https://doi.org/10.1038/s44401-025-00035-2

FairFML: fair federated machine learning with a case study on reducing gender disparities in cardiac arrest outcome prediction

Siqi Li[1,9], Qiming Wu[1,9], Doudou Zhou[2], Xin Li[1], Di Miao[1], Chuan Hong[3], Wenjun Gu[1], Yuqing Shang[1], Yohei Okada[4,5], Michael Hao Chen[1], Mengying Yan[3], Yilin Ning[1], Marcus Eng Hock Ong[6,7] & Nan Liu[1,4,8] ✉

Health equity is a critical concern in clinical research and practice, as biased predictive models can exacerbate disparities in clinical decision-making and patient outcomes. As healthcare systems increasingly rely on data-driven models, ensuring fairness in these systems is essential to prevent perpetuating existing disparities. While large-scale healthcare data exists across multiple institutions, cross-institutional collaborations often face privacy constraints, highlighting the need for privacy-preserving solutions that also promote fairness. We present Fair Federated Machine Learning (FairFML), a model-agnostic solution designed to reduce algorithmic bias in cross-institutional healthcare collaborations while preserving patient privacy. Validated through a real-world case study on reducing gender disparities in cardiac arrest outcome prediction, FairFML improved fairness metrics by up to 90% without compromising predictive performance. FairFML is flexible and compatible with various FL frameworks and models, from traditional statistical methods to deep learning, offering a robust and scalable solution for equitable model development in clinical settings.

Machine learning (ML) and artificial intelligence (AI) methods have been rapidly adopted in healthcare for a broad range of data-driven applications, such as predictive modeling[1], personalized treatment recommendations[2], and resource allocation in health systems[3]. However, ensuring health equity remains a critical challenge, particularly when algorithmic findings directly impact clinical decision-making and patient care within health systems[4]. Concerns have grown regarding the underperformance of ML and AI systems for historically underserved populations, including women and individuals from lower socioeconomic backgrounds[5]. For instance, studies have shown that Black patients are more frequently underdiagnosed with chronic obstructive pulmonary disease (COPD) compared to Hispanic White patients, emphasizing the need to address these disparities[5,6]. Similarly, in postpartum depression, ML models trained on gender-imbalanced data favored White women, even when Black women were predicted to be at higher risk, illustrating racial disparities in healthcare outcomes[7]. Medical imaging datasets, such as X-rays, also produce biased classifiers due to gender-imbalanced data, leading to consistently poorer performance for underrepresented genders[8].

Algorithmic disparity[9], often referred to as "biased" or "unfair" decision-making, arises when predictive models perform unequally across subgroups[10,11] defined by sensitive attributes such as gender, race/ethnicity, and socioeconomic status[12]. These inequities span various healthcare domains, including COVID-19[4,9,10], stroke[13], emergency medicine[14–16], cardiovascular disease[17], cancer[18], and organ transplants[19]. Despite growing efforts to develop fair models[12], most studies rely on single, centralized datasets. However, healthcare data are often distributed across multiple institutions, such as electronic health records (EHRs) from different hospitals or mobile health data from users' devices[20,21]. Aggregating these diverse data sources could accelerate research and improve care quality[22], but privacy regulations pose significant barriers[23].

[1]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore. [2]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore. [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. [4]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore. [5]Department of Preventive Services, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [6]Health Services Research Centre, Singapore Health Services, Singapore, Singapore. [7]Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore. [8]NUS Artificial Intelligence Institute, National University of Singapore, Singapore, Singapore. [9]These authors contributed equally: Siqi Li, Qiming Wu. ✉e-mail: liu.nan@duke-nus.edu.sg

Federated learning (FL), or federated ML (FML), offers a promising solution by enabling participants to collaboratively train models without sharing sensitive data[21], making it an increasingly popular approach in medical research[24,25]. However, while FL adoption is increasing, most studies focus primarily on overall predictive performance, often overlooking its potential to address algorithmic disparities[25]. Evidence suggests that standard FL algorithms struggle to reduce algorithmic biases[26,27], leading to models that retain their unfairness when transitioning from single-site analyses to FL settings. Local biases may persist or even be amplified due to the lack of centralization, as each institution contributes heterogeneous data that reflect varying socio-demographic characteristics and clinical practices, introducing diverse biases.

Although some studies have investigated these disparities within FL contexts, they predominantly rely on conventional ML datasets rather than real-world clinical data[27–29], raising concerns about the generalizability of their findings to actual healthcare systems. To address these gaps, we propose Fair Federated Machine Learning (FairFML), a unified solution to promote fairness in FL among distributed healthcare systems. As a proof of concept, we used real-world out-of-hospital cardiac arrest (OHCA) data from the United States, focusing on gender disparities—a critical concern for equity in OHCA care[15,30,31]. These disparities are often attributed to complex factors, including differences in layperson bystander cardiopulmonary resuscitation (CPR)[15]. This case study aims to demonstrate FairFML's effectiveness in mitigating such disparities, while maintaining prediction performance comparable to both local and centralized analyses.

Results

Model performance and fairness evaluation

Figure 1 illustrates the partitioning of 7425 individual episodes into four or six sites following the cohort formation process, with a 7:3 split for training and testing data. Supplementary Table 1 in the Supplementary Materials summarizes the baseline characteristics of the overall cohort and each site under different experimental conditions. In cases I and III, where clients were partitioned by race/ethnicity, significant distribution differences were observed, with the proportion of White individuals ranging from 88.9% to 48.2%. In cases II and IV, where clients were partitioned by age, the mean age varied considerably, ranging from approximately 60 to approximately 70 years. Outcome prevalence varied from 7.5% to 12.6%, and other variables also exhibited heterogeneous distributions, reflecting the real-world demographic differences across regions.

Details of the experimental setup, including the tuning of λ and γ and other general hyperparameters for FL, are provided in Supplementary Fig. 2 and Supplementary Table 2 in the Supplementary Materials. This tuning process is crucial for managing the inherent trade-off between model fairness and predictive performance, as increasing λ prioritizes fairness, typically leading to a controlled decrease in overall accuracy.

We assessed the performance of the federated model developed using FairFML by comparing it to the centralized model, local models trained independently at each site, and general FL models (FedAvg and Per-FedAvg). Specifically, FairFML integrates with these FL frameworks by replacing their standard model loss function with a λ-weighted fairness loss during training, forming FairFML (FedAvg) and FairFML (Per-FedAvg). This process is visually depicted in Fig. 2 and detailed algorithmically in

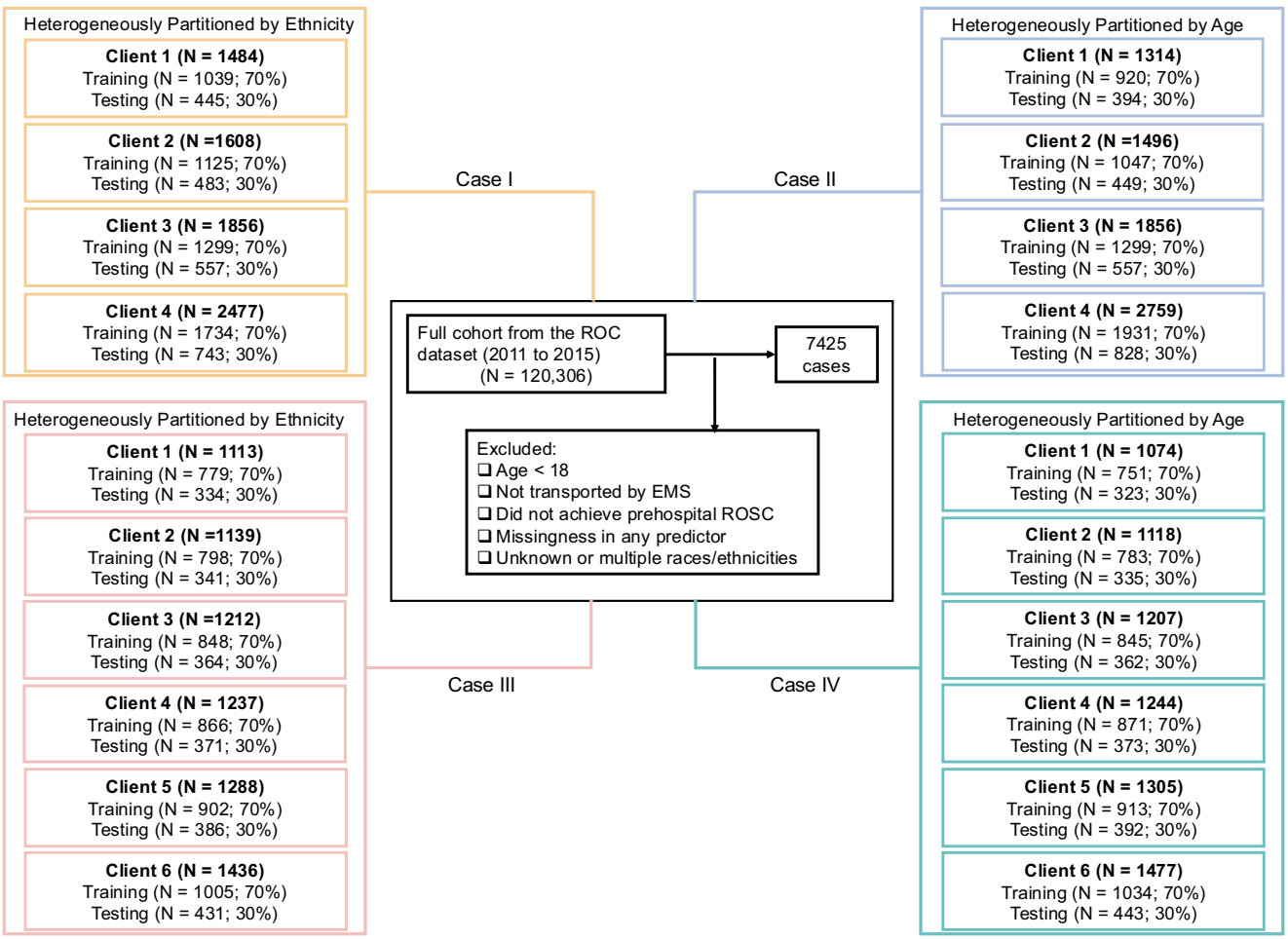

Fig. 1 | Cohort formation flow diagram. A total of 7425 episodes were partitioned heterogeneously across clients by race/ethnicity.

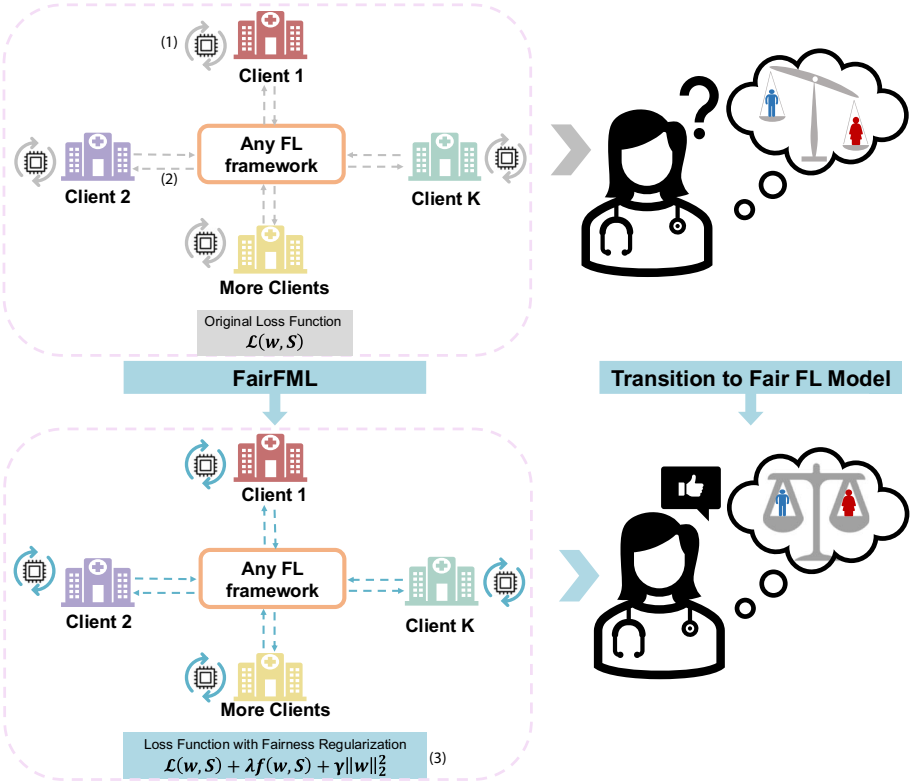

Fig. 2 | Workflow of FairFML. (1) Client-side training; (2) Federated parameter exchange; (3) Fairness loss incorporation.

Supplementary Fig. 1 in the Supplementary Materials. Figure 3 illustrates the performance of each model across the testing datasets for all sites in the Case IV experimental scenario. Results for the other three cases are provided in Supplementary Figs. 3–5 of the Supplementary Materials, detailing the performance comparison and fairness metrics for Case I (race/ethnicity, 4 sites), Case II (age, 4 sites), and Case III (race/ethnicity, 6 sites), respectively. These figures all show that fairness metrics generally improved across all clients, aligning with the overall trend across sites, with only a minor trade-off in predictive performance.

Key findings from Fig. 3 and Supplementary Figs. 3–5 include: (1) FairFML consistently outperformed other baseline models in fairness, demonstrating substantial improvements across metrics such as decreases in DPD and EOD (moving closer to 0) and increases in DPR and EOR (moving closer to 1). It maintained predictive performance nearly identical to other baseline models, with a maximum AUC decrease of less than 0.02 relative to the centralized model. (2) FairFML sometimes narrowed the confidence interval compared to baseline models, suggesting more stable model performance in terms of the corresponding fairness metrics. (3) Although FedAvg and Per-FedAvg occasionally outperformed central and local models on specific fairness metrics for certain clients, their improvements were less substantial. In contrast, FairFML-based model consistently demonstrated significant and superior performance across all fairness metrics.

Discussion

FairFML offers a unified, model- and framework-agnostic solution[25,32] for enhancing fairness in FL collaborations. Its adaptability to various FL frameworks and ML models—ranging from traditional statistical regressions and support vector machines to deep neural networks—makes it highly versatile for clinical and biomedical prediction tasks[25]. By reducing algorithmic disparities, as shown in our case study on gender disparities in cardiac arrest outcomes, FairFML mitigates bias for underserved populations when integrated with standard FL frameworks. This provides significant value to health systems by improving fairness in predictive models, which directly impacts clinical decision-making. At the system level, FairFML helps reduce care delivery disparities, enhances resource allocation, and optimizes healthcare services, particularly in distributed systems where data is private and cannot be shared.

Given that clients in cross-institutional FL collaborations often expect direct benefits for their research or clinical practice[22,25], it is essential to evaluate models against both client-level (local) and central models. Our results show that FairFML consistently outperforms traditional FL and local models in terms of fairness between the two genders, as seen in Fig. 3 and Supplementary Table 3. While the maximum AUC decrease compared to centralized or standard FL models was <0.02, this modest reduction is a clinically acceptable trade-off, outweighed by the substantial gains in equitable care achieved through bias mitigation. Establishing the real-world impact necessitates future prospective analyses and close collaborations with clinicians to directly evaluate patient-level outcomes where fairness is explicitly prioritized in model predictions.

Beyond its strong performance characteristics, FairFML's design ensures broad compatibility with a variety of FL algorithms, including FedAvg, FedProx, and Per-FedAvg, without requiring modifications to their underlying mechanics. While Per-FedAvg is known to improve client-level personalization through meta-learning[33], our experiments show that FairFML (Per-FedAvg) often achieves superior fairness outcomes compared to FairFML (FedAvg), highlighting its alignment with established personalization benefits. These properties make FairFML highly scalable, adaptable, and practical for real-world biomedical FL scenarios where fairness, interpretability, and implementation feasibility are critical.

FairFML's convex formulation enables efficient optimization using standard stochastic gradient descent and supports seamless integration into a wide range of predictive models, including logistic regression, ridge regression, support vector machines, and neural networks. Importantly, FairFML does not increase the underlying optimization complexity of the base model. For convex models, the overall objective remains convex; for

Fig. 3 | Results of case IV in our experiment. Performance comparison of the proposed FairFML method against baseline models using gender as the sensitive attribute. Area Under the Curve (AUC) measures predictive performance, while demographic parity difference (DPD), equalized odds difference (EOD), demographic parity ratio (DPR), and equalized odds ratio (EOR) assess fairness. Error bars represent the 95% confidence interval (CI) for each metric.

non-convex models, the fairness penalty introduces no additional non-convexity, allowing training to proceed as usual. In such cases, practitioners may adopt more robust FL frameworks—such as FedProx—to better handle convergence in non-convex settings.

Another benefit of FairFML's design is its preservation of model interpretability: while the convex fairness loss modifies model parameters to enhance fairness, it does so without fundamentally altering the model's core architecture. Consequently, commonly used explainable AI tools, such as LIME[34] and Shapley-Value-based ones[35,36], remain fully applicable, as their methodologies primarily depend on analyzing the model's input-output behavior rather than being sensitive to exact internal parameter values.

Building on its ability to enhance fairness for specific attributes, an important next step towards comprehensive health equity involves addressing multi-group fairness. Indeed, although gender disparities in cardiac arrest are a key focus, they are not the only relevant partition for group fairness in this context[37]. Studies show that individuals from Black, Hispanic, or lower socioeconomic status backgrounds experience pronounced disparities throughout the resuscitation pathway[38]. Our findings, presented in Supplementary Table 4 of the Supplementary Materials, highlight significant variations in gender disparities when further partitioned by race/ethnicity and age (≥65 vs. <65), demonstrating the relevance of intersectional multi-group fairness (i.e., multiple intersecting sensitive variables[39]) to further mitigate unfairness. Despite more than a decade of discussion on multi-group fairness[40,41], it has received limited attention in FL settings. This is particularly challenging when group partitions are imbalanced or entirely absent from some clients; in such scenarios, the fairness penalty may become unstable or undefined due to the lack of valid group comparisons.

While these aspects present considerable challenges, FairFML's model-agnostic and convex formulation provides a foundational framework that could be extended to address multi-group fairness in future work, allowing its penalty term to be integrated and trained using standard stochastic gradient descent strategies in FL. In addition, incorporating robust strategies such as oversampling (e.g., ROSE[42]), conditional data synthesis[43,44], or generative models like GANs[45] could help deal with imbalanced data. These adaptations would enable FairFML to operate more effectively in imbalanced or incomplete real-world clinical datasets.

Beyond the algorithmic fairness considerations discussed thus far, the concept of fairness in FL also encompasses broader aspects. This often involves client resource allocation and ensuring performance uniformity across clients[46,47], commonly referred to as "system fairness"[48]. This is particularly relevant in scenarios involving client selection to optimize convergence speed and reduce computational costs[49], as seen in cross-device FL[21]. However, cross-institutional FL[21]—which is more prevalent in healthcare settings and often involves fewer clients (typically fewer than five)[25] —the focus shifts to algorithmic fairness. While various strategies have been proposed to enhance fairness in clinical models, including privacy-preserving collaborations, McCradden et al.[50] caution that relying solely on technical solutions may inadvertently harm vulnerable groups. Thus, FairFML should be viewed as a starting point, followed by further analysis of downstream patient impacts, rather than assuming that fairness can be achieved solely through ML/AI metrics[50].

Translating FairFML into real-world clinical practice also requires overcoming significant logistical and operational challenges. These include securing multi-site IRB approvals and data-sharing agreements, coordinating domain experts to harmonize heterogeneous variable definitions, and establishing secure infrastructure with sustained engineering support. Potential solutions include developing modular, GUI-based tools that minimize coding burdens and standardized governance frameworks—such as the FAIR-EC[51] collaboration—to ethically and efficiently streamline privacy-preserving multi-site analyses.

Our clinical case study uses simulated partitioned clients for FL experiments as a proof of concept, in preparation for real-world applications. Although we simulated cross-site data heterogeneity, real-world collaborations may introduce additional complexities, particularly regarding model heterogeneity[25,32]. Further research is required to validate FairFML's robustness and applicability in real-world cross-institutional collaborations.

While this study focused on group fairness, our proposed method can be extended to improve individual fairness[52] by incorporating an individual fairness penalty within the convex framework[53]. A hybrid penalty combining both group and individual fairness metrics could offer a more comprehensive approach to mitigating unfairness in clinical research. In addition, the fairness penalty could be extended to handle multi-group fairness metrics, such as gender, ethnicity, and socioeconomic status, simultaneously by using the intersection of subgroups, given the convex property of the proposed fairness penalty, which could be easily trained using simple stochastic gradient descent strategies in FL training.

Moreover, to handle temporal distribution shifts in longitudinal or real-time settings, transfer learning[54] techniques can be integrated into FairFML to update the model as new batches of data arrive, preserving previously learned fairness constraints while adapting to evolving patient populations. Future work will aim to explore these extensions and validate FairFML in real-world settings to ensure its robustness and applicability across diverse clinical environments.

FairFML effectively mitigates bias and enhances fairness in model co-training across multiple healthcare data owners while preserving privacy. In our proof-of-concept case study using real-world emergency medicine data, FairFML reduced fairness disparities and improved outcomes for underserved populations without compromising predictive performance. These findings highlight the clinical value of FairFML in fostering equitable decision-making within health systems, ensuring that AI models benefit all patient groups fairly. By embedding fairness into FL frameworks, FairFML supports healthcare systems in optimizing resource allocation and improving care delivery, particularly in settings where data is distributed and privacy concerns are critical.

Methods

Notation and problem setup

In this study, we adopt the notation introduced by Berk et al.[53]. Let $y \in \mathcal{Y} = [-1, 1]$ represent the binary outcome and $x \in \mathcal{X} = R^d$ denote the feature vectors. Each instance is categorized into one of two groups based on a sensitive variable, denoted as χ_{d+1}. The joint distribution of $\mathcal{X}$ and $\mathcal{Y}$ is represented by $\mathcal{P}$. We consider a training set $S = \{(x_i, y_i)\}_{i=1}^{n}$, consisting of n independent and identically distributed (i.i.d.) samples drawn from $\mathcal{P}$. This training set is divided into two groups, S_1 and S_2, based on the sensitive variable, with n_1 and n_2 representing the respective sizes of these groups, such that $n_1 + n_2 = n$.

The λ-weighted fairness loss for a given model is defined as $\mathcal{L}(w, S) + \lambda f(w, S)$, where $\mathcal{L}$ represents the standard model loss function, w represents model parameters, and λ is a regularization parameter for the fairness penalty. Consistent with Berk et al.[53], we focus on a group fairness penalty, defined as

$$f(\boldsymbol{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x_i}, \boldsymbol{y_i}) \in S_1 \\ (\boldsymbol{x_j}, \boldsymbol{y_j}) \in S_2}} d\left(y_i, y_j\right)\left(\boldsymbol{w} \cdot \boldsymbol{x_i} - \boldsymbol{w} \cdot \boldsymbol{x_j}\right) \tag{1}$$

Here, $d(y_i, y_j) = \mathbb{1}[y_i = y_j]$ serves as the cross-group fairness weight.

4.2 Group fairness metrics

Demographic parity (DP), also known as statistical parity, and equalized odds (EO) are two widely used algorithmic fairness definitions for binary classifications:

- A model satisfies DP over a distribution $\mathscr{P}$ if its prediction $\hat{Y}$ is statistically independent of the sensitive feature:

$$P\left[\hat{Y}=1|\chi_{\mathscr{A}+1}=a\right]=P\left[\hat{Y}=1\right],\ \forall a \quad (2)$$

- A model satisfies EO over a distribution $\mathscr{P}$ if its prediction $\hat{Y}$ is conditionally independent of the sensitive feature given the true outcome label:

$$P\left[\hat{Y}=1|\chi_{\mathscr{A}+1}=a,Y=y\right]=P\left[\hat{Y}=1,=|Y=y\right],\forall a,y \quad (3)$$

In this study, we focused on a total of four fairness metrics: demographic parity difference (DPD), demographic parity ratio (DPR), equalized odds difference (EOD), and equalized odds ratio (EOR), which are calculated using the definitions of DP and EO as follows:

- DPD $=\max_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a\right]-\min_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a\right]$ measures the maximum difference in predicted outcomes across groups. A DPD closer to 0 indicates more equal predictions across groups.
- DPR $=\frac{\min_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a\right]}{\max_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a\right]}$ measures the ratio of the minimum to maximum predicted outcomes across groups. A DPR closer to 1 indicates more balanced prediction rates.
- EOD $=\max_{y\in\{-1,1\}}\left(\max_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a,Y=y\right]-\min_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a,Y=y\right]\right)$ measures the difference in prediction errors (false positives/negatives) across groups. An EOD closer to 0 indicates more equal predictions across groups.

EOR $=\min_{y\in\{-1,1\}}\frac{\min_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a,Y=y\right]}{\max_a E\left[\hat{Y}|\chi_{\mathscr{A}+1}=a,Y=y\right]}$ measures the ratio of error rates between groups. An EOR closer to 1 indicates more balanced prediction rates.

FairFML

We integrated the λ-weighted fairness loss described in "Natation problem setup" into the FL model training, and the workflow of our proposed FairFML is illustrated in Fig. 2. As shown, incorporating FairFML into any FL framework enhances the fairness of existing FL solutions by replacing the standard model loss function $\mathscr{L}$ with the λ-weighted fairness loss function during FL model training. The fairness regularizer f is convex[53], meaning that it has a single global minimum and no local minima. This property is crucial for optimization because it guarantees that the combined objective function $\mathscr{L}(w,S)+\lambda f(w,S)$ can be efficiently minimized without the risk of converging to suboptimal solutions. Convexity ensures that as we adjust λ, the trade-off between fairness and model accuracy is stable and predictable, which is essential for effective optimization in typical FL frameworks, such as FedAvg[55]. To prevent overfitting, we incorporate l_2 regularization, resulting in the final loss function: $\mathscr{L}(w,S)+\lambda f(w,S)+\gamma||w||_2^2$.

The trade-off between model accuracy and fairness, regulated by λ, varies significantly across datasets[53,56] where higher λ values impose greater fairness penalties. As λ increases from 0 to ∞, model accuracy tends to decrease. Therefore, users need to select an appropriate λ value for each dataset to balance improved fairness with an acceptable reduction in model accuracy. To address this challenge, we propose a data-driven approach for efficiently selecting λ while minimizing computational costs. As outlined in the pseudocode (Supplementary Fig. 1, Supplementary Materials), λ_k is initially chosen independently for each client k by plotting prediction metrics (e.g., accuracy or mean square error (MSE)) against λ_k. A practical method involves incrementing λ_k in fixed steps until the prediction metrics degrade beyond a set threshold compared to the unregularized model (e.g., when accuracy falls below $0.995*\mathrm{Acc}_0$, where Acc_0 is the accuracy of the model without the fairness penalty). The maximum λ_k across all clients is then used to define the range for FL training, from which a user-defined set of equally spaced λ values is selected.

For each λ value, we use a two-step strategy to determine the optimal γ. First, we explore broad, equally spaced γ values starting from zero. The user selects the best γ based on changes in predictive performance and fairness metrics. We then narrow the search range around that value and repeat the process to finalize γ for the given λ. Detailed pseudocode for selecting γ is provided in Supplementary Fig. 1.

Dataset and experiments

Our study population comprised OHCA patients treated by emergency medical services (EMS) providers, as recorded in the Resuscitation Outcomes Consortium (ROC) Cardiac Epidemiologic Registry (Epistry) (Version 3, covering the period from April 1, 2011, to June 30, 2015). The ROC, a North American database established in 2004, aims to advance clinical research on cardiopulmonary arrest[57]. Ethical approval was obtained from the National University of Singapore Institutional Review Board (IRB), which granted an exemption for this study (IRB Reference Number: NUS-IRB-2023-451).

For cohort formation and predictor selection, we followed established methodologies in out-of-hospital cardiac arrest (OHCA) research[57,58]. We included patients aged 18 and older who were transported by EMS, achieved return of spontaneous circulation (ROSC) at any point prehospital, and had complete data on gender, race, etiology, initial rhythm, witness status, response time, adrenaline use, and neurological status. The primary outcome was neurological status at discharge, measured by the Modified Rankin Scale (MRS), where scores of 0, 1, or 2 were classified as a good outcome. Variables used for outcome prediction included age (in years), etiology of arrest (cardiac/non-cardiac), witness presence (yes/no), initial rhythm (shockable/non-shockable), bystander cardiopulmonary resuscitation (CPR) (yes/no), response time (in minutes), and adrenaline use (yes/no).

We conducted four sets of experiments to simulate real-world cross-site data by partitioning the full study cohort heterogeneously: (I) by race/ethnicity into four sites, (II) by age into four sites, (III) by race/ethnicity into six sites, and (IV) by age into six sites. Specifically, the probability of an observation being assigned to each site depends on the variable used for partitioning (age or race/ethnicity). As a result, the marginal distributions of predictors and outcomes become heterogeneous across sites. Continuous variables were standardized using the mean and standard deviation from the full cohort, and logistic regression was employed for outcome prediction. We focused on two representative FL frameworks, FedAvg and Per-FedAvg[33]. FedAvg is the foundational FL framework, being the first proposed in the FL domain[32,54], while Per-FedAvg is a widely adopted solution for personalized FL. The latter is particularly relevant in healthcare data analysis, as it allows researchers to determine whether FL can offer localized benefits that enhance the performance of existing models for individual institutions[25]. Its effectiveness for personalized improvements on local datasets has also been demonstrated with healthcare data[59].

For each scenario, we conducted three types of analyses: (1) a central model trained on the full cohort and local models trained independently at each site, (2) federated logistic regression using FedAvg and Per-FedAvg, and (3) fairness-enhanced federated logistic regression using the proposed FairFML method with the two FL frameworks–FairFML (FedAvg) and FairFML (Per-FedAvg). We evaluated model performance using the area under the receiver operating characteristic curve (AUROC) and four fairness metrics, as described in "Group fairness metrics," with gender as the sensitive variable, using the 'Fairlearn' package[60].

Data availability

This study used data from the publicly available Resuscitation Outcomes Consortium (ROC) Epistry database (Version 3, April 2011–June 2015).

The dataset can be requested through the NIH website at: https://biolincc.nhlbi.nih.gov/studies/roc_cardiac_epistry_3/.

Code availability

The Python code for FairFML is available at https://github.com/nliulab/FairFML.

Received: 23 March 2025; Accepted: 16 July 2025;
Published online: 12 August 2025

References

1. de Hond, A. A. H. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Npj Digit Med.* **5**, 1–13 (2022).
2. Johnson, K. B. et al. Precision medicine, AI, and the future of personalized health care. *Clin. Transl. Sci.* **14**, 86–93 (2021).
3. Wu, H., Lu, X. & Wang, H. The application of artificial intelligence in health care resource allocation before and during the COVID-19 pandemic: scoping review. *JMIR AI* **2**, e38397 (2023).
4. Yang, J. et al. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nat. Mach. Intell.* **5**, 884–894 (2023).
5. Seyyed-Kalantari, L. et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
6. Mamary, A. J. et al. Race and gender disparities are evident in COPD underdiagnoses across all severities of measured airflow obstruction. *Chronic Obstr. Pulm. Dis. J. COPD Found.* **5**, 177–184 (2018).
7. Park, Y. et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw. Open* **4**, e213909 (2021).
8. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. https://doi.org/10.1073/pnas.1919012117.
9. Cui, S. et al. Addressing algorithmic disparity and performance inconsistency in federated learning. In *Proc. 35th International Conference on Neural Information Processing Systems* 26091–26102 (Curran Associates, Inc. 2021).
10. Yang, J. et al. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *Npj Digit Med.* **6**, 1–10 (2023).
11. Krasanakis, E. et al. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proc. 2018 World Wide Web Conference on World Wide Web - WWW'18* 853–862. (ACM Press, 2018).
12. Liu, M. et al. A translational perspective towards clinical AI fairness. *Npj Digit Med.* **6**, 1–6 (2023).
13. Hong, C. et al. Trans-balance: reducing demographic disparity for prediction models in the presence of class imbalance. *J. Biomed. Inf.* **149**, 104532 (2024).
14. Wu, H. et al. Quantifying Health Inequalities Induced by Data and AI Models. In *Proc. Thirty-First International Joint Conference on Artificial Intelligence* 5192–5198 (International Joint Conferences on Artificial Intelligence Organisation, 2022).
15. Liu, N. et al. Gender disparities among adult recipients of layperson bystander cardiopulmonary resuscitation by location of cardiac arrest in Pan-Asian communities: a registry-based study. *eClinicalMedicine* **44**, 101293 (2022).
16. Li, C. et al. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. *J. Biomed. Inf.* **143**, 104399 (2023).
17. Barda, N. et al. Addressing bias in prediction models by improving subpopulation calibration. *J. Am. Med. Inf. Assoc.* **28**, 549–558 (2021).
18. Bradshaw, R. L. et al. Enhanced family history-based algorithms increase the identification of individuals meeting criteria for genetic testing of hereditary cancer syndromes but would not reduce disparities on their own. *J. Biomed. Inf.* **149**, 104568 (2024).
19. Li, C., Jiang, X. & Zhang, K. A transformer-based deep learning approach for fairly predicting post-liver transplant risk factors. *J. Biomed. Inf.* **149**, 104545 (2024).
20. Zhou, D. et al. Federated offline reinforcement learning. *J. Am. Stat. Assoc*. **0**, 1–12. https://doi.org/10.1080/01621459.2024.2310287 (2024).
21. Kairouz, P. et al. Advances and open problems in federated learning. Foundations and trends® in machine learning, **14**, 1–210 (2021).
22. Li, S. et al. FedScore: a privacy-preserving framework for federated scoring system development. *J. Biomed. Inf.* **146**, 104485 (2023).
23. Teo, Z. L. et al. Federated machine learning in healthcare: a systematic review on clinical applications and technical architecture. *Cell Rep. Med.* **5**, 101419 (2024).
24. Dayan, I. et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* **27**, 1735–1743 (2021).
25. Li, S. et al. Federated and distributed learning applications for electronic health records and structured medical data: a scoping review. *J. Am. Med. Inform. Assoc*. **30**, 2041–2049 (2023).
26. Zhang, F. et al. Unified fair federated learning for digital healthcare. *Patterns* **5**, 1009077 (2024).
27. Zhang, D. Y., Kou, Z. & Wang, D. FairFL: a fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *Proc. IEEE International Conference on Big Data (Big Data)*, 1051–1060 (IEEE, 2020).
28. Li, J. et al. Improve individual fairness in federated learning via adversarial training. *Comput. Secur*. **132**, 103336 (2023).
29. Du, W., Xu, D., Wu, X. & Tong, H. Fairness-aware agnostic federated learning. In *Proceedings of the 2021* SIAM International *Conference on Data Mining (SDM).* (pp. 181−189). Society for Industrial and Applied Mathematics (2021).
30. Gramm, E. R., Salcido, D. D. & Menegazzi, J. J. Disparities in out-of-hospital cardiac arrest treatment and outcomes of males and females. *Prehosp. Emerg. Care* **27**, 1041–1047 (2023).
31. Grunau, B. et al. Public access defibrillators: gender-based inequities in access and application. *Resuscitation* **150**, 17–22 (2020).
32. Li, S. et al. Federated learning in healthcare: a benchmark comparison of engineering and statistical approaches for structured data analysis. *Health Data Sci.* **0**, 0196, https://doi.org/10.34133/hds.0196 (2024).
33. Fallah, A., Mokhtari, A. & Ozdaglar, A. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. In *Proc. Advances in Neural Information Processing Systems* 3557–3568 (Curran Associates, Inc., 2020).
34. Garreau, D. & Luxburg, U. Explaining the explainer: a first theoretical analysis of LIME. In *Proc. Twenty Third International Conference on Artificial Intelligence and Statistics* 1287–1296 (PMLR, 2020).
35. Covert, I. & Lee, S.-I. Improving KernelSHAP: practical Shapley value estimation using linear regression. In *Proc. 24th International Conference on Artificial Intelligence and Statistics* 3457–3465 (PMLR, 2021).
36. Fan, W. et al. SIM-Shapley: a stable and computationally efficient approach to Shapley value approximation. *ArXiv*. https://doi.org/10.48550/arXiv.2505.08198 (2025).
37. Galea, S. et al. Explaining racial disparities in incidence of and survival from out-of-hospital cardiac arrest. *Am. J. Epidemiol.* **166**, 534–543 (2007).
38. Mehta, N. K. et al. Racial, ethnic, and socioeconomic disparities in out-of-hospital cardiac arrest within the United States: now is the time for change. *Heart Rhythm O2* **3**, 857–863 (2022).
39. Sonboli, N. et al. "And the winner is…": dynamic lotteries for multi-group fairness-aware recommendation. *ArXiv*. https://doi.org/10.48550/arXiv.2009.02590 (2020).
40. Zhang, L., Roth, A. & Zhang, L. Fair risk control: a generalized framework for calibrating multi-group fairness risks. In *Forty-first*

International Conference on Machine Learning. *ArXiv*. https://doi.org/10.48550/arXiv.2405.02225 (2024).
41. Byrne, Z. S. & Miller, B. K. Is justice the same for everyone? examining fairness items using multiple-group analysis. *J. Bus. Psychol.* **24**, 51–64 (2009).
42. Demir, S. & Şahin, E. K. Evaluation of oversampling methods (OVER, SMOTE, and ROSE) in classifying soil liquefaction dataset based on SVM, RF, and Naïve Bayes. *Avrupa Bilim Ve Teknol Derg*. 142–147 https://doi.org/10.31590/ejosat.1077867 (2022).
43. Gu, T. et al. Synthetic data method to incorporate external information into a current study. *Can. J. Stat.* **47**, 580–603 (2019).
44. Gu, T., Taylor, J. M. G. & Mukherjee, B. A synthetic data integration framework to leverage external summary-level information from heterogeneous populations. *Biometrics* **79**, 3831–3845 (2023).
45. Karthika, S. & Durgadevi, M. Generative Adversarial Network (GAN): a general review on different variants of GAN and applications. In *Proc. 6th International Conference on Communication and Electronics Systems (ICCES)*. 1–8 (IEEE, 2021).
46. Huang, W. et al. Federated learning for generalization, robustness, fairness: a survey and benchmark. In *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–20. https://doi.org/10.1109/TPAMI.2024.3418862 (IEEE, 2024).
47. Li, T., Sanjabi, M., Beirami, A. & Smith, V. Fair resource allocation in federated learning. In *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1905.10497 (2020).
48. Vucinich, S. & Zhu, Q. The current state and challenges of fairness in federated learning. *IEEE Access* **11**, 80903–80914 (2023).
49. Shi, Y., Yu, H. & Leung, C. Towards fairness-aware federated learning. *IEEE Trans. Neural Netw. Learn Syst.* **35**, 11922–11938 (2024).
50. McCradden, M. D. et al. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health* **2**, e221–e223 (2020).
51. FAIR-EC: A. Global research network for fair, accountable, interpretable, and responsible AI in emergency care. *JMIR Prepr*. https://preprints.jmir.org/preprint/74202.
52. Mukherjee, D. et al. Two simple ways to learn individual fairness metrics from data. In *Proc. 37th International Conference on Machine Learning* 7097–7107 (PMLR, 2020).
53. Berk, R. et al. A convex framework for fair regression. arXiv preprint arXiv:1706.0240 (2017).
54. Siqi Li, et al. Bridging data gaps in healthcare: a scoping review of transfer learning in structured data analysis. *Health Data Sci.* https://doi.org/10.34133/hds.0321
55. McMahan, B. et al. Communication-efficient learning of deep networks from decentralized data. In *Proc. 20th International Conference on Artificial Intelligence and Statistics* 1273–1282 (PMLR, 2017).
56. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On fairness and calibration. Advances in neural information processing systems, 30 (2017).
57. Morrison, L. J. et al. Rationale, development and implementation of the resuscitation outcomes consortium epistry-cardiac arrest. *Resuscitation* **78**, 161–169 (2008).
58. Nishioka, N. et al. External validation of updated prediction models for neurological outcomes at 90 days in patients with out-of-hospital cardiac arrest. *J. Am. Heart Assoc.* **13**, e033824 (2024).
59. Shang, Y., Wu, Q., Li, S. & Miao, D. Empirical evaluations of personalized federated learning on heterogeneous electronic health records. In The Second Tiny Papers Track at ICLR 2024 (2024).
60. Weerts, H. et al. Fairlearn: assessing and improving fairness of AI systems. *J. Mach. Learn. Res.* **24**, 1–8 (2023).

Acknowledgements

This work was supported by the Duke/Duke-NUS Collaboration grant and the KPFA scholarship (Duke-NUS-KPFA/2025/0081). The funder played no role in study design, data collection, analysis, and interpretation of data, or the writing of this manuscript.

Author contributions

Siqi Li: Conceptualization, Project administration, Supervision, Method design, Algorithm development, Formal analysis, Data curation, Writing—original draft. Qiming Wu: Algorithm development, Formal analysis, Software, Writing —original draft. Doudou Zhou: Algorithm development, Software, Writing—review and editing. Xin Li: Data analysis, Writing—original draft. Di Miao: Formal analysis, Writing—original draft. Chuan Hong: Algorithm development, Writing —review and editing. Wenjun Gu: Data curation, Investigation, Writing—review and editing. Yohei Okada: Investigation, Validation, Writing—review and editing. Michael Hao Chen: Investigation, Validation, Writing—review and editing. Mengying Yan: Investigation, Validation, Writing—review and editing. Yuqing Shang: Algorithm development, Investigation, Writing—review and editing. Yilin Ning: Investigation, Validation, Writing—review and editing. Marcus Eng Hock Ong: Investigation, Validation, Writing—review and editing. Nan Liu: Conceptualization, Supervision, Funding acquisition, Resources, Writing—review and editing.

Competing interests

N.L., S.L., and M.E.H.O. hold a patent related to the federated scoring system. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44401-025-00035-2.

Correspondence and requests for materials should be addressed to Nan Liu.