (\lambda^g)^{(t)}\right) + \langle \boldsymbol{v}^{(t)}, (\lambda^g)^{(t)} - \lambda \rangle - \frac{1}{\gamma} D_f\left(\lambda \parallel (\lambda^g)^{(t)}\right)\right\},$$

where $\gamma > 0$ is the stepsize and $\boldsymbol{v}^{(t)}$ is the gradient of weight. The first two terms are a linear approximation of $F(\theta^{(t+1)}, \lambda)$, and the last term is a Bregman distance between λ and $(\lambda^g)^{(t)}$. A suitable convex function $f(\cdot)$ can be chosen to efficiently solve for $(\tilde{\lambda}^g)^{(t+1)}$. By choosing the negative entropy function, $(\tilde{\lambda}^g)^{(t+1)}$ has an explicit solution:

$$(\tilde{\lambda}^g)^{(t+1)} = \frac{(\lambda_i^g)^{(t)} e^{\gamma v_i^{(t)}}}{\sum_{i=1}^{|\mathcal{G}|} (\lambda_i^g)^{(t)} e^{\gamma v_i^{(t)}}}.$$

After each mirror gradient ascent of weight, we compute $(\lambda^g)^{(t+1)}$ by projecting $(\tilde{\lambda}^g)^{(t+1)}$ into $\{\lambda : D_f(|\mathcal{G}| \cdot \lambda^g \| (1, 1, \cdots, 1)) \leq \rho\}$ such that the constrains of the uncertainty set for the radius will be satisfied.[44]

We obtained the stochastic mirror descent ascent (SMDA) algorithm as the solution of objective $\mathcal{R}_{\text{dro}}(\theta)$ in centralized setting, and we extended it to the decentralized setting.

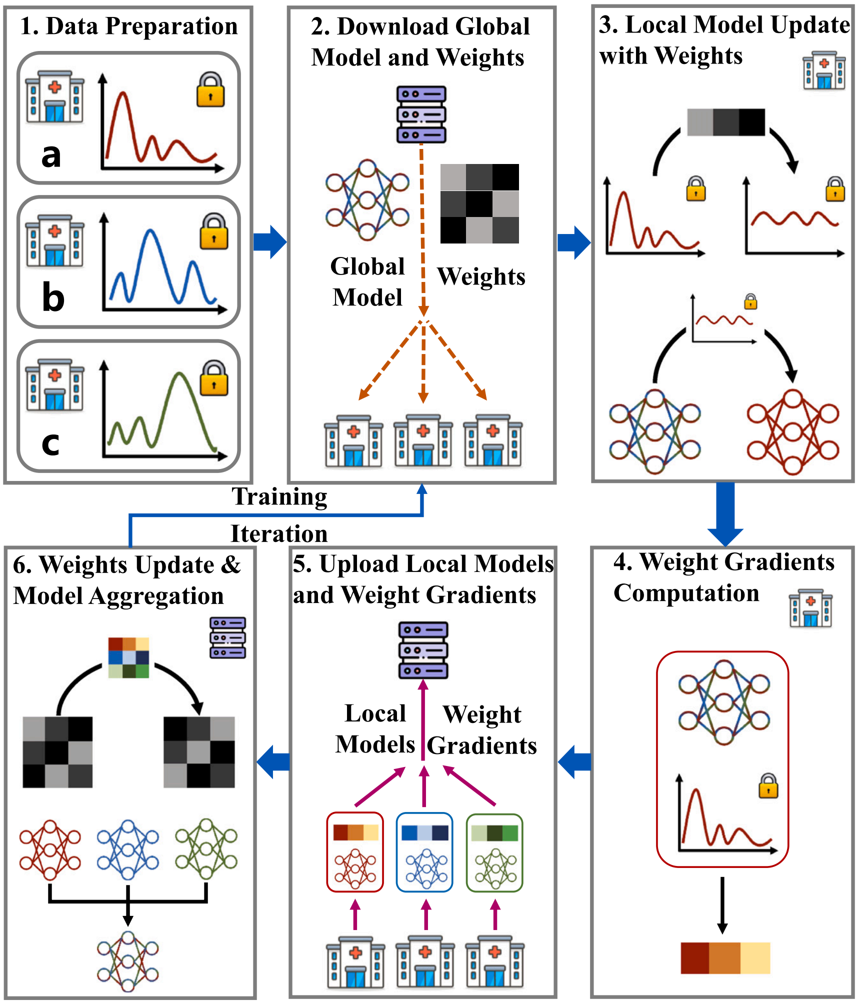

1. Data Preparation
a
b
c
2. Download Global Model and Weights
Global Model
Weights
3. Local Model Update with Weights
Training
Iteration
4. Weight Gradients Computation
5. Upload Local Models and Weight Gradients
Local Models
Weight Gradients
6. Weights Update & Model Aggregation

Figure 2. FedUFO training workflow

Along with the conventional model aggregation process in standard federated learning, we incorporate an additional step to enhance model fairness by maintaining a dedicated set of weights. After training, the federated model with fairness guarantees can be deployed at different hospitals to assist in the medical diagnosis of patients belonging to different demographics.

Efficient federated optimization algorithm

In FL, each client accesses its own local data. In each communication, the clients use their local data to update local models with the corresponding weights, and then the server aggregates the local models into a global model. In addition, clients must compute the partial gradients of the weights and send them to the server for updating the weights. The key challenge is that high-frequency communication is not allowed in FL because of communication costs. Therefore, we adopted the snapshot mechanism for both the model and weight updates. Specifically, we updated the weights as follows:

$$\left\{\left(\lambda_i^g\right)^{(r+1)}\right\} = \mathrm{Proj}\left(\left\{\frac{\left(\lambda_i^g\right)^{(r)} e^{\gamma E v_i^{g(r)}}}{\sum_{i=1}^{|\mathcal{G}|}\left(\lambda_i^g\right)^{(r)} e^{\gamma E v_i^{g(r)}}}\right\}, \rho\right),$$

where r is the current round of communication and E is the number of local iterations at each communication. The above equations can be viewed as an unbiased estimation of $\left(\tilde{\lambda}^g\right)^{(t+1)}$. We allowed multiple iterations of the model parameters and weights in a single communication.

This completes the federated mirror descent ascent (FedUFO) algorithm. Taking multilevel fairness as an example, we present the details of FedUFO in Algorithm 1. Please refer to Appendix A for the proof of convergence of Algorithm 1.

An overview of the workflow is shown in Figure 2. In each communication, hospitals first download the global model and weights from the server and then update the local models with their private data with the corresponding weights. After computing the gradients of the weights, hospitals send them and the local models to the server. Then, the server aggregates the local models into a global model and updates the weights. These steps are repeated until the global model converges. The trained federated model can be deployed in various locations to aid in medical decisions. We emphasize that the proposed unified solution is general and can be specified to guarantee a given level of fairness. We would like to clarify here that our contributions focus on the diverse levels of fairness in FL, while the privacy protections of the proposed unified framework are entirely inherited from the FL paradigm itself.

Main results

Evaluation metrics

Our aim is to guarantee that the model's performance is good and fair across subpopulations. Suppose that the full dataset D is divided into $|\mathcal{G}|$ groups, $D = \{D_1^g, D_2^g, \ldots, D_{|\mathcal{G}|}^g\}$. We first define the disparity of an FL model across groups $\{D_i^g | i = 1, 2, \ldots, |\mathcal{G}|\}$ as

$$Disparity = \sqrt{\frac{1}{|\mathcal{G}| - 1}\sum_{i=1}^{|\mathcal{G}|}\left(Acc\left(D_i^g\right) - Avg\_Acc\right)^2},$$

where $Acc(D_i^g)$ is the predictive accuracy for group D_i^g and $Avg\_Acc = \frac{1}{|\mathcal{G}|}\sum_{i=1}^{|\mathcal{G}|} Acc(D_i^g)$. In this study, following the difference principle of distributive justice and stability,[45] we view the performance of the federated model as a resource that should be allocated fairly among the various groups. Specifically, we

Table 1. Client-level fairness: Disparity over clients[a]

	Heterogeneity	w.o. FL	Federated baselines					Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FairFed[b]	Poulain's[b]	FedUFO[c]	Global
Fetal	10,000	0.0324	0.0205	0.0121	0.0245	–	–	0.0081[c]	0.0288
Prostate	10	0.0162	0.0240	0.0397	0.0206	–	–	0.0071[c]	0.0199
COVID-19	0.5	0.0109	0.0306	0.0122	0.0194	0.0144	0.0114	0.0010[c]	0.0120
Support	0.1	0.0400	0.1026	0.0251	0.0121	0.0913	0.0339	0.0065[c]	0.0275

[a] Lower numbers are better.

[b] The federated baselines, FairFed and Poulain's, only support settings of binary classification.

[c] Best federated learning results.

measured fairness using disparity. The smaller the value of disparity, the fairer the FL model. We focus on four levels of fairness in FL: client-level, attribute-level, multilevel, and agnostic distribution fairness. We measured client-level fairness using $\sqrt{\frac{1}{|\mathcal{C}|-1}\sum_{i=1}^{|\mathcal{C}|}\left(Acc\left(D_i^c\right) - Avg\_Acc\right)^2}$, where $\mathcal{C}$ is a set of client indices. Given a set of protected attributes $\mathcal{A}$, the attribute-level fairness can be measured using $\sqrt{\frac{1}{|\mathcal{A}|-1}\sum_{i=1}^{|\mathcal{A}|}\left(Acc\left(D_i^a\right) - Avg\_Acc\right)^2}$. For multilevel fairness, we used the harmonic average of client-level fairness and attribute-level fairness as evaluation metrics. Notably, the harmonic average is just one of the indicators used to measure fairness at multiple levels. Different fairness metrics for various levels are presented in detail in Tables S1 and S2. Following the previous studies,[46–48] we used a Dirichlet distribution with hyperparameter α to define data heterogeneity among different clients in FL. A lower value of α means strong data heterogeneity, and we set different values of α for different datasets to evaluate how our method performs at different degrees of client drift. For agnostic distribution fairness, we first trained a federated model with a specific value of α and then performed a simulation to test its multilevel fairness in federated settings for different values of α. Additionally, we used the accuracy, *Acc*, to measure the utility of the FL model. As part of our main results, we report the overall and worst-case performances for different scenarios.

Client-level fairness

The experimental results for client-level fairness are presented in Tables 1 and 2, respectively. We drew several conclusions based on these results. First, compared with local training (without [w.o.] FL) on the COVID-19 dataset, our solution decreased the disparity over clients from 0.0109 to 0.0010 (i.e., the fairness was improved by $\frac{0.0109-0.0010}{0.0109} = 90.83\%$). Similarly, our solution, FedUFO, also improved fairness by 75.00%, 56.17%, and 83.75% for the fetal, prostate, and support datasets, respectively. Second, by comparing the federated baselines, we achieved a state-of-the-art performance. Taking the results for the support dataset as an example, the traditional federated algorithm FedAvg yields an unfair federated model with a disparity of 0.1026. Moreover, its accuracies for two hospitals were found to be 68.80% and 54.29%, indicating that the model was 14.51% less accurate for one hospital than the other. Our solution, FedUFO, limits the gap to 0.92% and significantly promotes fairness among hospitals. Third, compared with all the baselines, our client-level solution improved the worst client performance when tested on four different datasets, as shown in Table 2. Particularly, the accuracy of the proposed FedUFO improved by more than 10% compared with FedAvg in terms of the worst-performing client. In conclusion, our proposed FedUFO significantly improves client-level fairness, which helps prevent some hospitals from deploying federated models that perform poorly for their data distribution. Furthermore, this will encourage more hospitals to participate in FL and promote its wider application in the field of digital health.

Attribute-level fairness

The experimental results for attribute-level fairness are presented in Tables 3 and 4. An attribute can be specified as any variable (e.g., target variable, sensitive attribute, and their combinations). We chose the target variable, death or not, as the attribute for the COVID-19 dataset, and we expect that the federated model will satisfy the accuracy parity (AP) (i.e., similar false positives and false negatives). However, the false-negative and false-positive rates of the federated model trained by FedAvg were 82.67% and 45.89%, respectively (with disparity of 0.2601). The above results indicate that the federated model trained by FedAvg predicts that patients will not die, as the false-negative rate was 36.78% higher than the false-positive rate. If such a model is deployed on a large scale, many patients will die because of the lack of timely treatment, which will lead to

Table 2. Worst-case performance in client-level fairness: *Acc* for the worst client[a] (%)

	Heterogeneity	w.o. FL	Federated baselines					Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FairFed[b]	Poulain's[b]	FedUFO$_c$	Global
Fetal	10,000	87.72	95.32	94.15	92.98	–	–	96.49[c]	94.74
Prostate	10	71.04	79.11	77.41	79.22	–	–	79.98[c]	83.93
COVID-19	0.5	58.12	62.85	62.15	62.15	61.58	62.01	64.12[c]	64.97
Support	0.1	57.14	54.29	62.86	64.00	54.29	60.00	64.80[c]	65.71

[a] Higher numbers are better.

[b] The federated baselines, FairFed and Poulain's, only support settings of binary classification.

[c] Best federated learning results.

Table 3. Attribute-level fairness: Disparity over attributes[a]

	Heterogeneity	w.o. FL	Federated baselines					Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FairFed[b]	Poulain's[b]	FedUFO$_a$	Global
Fetal	10,000	0.1316	0.0778	0.0752	0.0643	–	–	0.0552[c]	0.0591
Prostate	10	0.3040	0.2420	0.2031	0.3175	–	–	0.1192[c]	0.2303
COVID-19	0.5	0.4716	0.2601	0.0660	0.1549	0.0236	0.0194	0.0174[c]	0.3424
Support	0.1	0.0487	0.1078	0.1235	0.0905	0.0637	0.0320	0.0134[c]	0.0349

[a] Lower numbers are better.
[b] The federated baselines, FairFed and Poulain's, only support settings of binary classification.
[c] Best federated learning results.

serious social problems. Fortunately, the proposed solution can reduce the accuracy gap by less than 3% and, therefore, promote social fairness. Our algorithm, FedUFO, also achieved the lowest disparity and highest *Acc* of the worst attributes for the other datasets, demonstrating the effectiveness of our solution. Moreover, FedUFO can also avoid discrimination against specific races or genders. We present these results and discuss them in detail later.

Multilevel fairness

In practice, breaches at any level of fairness can raise serious ethical issues; therefore, the federated model must be encouraged to be fair at multiple levels simultaneously. Specifically, we considered both client-level fairness and attribute-level fairness and used their harmonic average as the evaluation metric for multilevel fairness. Tables 5 and 6 list the experimental results for multilevel fairness. More detailed experimental results (client- and attribute-level metrics) are provided in the supplemental information. We can observe that FedUFO achieves the best FL results for all datasets. Notably, our solution is flexible because we allow users to balance the measure for multiple levels of fairness according to their needs by setting a hyperparameter. We present the results of the trade-offs in a later section.

Agnostic distribution fairness

To maximize social and medical benefits, powerful models derived from FL should be widely applicable. Therefore, the model should be deployable in hospitals that do not participate in FL and whose data have an unknown distribution. The experimental results for agnostic distribution fairness are listed in Tables 7 and 8. We set the heterogeneity hyperparameter α to 10, 5, and 2 during training and tested the trained federated model using $\alpha = 1$ by simulating a variety of differences between training environments and deployment environments. Although the performance of other methods greatly fluctuates with the degree of heterogeneity, our solution shows strong stability. Empirically, the results illustrate that FedUFO can effectively improve the fairness and worst-case performance (in terms of the harmonic average) of federated models, even if the target distributions are unknown.

Model utility

A fair but poorly performing model is meaningless.[49] For example, if a model has an accuracy of zero for all groups, then the model meets the fairness requirement but does not contribute to medical decisions. Therefore, we also focused on model utility measured by overall accuracy (in terms of macro-averaged performance weighted by sample size). Our solutions achieved FL results comparable with the standard FL algorithm FedAvg (without consideration of fairness), as shown in Table 9.

Trade-off analysis

The general framework we proposed can not only constrain any given level(s) of fairness in FL but also allow users the flexibility to make trade-offs among different goals based on their own needs. Specifically, users can balance (1) accuracy and fairness, (2) multiple levels of fairness, and (3) in-distribution fairness and out-of-distribution fairness. We ran the analytical experiments on the COVID-19 dataset with a heterogeneity of $\alpha = 0.1$, and the experimental results are shown in Figure 3.

Trade-off between accuracy and single-level fairness

First, we can balance the model's utility and fairness via the uncertainty set radius ρ. An uncertainty set with a larger radius ρ considers more potential distribution shifts; thus, it can provide a fairness guarantee for the worst cases. However, this also places a looser upper bound on the ERM, leading to a decrease in accuracy.

As shown in Figures 3A and 3B, when we decrease the radius ρ from 1e$-$1 to 1e$-$8, the disparity among clients and attributes increases while the accuracy is improved.

Table 4. Worst-case performance in attribute-level fairness: *Acc* for the worst attribute[a] (%)

	Heterogeneity	w.o. FL	Federated baselines					Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FairFed[b]	Poulain's[b]	FedUFO$_a$	Global
Fetal	10,000	62.50	75.00	78.57	81.25	–	–	86.49[c]	83.33
Prostate	10	20.89	17.71	31.49	18.55	–	–	56.56[c]	19.37
COVID-19	0.5	25.54	45.89	58.12	52.10	60.65	61.43	62.52[c]	41.52
Support	0.1	57.89	60.19	54.39	56.14	61.17	62.14	63.16[c]	66.99

[a] Higher numbers are better.
[b] The federated baselines, FairFed and Poulain's, only support settings of binary classification.
[c] Best federated learning results.

Table 5. Multilevel fairness: Harmonic average[a] of disparity over clients and disparity over attributes[b]

	Heterogeneity	w.o. FL	Federated baselines					Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FairFed[c]	Poulain's[c]	FedUFO$_m$	Global
Fetal	10,000	0.0260	0.0162	0.0104	0.0177	–	–	0.0003[d]	0.0194
Prostate	10	0.0154	0.0218	0.0332	0.0193	–	–	0.0136[d]	0.0183
COVID-19	0.5	0.0107	0.0274	0.0103	0.0172	0.0089	0.0072	0.0016[d]	0.0116
Support	0.1	0.0220	0.0526	0.0209	0.0107	0.0375	0.0292	0.0021[d]	0.0154

[a] Disparity over clients and disparity over attributes are shown in the supplemental information (Table S1).

[b] Lower numbers are better.

[c] The federated baselines, FairFed and Poulain's, only support settings of binary classification.

[d] Best federated learning results.

Trade-off between multiple levels of fairness

We also set a hyperparameter $\beta \in [0,1]$ to balance client-level and attribute-level fairness. We varied the value of β from 0.1 to 0.9, as shown in Figure 3C, and found that client-level fairness benefits from a large coefficient β, whereas attribute-level fairness benefits from a small β.

Trade-off between in-distribution fairness and out-of-distribution distribution fairness

A wide uncertainty set can be used to improve the generalization ability of fairness. However, an overly wide uncertainty set results in overly pessimistic problems, resulting in a lot of accuracy sacrifice. Empirically, we considered three uncertainty sets of different sizes formed by the client index, attribute, and uncertainty set radius. For a small uncertainty set, we used the union of the client-level uncertainty set and attribute-level uncertainty set with $\beta = 0.5$. For the two large uncertainty sets, formed by the combination of client index and target variable, we set the uncertainty set radius ρ as 1e$-$7 and 1e$-$1, respectively. The large uncertainty set yields a model with good out-of-distribution fairness but compromises in-distribution fairness. The empirical evidence supports our analysis, as shown in Figure 3D. We recommend the construction of an uncertainty set based on specific requirements and expert knowledge.

We wish to elucidate that the flexibility extended here aims to offer users a realm of choice, enabling them to judiciously forego certain indicators of lesser importance to them in favor of enhancing those they deem significant, based on their actual needs. The extent of alterations in the ultimate outcome of the indicator is, to a degree, shaped by the pre-set data distribution.

Relation with other fairness notions

In this study, our aim was to encourage the federated model to have a low accuracy disparity among different groups. The definition of fairness in this paper is different from traditional algorithmic fairness, which requires the independence of model decisions and sensitive attributes.[47,50–56] However, we demonstrated that our proposed unified fairness notions can also be used to improve algorithmic fairness. Specifically, equal opportunity (EO),[57] the most commonly used metric for measuring fairness (in terms of discrimination against certain groups), can be viewed as a relaxed version of fairness. The EO requires the following:

$$P(Y=1|\widehat{Y}=1, A=0) = P(Y=1|\widehat{Y}=1, A=1),$$

If we specify S as a combination of the target label and sensitive attribute, our proposed unified fairness notion requires that

$$\begin{aligned} &P(Y=\widehat{Y}|\widehat{Y}=0, A=0) = P(Y=\widehat{Y}|\widehat{Y}=0, A=1) \\ &= P(Y=\widehat{Y}|\widehat{Y}=1, A=0) = P(Y=\widehat{Y}|\widehat{Y}=1, A=1), \end{aligned}$$

Therefore, our optimization objective provides an upper bound for EO.

We empirically evaluated the models trained using different algorithms in terms of EO. The sensitive attribute was race, and the target variable was treatment outcome. FairFed[47] and Poulain's FairFedAvg[48] are two state-of-the-art methods that aim to mitigate discrimination (in terms of sensitive attributes) in FL. Following the setting in the previous work,[48] we set the heterogeneity hyperparameter $\alpha = 1$. We also report the worst-case performance, which was measured using the following metrics:

$$\mathrm{WorstTPR} = \min_n \Pr(\widehat{Y}=1|A=n, Y=1).,$$

The experimental results are listed in Table 10. We can observe that our solution, FedUFO, achieves the best results in terms of EO and WorstTPR, while maintaining comparable

Table 6. Worst-case in multilevel fairness: Harmonic average[a] of Acc for the worst client and Acc for the worst attribute[b] (%)

	Heterogeneity	w.o. FL	Federated baselines					Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FairFed[c]	Poulain's[c]	FedUFO$_m$	Global
Fetal	10,000	72.99	83.95	85.66	86.72	–	–	91.76[d]	88.67
Prostate	10	32.29	28.94	44.77	30.06	–	–	64.84[d]	31.48
COVID-19	0.5	35.49	53.05	60.07	56.68	61.11	61.72	63.17[d]	50.66
Support	0.1	57.51	57.09	58.32	59.81	57.53	61.05	62.50[d]	66.34

[a] *Acc* for the worst client and *Acc* for the worst attribute are shown in the supplemental information (Table S1).

[b] Higher numbers are better.

[c] The federated baselines, FairFed and Poulain's, only support settings of binary classification.

[d] Best federated learning results.

Table 7. Agnostic distribution fairness: Harmonic average[a] of Disparity over clients and Disparity over attributes[b]

	Heterogeneity[c]	w.o. FL	Federated baselines			Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FedUFO$_u$	Global
	10	0.0604	0.0568	0.0507	0.0495	0.0423[d]	0.0333
Prostate	5	0.0567	0.0676	0.0609	0.0668	0.0487[d]	0.0566
	2	0.0543	0.0556	0.0455	0.0450	0.0415[d]	0.0379

[a] Disparity over clients and disparity over attributes are shown in the supplemental information (Table S2).
[b] Lower numbers are better.
[c] We trained the federated models under various degrees of heterogeneity (including α = 10, 5, and 2) and evaluated the models for α = 1.
[d] Best federated learning results.

overall performance with the standard FL algorithm, FedAvg (without consideration of fairness). These results illustrate that the fairness framework is unified and flexible.

We recommend setting S on the basis of expert knowledge in practice.

Sensitivity analysis for FL

Our method is robust to changes in the hyperparameters of the FL settings. The general framework of our method is not sensitive to variations in the number of clients or local iterations. In this section, we modify these two hyperparameters and test the performance of our model on a fetal dataset.

Number of clients N

Our model was robust to the change in the number of clients. We set the heterogeneity hyperparameter α of the dataset to 10 and the number of clients to 4, 6, 8, and 10 and recorded the variation in accuracy and fairness to investigate the sensitivity to the change in the number of clients in our model. As shown in Figure 4A, the accuracy of our model fluctuates only slightly, within a range of 5%, around 79%, and the fairness of our model holds up to 0.013 when the number increases from four to six.

Number of local iteration epochs E

Our model is insensitive to the number of local iterations. We set the heterogeneity hyperparameter α of the dataset to 1 and changed the number of iterations to 1, 3, 5, and 10 to verify the stability of the performance of our model. In Figure 4B, the range of accuracy is approximately 3% and that of fairness is less than 0.001, indicating that our model maintains high performance and a high level of fairness despite variation in the number of local iterations.

DISCUSSION

This study investigated fairness considerations in the realm of FL, particularly within the domain of digital healthcare. Our aim was to establish a consistent performance standard across various subpopulations using the federated model. To achieve this, we developed a unified framework for fair FL that offers adaptability for enhancing fairness at multiple levels, ranging from client- and attribute-level fairness to multilevel fairness and fairness extensions to uncharted data distributions, based on user requirements. In addition, we introduced an efficient optimization algorithm tailored for FL, which is essential for implementing the aforementioned framework and has been substantiated by comprehensive theoretical analyses.

We conducted extensive experimentation on four real-world medical datasets, encompassing a spectrum of FL scenarios. When compared with a diverse array of advanced federated baseline methodologies, our approach consistently demonstrated superior fairness outcomes at the desired fairness levels in most instances. Additionally, we elucidated the adaptability inherent in our framework through experimental scrutiny, showcasing its capacity to enable models to strike an optimal balance between fairness and accuracy. This adaptability also permits trade-offs in fairness levels across varying strata.

Moreover, our investigation addresses the issue of potential bias or discrimination arising from federated models. Theoretically, we posit that mitigating such concerns can be approached as a sub-problem inherent in the broader, unified framework that we propose. Empirical assessments demonstrated the effectiveness of our approach in mitigating discrimination, accomplished through the prudent selection of uncertainty set ranges while maintaining model performance comparable with the standard FL benchmark, FedAvg.

However, the proposed approach has certain limitations. For instance, the precise determination of uncertainty set ranges prior to model training poses a challenge, despite its pronounced influence on model outcomes. In practical applications, we recommend that users integrate domain expertise with

Table 8. Worst-case performance in agnostic distribution fairness: Harmonic average[a] of *Acc* for the worst client and *Acc* for the worst attribute[b]

	Heterogeneity[c]	w.o. FL	Federated baselines			Ours	Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FedUFO$_u$	Global
	10	33.49	50.99	55.57	49.87	67.18[d]	47.14
Prostate	5	40.71	57.11	43.66	59.86	66.85[d]	63.81
	2	36.50	55.28	58.76	58.59	68.77[d]	54.57

[a] *Acc* for the worst client and *Acc* for the worst attribute are shown in the supplemental information (Table S2).
[b] Higher numbers are better.
[c] We trained the federated models under various degrees of heterogeneity (including α = 10, 5, and 2) and evaluated the models for α = 1.
[d] Best federated learning results.

Table 9. Overall performance: Macro-averaged accuracy[a] (%)

	Heterogeneity	w.o. FL	Federated baselines					Ours				Centralized
Dataset	α	Local	FedAvg	AFL	q-FedAvg	FairFed[b]	Poulain's[b]	FedUFO$_c$	FedUFO$_a$	FedUFO$_m$	FedUFO$_u$	Global
Fetal	10,000	90.00	96.76	95.00	94.71	–	–	97.06	94.71	96.47	95.41	96.76
Prostate	10	72.18	80.74	79.48	80.29	–	–	80.46	80.48	80.60	80.19	85.28
COVID-19	0.5	58.67	64.40	62.76	63.12	62.31	62.58	64.21	63.76	63.49	63.21	65.58
Support	0.1	61.56	65.62	65.62	64.37	64.37	63.75	65.00	64.38	63.13	64.28	68.75

[a] Higher numbers are better.
[b] The federated baselines, FairFed and Poulain's, only support settings of binary classification.

predetermined hyperparameter settings for the uncertainty set and test the parameters on a limited subset of data before embarking on large-scale training. We expect that further investigation into determination of the uncertainty set will be conducted in future studies.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information, questions, and requests should be sent to Kun Kuang (kunkuang@zju.edu.cn).

Materials availability

This study did not involve any physical materials.

Data and code availability

Our source code is available at GitHub (https://github.com/Zitao-Shuai/FedUFO) and has been archived at Zenodo.[58]

Datasets

We conducted our experiments on four medical datasets: (1) prostate cancer datasets from the US (prostate),[59] (2) a fetal state dataset of cardiotocography (fetal),[60] (3) a COVID-19 dataset of Brazilian patients (COVID-19);[61] and (4) a support dataset of seriously ill hospitalized adults (support).[62]

All the datasets were tabular and were used for classification. (1) The prostate dataset has been widely used for forecasting tumor types. We selected features, such as age at diagnosis, race, sex, year of diagnosis, site, morphology group, and therapy group, in the period 2017–2020, from the Surveillance, Epidemiology, and End Results (SEER) database. For data processing, we used the site recode ICD-O-3/WHO 2008 as the target variable. We recoded the top nine attributes based on the number of samples as nine classes and recoded the remaining attributes as the tenth class. Therefore, we constructed a 10-classification task. Then, we transformed the other input features into the one-hot coding form and dropped the data records with null value. Ultimately, 287,237 data records and 66 features were obtained. For the remaining datasets, we followed the method of data processing described by Seedat et al.[63] and dropped the rows with null values. (2) The fetal cardio dataset is a 10-classification dataset used to identify cardiovascular diseases. Ultimately, 2,123 data records and 35 features were obtained. (3) The COVID-19 dataset is a binary classification dataset used to analyze the relationship between several factors and deaths caused by COVID-19.[64] For this dataset, 6,882 data records and 44 features were obtained. (4) The support dataset is another popular dataset used in digital medical analyses. For this dataset, we obtained 1,000 data records and 30 features.

The fairness of algorithms for sensitive attributes (e.g., age, sex, and race) is an important concern. In this study, we took sex as an example and ran comprehensive experiments on the prostate and COVID-19 datasets to evaluate the performance of different algorithms in terms of fairness.

For training and evaluation, the dataset was randomly divided into two parts. The first part accounted for 20% of the total dataset and was used to evaluate the equalized opportunity of the models. The second part was used as the primary dataset for subsequent experiments. Given N clients, we used the Latent Dirichlet Allocation (LDA)[65,66] algorithm to divide the second part into N parts for each client. The non-IID (independent and identically distributed) degree of these datasets was controlled by the hyperparameter α of the algorithm. For each client, we randomly divided the dataset into sets with 80% and 20% of the data. The set with 80% of the data was used as the training set, and the set with 20% of the data was used for testing.

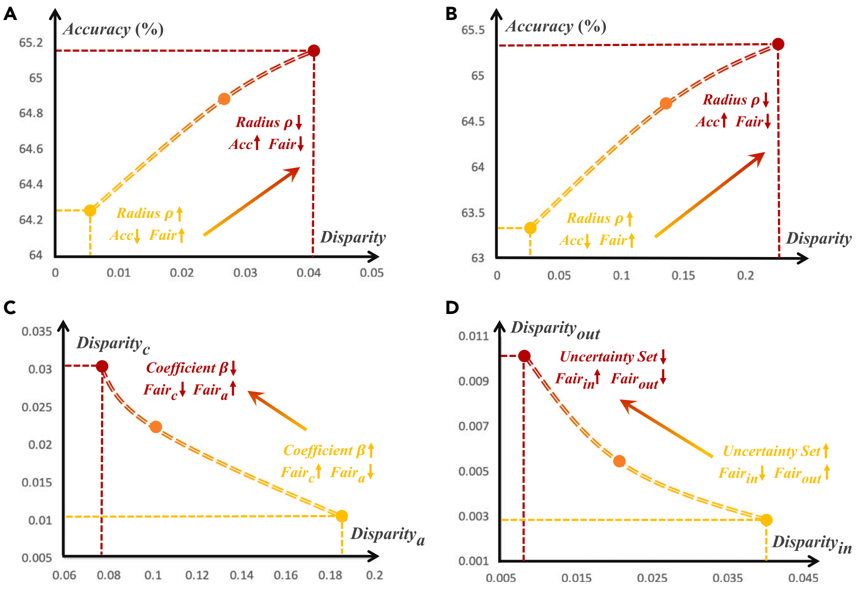


Figure 3. Trade-offs experimental results on COVID-19 dataset

Heterogeneity α = 0.1. Our solution FedUFO is flexible and can balance different goals.

(A) Trade-off between client-level fairness and accuracy.

(B) Trade-off between attribute-level fairness and accuracy.

(C) Trade-off between client-level fairness and attribute-level fairness.

(D) Trade-off between in-distribution fairness and out-of-distribution fairness.

Table 10. Federated results of discrimination against specific sensitive attribute

		w.o. FL	Federated baselines					Ours	Centralized
Dataset	Metrics	Local	FedAvg	AFL	q-FedAvg	FairFed	Poulain's	FedUFO$_a$	Global
COVID-19	EO[a]	0.300	0.318	0.267	0.299	0.211	0.174	0.160[d]	0.265
	WorstTPR[b] (%)	12.17	36.12	28.52	19.39	49.80	47.15	61.60[d]	39.92
	Overall *Acc*[c] (%)	57.49	62.34[d]	59.17	58.89	59.34	59.17	61.52	66.24

[a] Lower numbers are better.
[b] Higher numbers are better.
[c] Higher numbers are better.
[d] Best federated learning results.

The detailed breakdown of the distribution of each dataset is shown in Table 11.

Compared methods

We used the following baselines as our comparison methods: FedAvg,[10] AFL,[24] q-FedAvg,[23] FairFed,[47] and Poulain's FairFedAvg.[48] FedAvg is a classic baseline of the FL domain that simply aggregates local models with equal weights. AFL and q-FedAvg are two popular FL methods aiming to alleviate unfairness at the client level, while FairFed[47] and Poulain's FairFedAvg[48] are two state-of-the-art attribute-level methods. We also trained local models on the clients' own training datasets (called local). Additionally, we trained a model on the union of the training dataset for each client in a centralized setting (called global).

Experimental settings

Model structure and hyperparameter setting

For the smaller datasets, COVID-19, fetal, and support, we utilized a 64-dimension fully connected (FC) layer as our backbone network. We set the learning rate between 5e−1 and 5e−4 for the global method and 0.01 for the other methods, and the batch size to 32. The number of local training epochs, *E*, was set to five. For the Prostate dataset, we utilized a 256-dimensional FC layer as the backbone network. We set the learning rate of the global method to 5e−3 and 1e−3 for the other methods, and the batch size to 128. The number of epochs of the local iterations was set to two. The number of communications for each dataset was set to ten. All the methods shared the same network structure. We used the Adam[67] algorithm as the optimizer in our experiments. For the main experiment, the γ value for all the methods ranged from 1 to 10 for the prostate and support datasets and from 1e−6 to 1e−2 for the fetal and COVID-19 datasets. In the main experiment, the β value used to measure the attribute-level and client-level fairness was set to 5e−1, and the radius was set to 1e−4.

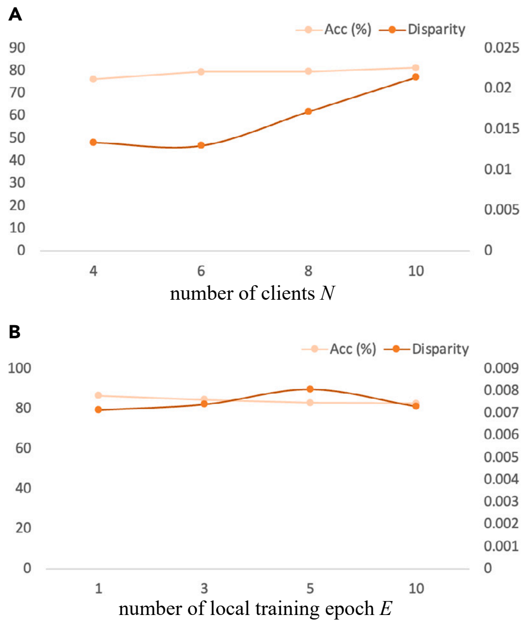


Figure 4. Sensitivity analysis experimental results

(A and B) Our solution, FedUFO, is robust to variations in the number of clients *N* and the number of local iterations *E*.

Model selection and evaluation

For each method, we restored the model at each checkpoint after five communications and selected the model for evaluation based on its total loss in all training datasets. We used the standard deviation of the accuracy of model on different clients to measure client-level fairness and the standard deviation of the accuracy of model on different attributes to measure attribute-level fairness. For multilevel fairness, we used the harmonic average of client-level fairness and attribute-level fairness as the evaluation metric. To measure agnostic fairness, we split a given dataset into two different non-independent and identically distributed (i.i.d.) degrees. First, we trained the models under one of the partitions in a federated setting and then evaluated the aggregated model for the other partition.

Supplemental information

Appendix A presents the convergence rates of the proposed FedUFO algorithm and its theoretical guarantees. In Appendix B, we discuss the effect of the size of the uncertainty set in FedUFO on the fairness and performance of the federated model, showing why we used an uncertainty set that is considerably large and may lead to failure theoretically. In Appendix C, we provide additional experimental results for multilevel fairness (Table S1) and agnostic distribution fairness (Table S2) to demonstrate the effectiveness of our solution, FedUFO.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100907.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Project of China (2021ZD0110700) and the National Natural Science Foundation of China (U19B2043, 62006207).

AUTHOR CONTRIBUTIONS

F.Z., Z.S., K.K., F.W., Y.Z., and J.X. conceived the study. The methods used in the study were developed by F.Z. and K.K. Z.S. conducted the experiments. All the authors contributed to the methodology and drafted the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Table 11. Detailed breakdown of the distribution of datasets

Fetal state dataset of cardiotocography: Training distribution

α = 10,000	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5	Attr 6	Attr 7	Attr 8	Attr 9	Attr 10
Client 1	126	179	17	24	20	109	80	35	23	63
Client 2	122	193	19	26	22	106	81	31	21	63

Fetal state dataset of cardiotocography: Test distribution

α = 10,000	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5	Attr 6	Attr 7	Attr 8	Attr 9	Attr 10
Client 1	27	53	5	8	6	24	18	6	3	19
Client 2	31	40	5	6	6	30	19	10	5	19

Prostate cancer datasets from the US: Training distribution

α = 10	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5	Attr 6	Attr 7	Attr 8	Attr 9	Attr 10	Attr 11
Client 1	23,632	7,288	5,697	5,265	6,127	5,074	3,673	3,809	3,021	3,270	28,575
Client 2	31,458	6,305	5,549	5,855	3,280	2,777	2,329	1,753	2,351	1,306	25,437

Prostate cancer datasets from the US: Test distribution

α = 10	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5	Attr 6	Attr 7	Attr 8	Attr 9	Attr 10	Attr 11
Client 1	5,789	1,847	1,438	1,375	1,507	1,274	941	1,009	782	820	7,076
Client 2	7,740	1,690	1,380	1,405	846	722	575	392	607	344	6,399

COVID-19 dataset: Training distribution			COVID-19 dataset: Test distribution		
α = 0.5	Attr 1	Attr 2	α = 0.5	Attr 1	Attr 2
Client 1	1,500	1,332	Client 1	359	349
Client 2	834	738	Client 2	195	198
Support dataset: Training distribution			Support dataset: Test distribution		
α = 0.1	Attr 1	Attr 2	α = 0.1	Attr 1	Attr 2
Client 1	171	329	Client 1	49	76
Client 2	35	105	Client 2	8	27

Received: August 3, 2023
Revised: September 1, 2023
Accepted: December 4, 2023
Published: December 28, 2023

REFERENCES

1. Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P.L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., and Firmin, D. (2018). Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. IEEE Trans. Med. Imag. *37*, 1310–1321.
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., and Davatzikos, C. (2017). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Sci. Data *4*, 170117.
3. Li, X., Dvornek, N.C., Zhou, Y., Zhuang, J., Ventola, P., and Duncan, J.S. (2019). Graph neural network for interpreting task-fmri biomarkers. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, Proceedings, Part V 22 (Springer), pp. 485–493.
4. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., and Duncan, J.S. (2021). Braingnn: Interpretable brain graph neural network for fmri analysis. Med. Image Anal. *74*, 102233.
5. Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V.C., Shi, L., and Heng, P.A. (2016). Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. IEEE Trans. Med. Imag. *35*, 1182–1195.
6. Wang, W.-Y., Zhou, H., Wang, Y.-F., Sang, B.-S., and Liu, L. (2021). Current policies and measures on the development of traditional chinese medicine in China. Pharmacol. Res. *163*, 105187.
7. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. NPJ Digit. Med. *3*, 119.
8. Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., and Duncan, J.S. (2020). Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. Med. Image Anal. *65*, 101765.
9. Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R., and Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. *10*, 12598.
10. 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 (PMLR), pp. 1273–1282.
11. Zhu, W., Luo, J., and White, A.D. (2022). Federated learning of molecular properties with graph neural networks in a heterogeneous setting. Patterns *3*, 100521.
12. Sav, S., Bossuat, J.-P., Troncoso-Pastoriza, J.R., Claassen, M., and Hubaux, J.P. (2022). Privacy-preserving federated neural network learning for disease-associated cell classification. Patterns *3*, 100487.
13. Li, Z., Mao, F., and Wu, C. (2022). Can we share models if sharing data is not an option? Patterns 3, 100603. https://www.sciencedirect.com/science/article/pii/S2666389922002288.
14. Li, X., Zhao, S., Chen, C., and Zheng, Z. (2023). Heterogeneity-aware fair federated learning. Inf. Sci. *619*, 968–986.
15. Li, Z., Shang, X., He, R., Lin, T., and Wu, C. (2023). No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. Preprint at arXiv.
16. Zhou, Z., Chu, L., Liu, C., Wang, L., Pei, J., and Zhang, Y. (2021). Towards fair federated learning. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 4100–4101.

17. Zhao, Y., Li, M., Lai, L.N., Suda, N., Civin, D., and Chandra, V. (2018). Federated learning with non-iid data. Preprint at arXiv.

18. Yu, F., Rawat, A.S., Menon, A., and Kumar, S. (2020). Federated learning with only positive labels. In International Conference on Machine Learning (PMLR), pp. 10946–10956.

19. Zhang, J., Li, C., Robles-Kelly, A., and Kankanhalli, M. (2020). Hierarchically fair federated learning. Preprint at arXiv.

20. Lyu, L., Xu, X., Wang, Q., and Yu, H. (2020). Collaborative fairness in federated learning. Federated Learning: Privacy and Incentive, 189–204.

21. Deng, Y., Lyu, F., Ren, J., Chen, Y.-C., Yang, P., Zhou, Y., and Zhang, Y. (2021). Fair: Quality-aware federated learning with precise user incentive and model aggregation. In IEEE INFOCOM 2021-IEEE Conference on Computer Communications (IEEE), pp. 1–10.

22. Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. (2022). Minimax demographic group fairness in federated learning. Preprint at arXiv.

23. Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2019). Fair resource allocation in federated learning. Preprint at arXiv.

24. Mohri, M., Sivek, G., and Suresh, A.T. (2019). Agnostic federated learning. In International Conference on Machine Learning (PMLR), pp. 4615–4625.

25. Li, T., Beirami, A., Sanjabi, M., and Smith, V. (2020). Tilted empirical risk minimization. Preprint at arXiv.

26. Horvath, S., Laskaridis, S., Almeida, M., Leontiadis, I., Venieris, S., and Lane, N. (2021). Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. Adv. Neural Inf. Process. *34*, 12876–12889.

27. Deng, Y., Kamani, M.M., and Mahdavi, M. (2020). Distributionally Robust Federated Averaging (NeurIPS).

28. Ro, J., Chen, M., Mathews, R., Mohri, M., and Suresh, A.T. (2021). Communicationefficient agnostic federated averaging. Preprint at arXiv.

29. Li, T., Hu, S., Beirami, A., and Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. In International Conference on Machine Learning (PMLR), pp. 6357–6368.

30. Chu, W., Xie, C., Wang, B., Li, L., Yin, L., Zhao, H., and Li, B. (2022). Focus: Fairness via agent-awareness for federated learning on heterogeneous data. Preprint at arXiv.

31. Hanzely, F., and Richt'arik, P. (2020). Federated learning of a mixture of global and local models. Preprint at arXiv.

32. Deng, Y., Kamani, M.M., and Mahdavi, M. (2020). Adaptive personalized federated learning. Preprint at arXiv.

33. Mansour, Y., Mohri, M., Ro, J., and Suresh, A.T. (2020). Three approaches for personalization with applications to federated learning. Preprint at arXiv.

34. Wang, Z., Fan, X., Qi, J., Wen, C., Wang, C., and Yu, R. (2021). Federated Learning with Fair Averaging (IJCAI).

35. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and open problems in federated learning. Preprint at arXiv.

36. Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) *10*, 1–19.

37. Rahman, K.M.J., Ahmed, F., Akhter, N., Hasan, M., Amin, R., Aziz, K.E., Islam, A.K.M.M., Mukta, M.S.H., and Islam, A.K.M.N. (2021). Challenges, applications and design aspects of federated learning: A survey. IEEE Access *9*, 124682–124700.

38. Duchi, J., and Namkoong, H. (2017). Variance-based regularization with convex objectives. NeurIPS.

39. Namkoong, H., and Duchi, J.C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. NIPS (News Physiol. Sci.) *29*, 2208–2216.

40. Cotter, A., Jiang, H., Gupta, M.R., Wang, S., Narayan, T., You, S., and Sridharan, K. (2019). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. J. Mach. Learn. Res. *20*, 1–59.

41. Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In International Conference on Machine Learning (PMLR), pp. 2029–2037.

42. Sagawa, S., Koh, P.W., Hashimoto, T.B., and Liang, P. (2020). Distributionally Robust Neural Networks ICLR.

43. Liu, J., Shen, Z., Cui, P., Zhou, L., Kuang, K., Li, B., and Lin, Y. (2021). Stable adversarial learning under distributional shifts. Proc. AAAI Conf. Artif. Intell. *35*, 8662–8670.

44. Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In Proceedings of the 25th international conference on Machine learning, pp. 272–279.

45. Rawls, J. (2001). Justice as Fairness: A Restatement (Harvard University Press).

46. Hsu, T.-M.H., Qi, H., and Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. Preprint at arXiv.

47. Ezzeldin, Y.H., Yan, S., He, C., Ferrara, E., and Avestimehr, A.S. (2023). Fairfed: Enabling group fairness in federated learning. Proc. AAAI Conf. Artif. Intell. *37*, 7494–7502.

48. Poulain, R., Bin Tarek, M.F., and Beheshti, R. (2023). Improving fairness in ai models on electronic health records: The case for federated learning methods. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1599–1608.

49. Pfohl, S.R., Foryciarz, A., and Shah, N.H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. J. Biomed. Inf. *113*, 103621.

50. Barocas, S., and Selbst, A.D. (2016). Big data's disparate impact. Calif. Law Rev. *104*, 671.

51. Woodworth, B., Gunasekar, S., Ohannessian, M.I., and Srebro, N. (2017). Learning non-discriminatory predictors. In Conference on Learning Theory (PMLR), pp. 1920–1953.

52. Zhang, D.Y., Kou, Z., and Wang, D. (2020). Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In 2020 IEEE International Conference on Big Data (Big Data) (IEEE), pp. 1051–1060.

53. Hao, W., El-Khamy, M., Lee, J., Zhang, J., Liang, K.J., Chen, C., and Duke, L.C. (2021). Towards fair federated learning with zero-shot data augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3310–3319.

54. Du, W., Xu, D., Wu, X., and Tong, H. (2021). Fairness-aware agnostic federated learning. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM) (SIAM), pp. 181–189.

55. Cui, S., Pan, W., Liang, J., Zhang, C., and Wang, F. (2021). Addressing algorithmic disparity and performance inconsistency in federated learning. Adv. Neural Inf. Process. *34*.

56. Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H.H., and Zhou, J. (2021). Federated adversarial debiasing for fair and transferable representations. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 617–627.

57. Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. NeurIPS *29*, 3315–3323.

58. Zhang, F., Shuai, Z., Kuang, K., Wu, F., Zhuang, Y., and Xiao, J. (2023). Unified Fair Federated Learning for Digital Healthcare (Zenodo). https://doi.org/10.5281/zenodo.10158014.

59. Duggan, M.A., Anderson, W.F., Altekruse, S., Penberthy, L., and Sherman, M.E. (2016). The surveillance, epidemiology and end results (seer) program and pathology: towards strengthening the critical relationship. Am. J. Surg. Pathol. *40*, e94–e102.

60. Ayres-de Campos, D., Bernardes, J., Garrido, A., Marques-de-Sá, J., and Pereira-Leite, L. (2000). Sisporto 2.0: a program for automated analysis of cardiotocograms. J. Matern. Fetal Med. *9*, 311–318.

61. Baqui, P., Bica, I., Marra, V., Ercole, A., and van Der Schaar, M. (2020). Ethnic and regional variations in hospital mortality from COVID-19 in brazil: a cross-sectional observational study. Lancet Global Health *8*, e1018–e1026.

62. 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. Ann. Intern. Med. *122*, 191–203.

63. Seedat, N.J., Crabb'e, J., Bica, I., and van der Schaar, M. (2022). Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data. Adv. Neural Inf. Process. *35*, 23660–23674.

64. Wang, W.-Y., Xie, Y., Zhou, H., and Liu, L. (2021). Contribution of traditional chinese medicine to the treatment of COVID-19. Phytomedicine *85*, 153279.

65. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.

66. Wang, Y., Tong, Y., and Shi, D. (2020). Federated latent dirichlet allocation: A local differential privacy based framework. Proc. AAAI Conf. Artif. Intell. *34*, 6283–6290.

67. Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. Preprint at arXiv.