Journal of Healthcare Informatics Research (2024) 8:181–205
https://doi.org/10.1007/s41666-024-00164-7

RESEARCH ARTICLE

Analyzing the Impact of Personalization on Fairness in Federated Learning for Healthcare

Tongnian Wang[1] · Kai Zhang[2] · Jiannan Cai[3] · Yanmin Gong[4] · Kim-Kwang Raymond Choo[1] · Yuanxiong Guo[1]

Received: 8 September 2023 / Revised: 29 February 2024 / Accepted: 7 March 2024 /
Published online: 23 March 2024

Abstract

As machine learning (ML) usage becomes more popular in the healthcare sector, there are also increasing concerns about potential biases and risks such as privacy. One countermeasure is to use federated learning (FL) to support collaborative learning without the need for patient data sharing across different organizations. However, the inherent heterogeneity of data distributions among participating FL parties poses challenges for exploring group fairness in FL. While personalization within FL can handle performance degradation caused by data heterogeneity, its influence on group fairness is not fully investigated. Therefore, the primary focus of this study is to rigorously assess the impact of personalized FL on group fairness in the healthcare domain, offering a comprehensive understanding of how personalized FL affects group fairness in clinical outcomes. We conduct an empirical analysis using two prominent real-world Electronic Health Records (EHR) datasets, namely eICU and MIMIC-IV. Our methodology involves a thorough comparison between personalized FL and two baselines: standalone training, where models are developed independently without FL collaboration, and standard FL, which aims to learn a global model via the FedAvg algorithm. We adopt Ditto as our personalized FL approach, which enables each client in FL to develop its own personalized model through multi-task learning. Our assessment is achieved through a series of evaluations, comparing the predictive performance (i.e., AUROC and AUPRC) and fairness gaps (i.e., EOPP, EOD, and DP) of these methods. Personalized FL demonstrates superior predictive accuracy and fairness over standalone training across both datasets. Nevertheless, in comparison with standard FL, personalized FL shows improved predictive accuracy but does not consistently offer better fairness outcomes. For instance, in the 24-h in-hospital mortality prediction task, personalized FL achieves an average EOD of 27.4% across racial groups in the eICU dataset and 47.8% in MIMIC-IV. In comparison, standard FL records a better EOD

✉ Yuanxiong Guo
yuanxiong.guo@utsa.edu

Extended author information available on the last page of the article

of 26.2% for eICU and 42.0% for MIMIC-IV, while standalone training yields significantly worse EOD of 69.4% and 54.7% on these datasets, respectively. Our analysis reveals that personalized FL has the potential to enhance fairness in comparison to standalone training, yet it does not consistently ensure fairness improvements compared to standard FL. Our findings also show that while personalization can improve fairness for more biased hospitals (i.e., hospitals having larger fairness gaps in standalone training), it can exacerbate fairness issues for less biased ones. These insights suggest that the integration of personalized FL with additional strategic designs could be key to simultaneously boosting prediction accuracy and reducing fairness disparities. The findings and opportunities outlined in this paper can inform the research agenda for future studies, to overcome the limitations and further advance health equity research.

Keywords Health disparities · Group fairness · Federated learning · Personalization · Privacy

1 Introduction

Today's discourse on the digitization of healthcare has moved beyond the potential transformative effects of artifacts such as Electronic Health Records (EHR) to a vision of the future in which Artificial Intelligence (AI) is poised to significantly enhance healthcare practices and delivery [1–4]. Research advances in AI, particularly Machine Learning (ML) and Deep Learning (DL), have led to groundbreaking innovations that disrupt various fields in healthcare, including radiology, pathology, and genomics [5]. As ML model usage proliferates into many aspects of our lives, there is growing concern regarding their ability to cause harm by introducing biases in decision-making [6]. ML models are known, for example, to be susceptible to algorithmic biases that can exacerbate existing health disparities [7].

Health equity can be broadly defined as minimizing unjustifiable disparities in health and its determinants among groups of people with varying levels of underlying social advantage or privilege, typically dictated by their relative positions in society in terms of power, wealth, or prestige in comparison to other groups [8]. However, existing studies have reported that ML models can potentially generate biased outcomes that disproportionately affect certain socio-demographic groups [7, 9]. For instance, the work of [6] has revealed significant disparities in the performance of state-of-the-art clinical prediction models. These models have been found to underperform when applied to women, ethnic and racial minorities, as well as individuals with public insurance. Additionally, other research has highlighted biases in popular language models when trained on scientific articles. Notably, these models have been observed to generate biased recommendations in clinical note templates, suggesting hospitals for violent white patients and prisons for violent black patients [10].

The design and utilization of ML systems are marked by a series of cascading effects resulting from health disparities and discrimination. First, ML models rely heavily on large datasets, and any biases from existing practices, institutional policies, norms, and social science factors that affect these datasets can be reproduced in the algorithmic models that they generate [7]. However, numerous factors contribute to

health disparities, inevitably introducing biases into datasets. These factors encompass widespread disparities in living and working conditions, differential access to and quality of healthcare, systemic racism, and other ingrained patterns of discrimination, all of which are likely to leave disadvantaged groups vulnerable to disproportionate health risks [7]. Moreover, the datasets used to develop and validate ML models are often insufficient in representing the general population, particularly with regard to variations in the prevalence and incidence of diseases and their risk factors [11–13]. If datasets do not cover adequate representation of populations at higher risk of a disease, the trained prediction models used in clinical AI decision support may not be able to accurately detect the target disease systematically [7, 13]. Additionally, the tendency to create health data silos can exacerbate this challenge, particularly when privacy regulations such as the General Data Protection Regulation (GDPR) [14] and the Health Insurance Portability and Accountability Act (HIPAA) [15] restrict the sharing of medical data between parties [5]. In addition to the lack of representativeness and patterns of discrimination, ML models are also susceptible to bias stemming from institutional racism and the implicit biases of both AI developers and users. Such biases can influence decision-making during the design and deployment process, potentially integrating discrimination and prejudice into the deployment process and resulting products [7]. Therefore, it is crucial to incorporate fairness-aware algorithms in the design of ML models to ensure group fairness, which refers to the idea that ML models should not be biased against specific groups and should not reveal real-world discrimination.

Extensive research efforts such as those presented in [16–19] have focused on studying how ML models can exhibit bias against socio-demographic groups in centralized scenarios where a single entity possesses all the data, and various fairness-related methods have been developed to mitigate the impact of such biases and promote group fairness [19–22]. However, these works rely on the availability of the entire dataset at a central entity during training, while in real-world healthcare applications, data are typically owned by multiple parties who are restricted from sharing it due to privacy concerns [5, 23]. Federated Learning (FL) provides a promising solution by enabling parties to collaboratively learn a global model without sharing their data [23–28]. Nevertheless, the decentralized nature and data heterogeneity of FL make it challenging to address this issue by applying fair training solutions from centralized settings to federated settings. One of the key obstacles is the fact that data cannot be shared between parties in a federated setting. Additionally, there are often significant differences in data distributions across parties, as well as across groups. For example, in healthcare, different hospitals will have different patient populations with distinct demographic compositions including race or gender. Although several fairness-aware FL approaches have been designed to mitigate bias by focusing on achieving group fairness using a single global model [29–31], their capacity for substantial improvements is limited due to the heterogeneity of data distributions in FL systems. This limitation is rooted in the disparate impact of the global model learned in FL on group fairness across participating parties, leading to unequal fairness benefits for different clients [32]. In this case, despite the importance of group fairness in the context of FL, it has not been fully investigated due to various challenges.

Personalization is a technique that has been widely utilized in FL to mitigate the data heterogeneity issue across clients, a universal characteristic inherent in all real-world datasets [33–35]. Unlike general FL methods that only have a global model, personalized FL methods allow each client to have their own local personalized models, which brings better adaptability on local private datasets. As a result, personalized FL often outperforms standard FL in terms of prediction accuracy, especially under practical non-IID (Independent and Identically Distributed) scenarios where data distributions vary significantly across different clients, leading to dependencies and heterogeneities in the data [34, 35]. In such contexts, the assumption of IID data, where each data point is drawn from a uniform distribution and each sample is independent of others does not hold, making standard FL approaches less effective [36]. The unique aspect of personalized FL, where clients can learn their own personalized models, makes us wonder whether it has the potential to reduce bias in these personalized models compared to the one-size-fits-all global model. Therefore, we investigate the following questions in this paper: Can personalization in FL improve fairness for parties compared to standalone training (i.e., each party conducts local training using its own data without collaboration via FL)? Can it provide more benefits than standard FL in terms of fairness? Can personalization mitigate the disparate impact of the global model on fairness in FL?

In this paper, we provide an empirical analysis based on two real-world EHR datasets: eICU and MIMIC-IV. We show that personalization, intended to handle performance degradation caused by data heterogeneity across parties, is insufficient for sustaining fairness benefits, as compared to its performance benefits. Specifically, we observe that the model trained using personalized FL can, on average, yield greater fairness benefits than standalone training. We also demonstrate that personalization does not necessarily guarantee enhanced fairness benefits compared to standard FL.

The rest of this paper is organized as follows. To begin, we briefly review the extant literature on group fairness and personalization in FL. We then proceed to elaborate on the background of FL, personalized FL, and the definition of group fairness within FL. Following this, we introduce our experimental setup, followed by a comprehensive analysis of our empirical findings. Subsequently, we proceed to discuss the contributions and practical implications of our work, highlighting the potential benefits it offers to various stakeholders. We also suggest potential solutions to mitigate the bias in personalized FL. Finally, we summarize our work and suggest promising future research directions.

2 Related Work

2.1 Group Fairness in Centralized ML

The proliferation of ML algorithms in decision-making processes has resulted in a significant focus on fairness. Several definitions of fairness have been proposed, with particular emphasis on *group fairness* [16, 18]. Specifically, group fairness focuses on ensuring that a model's predictions and outcomes should be equitable regardless of demographic groups that are defined by sensitive attributes such as race and sex.

Extensive research has investigated how ML models can exhibit bias against socio-demographic groups in centralized scenarios where a single entity possesses all the data [20, 21, 37, 38]. Common approaches for realizing group fairness in centralized settings can be classified into three categories: preprocessing [17], in-processing [20, 21, 39], and post-processing [40]. Notably, in the particular data context of EHRs, [41] involves the utilization of adversarial learning to mitigate bias in an ML model tasked with predicting the risk of cardiovascular diseases from EHRs. Regularization techniques have been used to achieve counterfactual fairness [42], where the model is required to generate consistent predictions for a patient even when the value of their sensitive attribute is changed. Additionally, methods such as fine-tuning and pruning of pre-trained models have been proposed, especially for chest X-ray applications [43]. Nevertheless, the exploration of group fairness within the realm of FL using EHRs remains an avenue yet to be thoroughly investigated.

2.2 Group Fairness in FL

Recently, considerable progress has been made in training group fair models in FL [29–31, 44]. Most of these studies focus on achieving group fairness by measuring and mitigating bias utilizing a single global model [29–31]. For example, [29] derive a framework from a constrained multi-objective optimization perspective, wherein they seek the Pareto optimal model that achieves fairness constraints across all clients while maintaining consistent performance. Another study introduced a FL approach that incorporates fairness-aware aggregation and local debiasing techniques to improve group fairness within the FL setting [44]. However, in practice, the data distributions of different parties (i.e., hospitals) in healthcare are often heterogeneous. In this case, by learning a global model, these approaches could result in a decline in prediction accuracy for individual parties, because a global model may not accurately reflect the fairness of FL concerning the local data distributions of the parties [32]. A recent study [32] has revealed that FL can have a disparate impact on parties where those having more bias in the standalone setting (caused by local-only training on their local data) could obtain a fairer model through FL.

2.3 Personalization in FL

The existing literature on personalization in FL has primarily concentrated on evaluating the performance accuracy of FL methods across clients (e.g., parties or user devices), disregarding socio-demographic groups. Personalized FL has garnered significant attention as a potential solution to address the data heterogeneity inherent in FL [33–35, 45–47]. In the conventional design of FL, the objective is to train a global model using clients' data in a privacy-preserving and communication-efficient manner [23]. However, if the data distributions across clients are distinct (i.e., non-IID), the learned global model may not generalize well to each client's data [48]. This phenomenon has been reported in literature [49, 50], where an increase in statistical diversity leads to a significant increase in generalization errors of the global model on clients' local data. Consequently, multiple approaches have been proposed

to achieve personalization in FL, such as multi-task learning [33, 34], meta-learning [45, 46], representation learning [35, 51], etc. From the universal learning perspective, we can divide the existing personalized FL algorithms into two categories [52]: full model-sharing (with a global shared model) [34, 48] and partial model-sharing (without a global shared model) [35, 53]. Typically, full model-sharing algorithms are mainly extended from the conventional FL methods, i.e., FedAvg [23] or FedProx [54], which combines the adaption of local personalized features on local training updates procedure, such as regularized loss function [34, 48], model mixture [55], and meta-learning [56]. Full model-sharing also indicates knowledge from every client's local dataset could be transferred to local models of other clients by sharing the global model. Partial model-sharing often advocates for learning a shared representation across various clients and indicates each client could only utilize knowledge of partial model parameters trained on the other clients' local datasets so that each client could gain certain degrees of personalization [35, 53]. In general, regardless of the personalization techniques, existing work has demonstrated that learning personalized models for clients in the FL setting could work better than the global shared model or the local individual models when the data distributions across clients are highly non-IID as in the real-world datasets [57, 58].

3 Background

3.1 Standard FL

In standard FL, the goal is to learn a global model that achieves uniformly good performance over all clients [23]. Motivated by this goal, many existing methods, with the most common one being FedAvg [23], adopt a process that involves the following steps at each communication round: (i) the server selects a random subset of clients to participate in training, and delivers the current copy of the global model to them; (ii) each selected client computes a local model using its local dataset; and (iii) the server aggregates the local models received from clients to update the global model. The above process is repeated for multiple communication rounds until convergence.

Formally, a FL system contains a server and K clients, where each client $k \in [K]$ holds a local dataset D_k sampled from a distribution $(X_k, Y_k) \sim \mathcal{D}_k$. Here $X_k \in \mathbb{R}^d$ denotes the input feature vector and $Y_k \in \mathcal{Y}$ denotes the corresponding label. The sample size of the local dataset D_k is n_k. The goal of standard FL is to fit a single global model f parameterized by w across all clients as follows:

$$\min_{w} \frac{1}{K} \sum_{k \in [K]} L_k(w), \tag{1}$$

where $L_k(w) := \mathbb{E}_{(X_k, Y_k) \sim \mathcal{D}_k}[\ell_k(f(X_k; w), Y_k)]$ is the empirical risk of client k, and ℓ_k is the loss function. The workflow of a standard FL system is shown in Fig. 1.

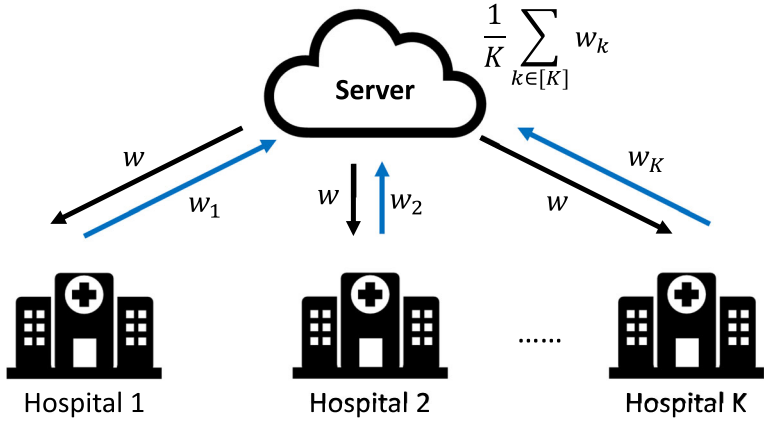

Fig. 1 The workflow of the standard FL system with K hospitals

3.2 Personalized FL

In practice, the local dataset D_k of each client may follow a distinct distribution $\mathcal{D}_k$. Therefore, it is common to consider learning personalized, client-specific models for all clients $k \in [K]$. In this work, we exclusively focus on full model-sharing among the two personalized FL categories due to its flexibility for personalization and strong connections with both global and local models. Under the full model-sharing category, we learn both a global model f parameterized by w and K local personalized models h_k parameterized by θ_k for each client k. The objective can be formulated as:

$$\min_{w,\{\theta_k\}_{k\in[K]}} \frac{1}{K} \sum_{k\in[K]} L_k(w; \theta_k), \tag{2}$$

where $L_k(w; \theta_k) := \mathbb{E}_{(X_k,Y_k)\sim\mathcal{D}_k}[\ell_k(f(X_k; w), Y_k)] + \lambda\mathcal{H}(w, \theta_k)$. Note θ_k represents local personalized model parameters owned by each client and w means the global reference model parameters shared among all clients. The new added $\mathcal{H}$ is the regularizer of similarity between w and θ_k and λ is the coefficient for each client.

Specifically, we adopt a common FL framework called Ditto [34] for personalization so that each client can obtain its own personalized model through multi-task learning in FL. We choose Ditto because of its flexible degree of personalization and its adaptability across different types of models as a full model-sharing method, in contrast to partial model-sharing methods [35, 53]. More importantly, compared to other full model-sharing methods [33, 48], Ditto's strength lies in the way it produces personalized models. The personalized models it produces acted as an interpolation between the global and local models, and Ditto can be viewed as a highly adaptable, lightweight personalization add-on for any global federated objective, which also maintains the privacy and communication efficiency of the global objective [34].

In Ditto, the global objective to obtain the global model is the same as that of the standard FL, and each local objective adds a regularization term to the empirical risk over the local dataset that encourages the personalized model of each client to be close to the optimal global model. Specifically, the bi-level optimization problem solved by each client $k \in [K]$ is defined as

$$\begin{aligned} &\min_{\{\theta_k\}_{k\in[K]}} L_k(\theta_k) + \frac{\lambda}{2}\|\theta_k - w\|^2 \\ &\text{s.t.} \quad w \in \arg\min_{w} \frac{1}{K}\sum_{k\in[K]} L_k(w), \end{aligned} \tag{3}$$

where λ is a hyperparameter that controls the interpolation between the global and local models, i.e., the personalization degree of each client. The training procedure for Ditto is presented in Algorithm 1. It involves the following steps at each round: (i) the server randomly chooses a subset of clients $\mathcal{S}^t$ for training and sends them the current global model parameter w^t; (ii) the selected clients update their personalized models' parameters θ_k through multiple local iterations using w^t as the reference with an added regularizer $\frac{\lambda}{2}\|\theta_k - w\|^2$, and update the received global model parameter w^t in the same way as in standard FL; and (iii) the updated global models with parameters w_k^t are sent back to the server for aggregation, while the personalized models with parameters θ_k are kept locally. This process is repeated for multiple rounds until convergence. At the end of the training, a global model with parameter w^* and K personalized models with parameters $\{\theta_k^*\}_{k\in[K]}$ are obtained.

Algorithm 1 Ditto for personalized FL with FedAvg being the aggregation strategy.

Input: Number of training rounds T, local iteration number s and r, global learning rate η_g, and personalized learning rate η.
Initialize w^0,$\{\theta_k^0\}_{k\in[K]}$
for *round* $t = 0, \ldots, T-1$ **do**
  Server randomly selects a subset of clients $\mathcal{S}^t$ and sends the current w^t to the selected clients
  `/* Client k Local Update */`
  for *client* $k \in \mathcal{S}^t$ *in parallel* **do**
    Update w_k^t for r local iterations:
      $w_k^t \leftarrow w^t - \eta_g \nabla L_k(w^t)$
    Update θ_k for s local iterations:
      $\theta_k \leftarrow \theta_k - \eta(\nabla L_k(\theta_k) + \lambda(\theta_k - w^t))$
    Send $\Delta_k^t := w_k^t - w^t$ back to the server
  end
  `/* Server Aggregation */`
  Server computes $w^{t+1} \leftarrow w^t + \frac{1}{|\mathcal{K}^t|}\sum_{k\in[\mathcal{K}^t]} \Delta_k^t$
end
return $\{\theta_k^T\}_{k\in[K]}$, w^T

3.3 Group Fairness in FL

In light of the definition of group fairness, which demands equitable and unbiased treatment of distinct groups by the model, we quantify group fairness within FL by focusing on three widely used group fairness notions, namely *Equal Opportunity* (EOPP) [18], *Equalized Odds* (EOD) [18], and *Demographic Parity* (DP) [16]. A model meets the DP fairness criteria when the predicted outcome doesn't rely on the sensitive attributes. However, pushing for DP might not always work well, especially if the actual outcome is tied to these sensitive attributes. To address this issue, EOPP aims to make sure that the predicted outcome is conditionally independent of the sensitive attributes, particularly when the target label is positive. EOPP goes a step further by ensuring that the True Positive Rates (TPRs) are the same for different groups. In the case of a binary target, EOD serves to guarantee that the predicted outcome is conditionally independent of the sensitive attributes, for every value of the target. This ensures both the TPRs and the False Positive Rates (FPRs) are equal, consequently ensuring that the False Negative Rates (FNRs) and the True Negative Rates (TNRs) are also equal.

Suppose there are K local clients, we denote a data sample from dataset D_k on the k-th client as (X_k, Y_k, A_k), where $A_k \in \mathcal{A}$ is the sensitive attribute, $X_k \in \mathbb{R}^d$ denote the input feature vector, and $Y_k \in \mathcal{Y}$ denote the true label. Recall that the goal of standard FL is to collaboratively learn a global model f with the parameters w to predict $\hat{Y}_k$ as $f(X_k; w)$ on each client. Similarly, personalized FL aims to learn a personalized model h_k with the parameters θ_k for each client k so that it can predict the target as $h_k(X_k; \theta_k)$ on each client. We can use the fairness gap with respect to the EOD difference on the k-th client to measure fairness. For example, when considering the personalized model h_k, it is defined as

$$\Delta^k_{\text{EOD}}(h_k; D_k) := \max_{a,a' \in \mathcal{A}, y \in \mathcal{Y}} |\Pr(\hat{Y}_k = 1 | A_k = a, Y_k = y) - \Pr(\hat{Y}_k = 1 | A_k = a', Y_k = y)|, \tag{4}$$

where $y \in \{0, 1\}$ for binary classification tasks. The EOD difference is essentially the largest gap in TPRs and FPRs between any two groups, denoted as a and a', considering all combinations of pairs from the involved groups. Consequently, in scenarios involving multiple groups, the EOD metric will identify the pair of groups exhibiting the most significant discrepancy. A classifier satisfies EOD if different groups have equal TPRs and FPRs. Δ^k_{EOPP} is a relaxed version of Δ^k_{EOD} that only considers positive labels, which is defined as

$$\Delta^k_{\text{EOPP}}(h_k; D_k) := \max_{a,a' \in \mathcal{A}} |\Pr(\hat{Y}_k = 1 | A_k = a, Y_k = 1) - \Pr(\hat{Y}_k = 1 | A_k = a', Y_k = 1)|. \tag{5}$$

EOPP ensures different groups have equal TPRs. Similarly, the fairness gap with respect to DP on client k is defined as follows

$$\Delta^k_{\text{DP}}(h_k; D_k) := \max_{a,a' \in \mathcal{A}} |\Pr(\hat{Y}_k = 1 | A_k = a) - \Pr(\hat{Y}_k = 1 | A_k = a')|. \tag{6}$$

DP ensures equal positive prediction rates across groups. In the rest of the paper, we use the fairness gap Δ (including Δ_{EOPP}, Δ_{EOD} and Δ_{DP}) to measure the bias (fairness) of a model. The model is considered fairer when the value of Δ is smaller.

4 Empirical Analysis

The goal of our empirical analysis is to understand how personalization in FL impacts group fairness for clients. We aim to answer the following questions:

- Can personalization in FL improve fairness for parties compared to standalone training?
- Can personalization provide more benefits than standard FL in terms of fairness?
- Can personalization mitigate the disparate impact of the global model on fairness in FL?

In this section, we first describe our experimental setup in Section 4.1, including the descriptions of the dataset, tasks, and model. Then we provide our empirical analysis and discuss our findings.

4.1 Experimental Setup

4.1.1 Datasets

The eICU Collaborative Research Database (**eICU-CRD**) [59] is a freely available, multi-center Intensive Care Unit (ICU) database. It comprises over 200,000 patient ICU encounters for 139,367 unique patients admitted between 2014 and 2015. Patients were admitted to one of the 335 units at 208 hospitals located throughout the United States. It is a collection of a number of tables, and the tables are all linked by a set of identifiers, and each instance in the database is a specific ICU stay. To simulate the distributed setting, we naturally partition the database into different hospitals. Figure 2

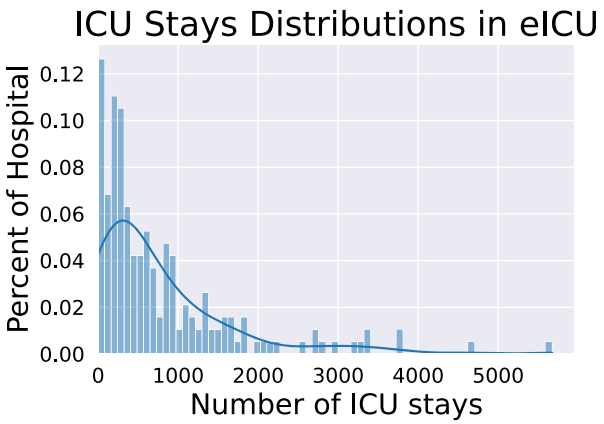

Fig. 2 ICU stay distributions across all hospitals in the eICU-CRD

depicts the distribution of the number of ICU stays among all hospitals. The database is highly imbalanced, with most hospitals having less than 1000 data samples.

To preprocess the raw data, we follow the data preprocessing pipeline outlined in [60]. Specifically, we select all adult patients (older than 18) with at least one recorded observation and an ICU stay lasting at least 5 h. In total, 17 static features are selected and subject to feature engineering, such as scaling numerical variables between -1 and 1 and converting categorical variables to one-hot encoding. Static features include patient and hospital attributes such as gender, race, age, admission location, etc., and all of them are preprocessed accordingly. Such static features could potentially impact the fairness of medical outcomes, as supported by existing literature [61, 62]. Therefore, their inclusion is also crucial for evaluating the model's performance across diverse patient groups and identifying potential biases that may arise from demographic disparities. Furthermore, diseases documented in the *pasthistory*, *admissiondx*, and *diagnoses* tables are extracted and represented through binary encoding. To maintain the hierarchical structure of diagnosis coding, separate features are assigned to each hierarchical level using binary encoding, as suggested in [60]. Each patient admission is represented by a single diagnosis encoding. Additionally, to enhance model performance, we include 87 time series features for each ICU stay from the following tables in the database: *lab*, *nursecharting*, *respiratorycharting*, *vitalperiodic* and *vitalaperiodic*. Time series features are extracted for every hour of the ICU stay, from 24 h before the ICU visit and up to the discharge time. Only variables presented in at least 12.5% of the total patient stays are included, or 25% for lab variables, as suggested in [60]. Moreover, time series features are then re-sampled according to 1-h intervals and then forward-filled over the gaps to cope with missing data, in order to handle the relatively sparsely sampled lab variables. Any data recorded before the ICU admission will be removed after forward-filling is complete. After that, corresponding decay indicators of time series features are added to specify how recently the observation was recorded, similar to the masking used in [63].

In this study, we conduct both 24-h and 48-h in-hospital mortality prediction. Specifically, one prediction will be made each hour, and we report the mortality prediction once per ICU stay (i.e., at 24 h or 48 h into the stay). For in-hospital mortality prediction, the distribution of the class label is highly skewed, with most data samples labeled as survival (i.e., negative class). To avoid lacking positive samples in each hospital, we only include hospitals that have more than 100 mortality cases (i.e., positive class) in the experiments. This threshold helps exclude hospitals that not only have fewer mortality cases but also tend to have a very limited overall data sample size. Such a criterion guarantees a diverse mix of positive and negative data samples in the training, validation, and testing sets, following a split of 70% for training, 15% for validation, and 15% for testing for each hospital's data. This diversity is crucial for avoiding datasets dominated by a single class, as the presence of varied data is fundamental for models to identify and learn from underlying patterns. After applying these steps, we ultimately include 42 hospitals in our empirical analysis. The distribution of mortality labels across the 42 hospitals is shown in Fig. 3.

We consider multi-value race, gender (i.e., male and female), and age (i.e., above 63 or not) as the sensitive attribute for both tasks. Figure 4 shows mortality distribution for each race subgroup across the 42 hospitals, and Fig. 5 shows the distributions of

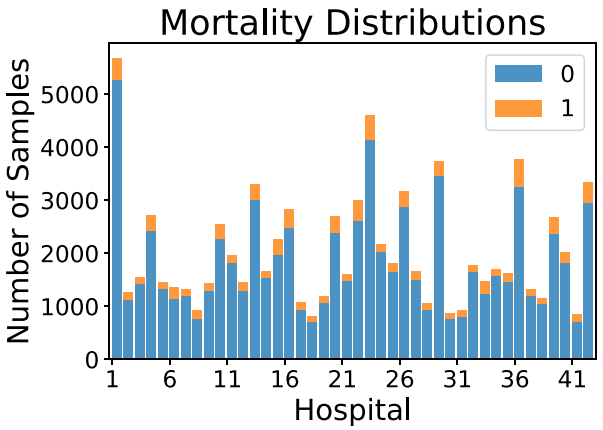

Fig. 3 Mortality distribution across selected hospitals in the eICU-CRD

mortality in gender and age subgroups across hospitals. The mortality distribution varies across race subgroups, and the Caucasian group has the most mortality data points. We observe that the mortality distributions for the two age groups are not evenly distributed, and the data is biased toward the group with age larger than 63. In contrast, the mortality rates are similar for the male and female groups, indicating that mortality is more evenly distributed and less bias exists.

We verify our results on a secondary dataset, the Medical Information Mart for Intensive Care (**MIMIC-IV**) database [64], a de-identified, publicly accessible EHR dataset sourced from the Beth Israel Deaconess Medical Center. It contains 69,619 ICU admissions involving 50,048 patients over the period from 2008 to 2019. We use the same cohort selection criteria as in eICU to extract 69,609 ICU stays from 50,042 patients. We followed the same feature selection process as in eICU to obtain a short

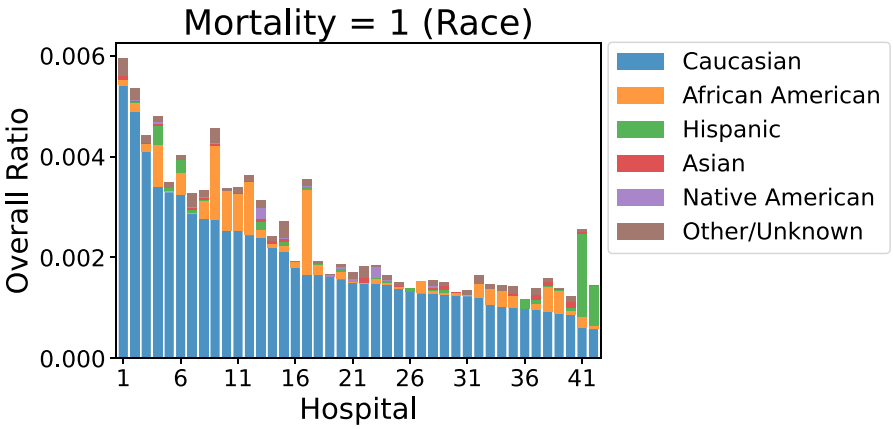

Fig. 4 Proportion of samples with positive mortality by race groups among samples from the selected hospitals in the eICU-CRD

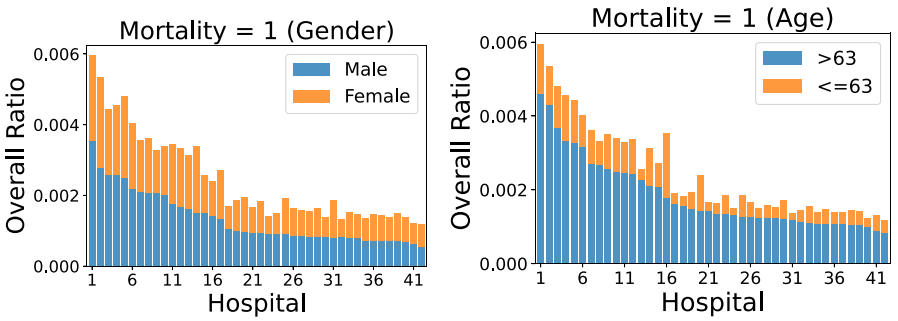

Fig. 5 Proportion of samples with positive mortality by gender and age groups among samples from selected hospitals in the eICU-CRD

list of 172 time series features from the *chartevents* and *labevents* tables. Out of these, 71 useless features from *chartevents* were manually excluded from the time series features because these variables do not capture the changes in a patient's condition, or because their distribution does not provide useful discrimination between patients, as introduced by [60]. The missing data is filled in the same way as in eICU. We eventually extracted 12 static features and 101 time series features for each ICU admission in MIMIC-IV. Given that MIMIC-IV contains data from a single medical center and it cannot be naturally partitioned, we created a non-IID cohort from this dataset, where we synthesize 20 non-IID clients through a Dirichlet distribution, in line with previous studies [65, 66]. The tasks performed on this dataset, as well as the sensitive attribute analyzed, are consistent with those used in the eICU dataset.

4.1.2 Model

For the eICU dataset, we employ the Temporal Pointwise Convolutional (TPC) model introduced by [60] in our experiments, which is the state-of-the-art model for patient outcome prediction using time series data. This model combines temporal convolutional layers that capture causal dependencies in the time domain, and pointwise convolutional layers that compute higher-level features from interactions in the feature domain to handle time series features. Specifically, the time series features and corresponding decay indicators are the initial inputs to the first TPC layer, and will be processed by N TPC layers, where the temporal convolution networks (TCN) [67, 68] will examine through regular timepoint t and map the X input channels into Y output channels, and the pointwise convolution will be applied separately to each timepoint t with information from static features. Besides, static features will be combined with time series representations among the feature domain using joint fusion. Finally, a two-layer pointwise convolution model is implemented, so that final predictions can be obtained.

In the context of the MIMIC-IV dataset, a Transformer-based model is employed, leveraging the capabilities of multi-head self-attention mechanisms, which has demonstrated superior performance across various tasks. Our model implementation is the same as [60].

4.1.3 Implementation Details

Our empirical analysis is conducted based on three methods briefly described below:

- **Standalone** training, where each client conducts local training using its own data without communicating with the central server using FL.
- **Standard** FL, where FedAvg is used to let clients collaboratively learn a global model for universal use.
- **Personalized** FL, where Ditto is implemented, in order to collaboratively learn local personalized models for individual use only, by interpolating the global model.

To compare standalone training, FedAvg (standard FL), and Ditto (personalized FL), we train the model for 30 communication rounds or epochs under each setting. The learning rate in Ditto for updating the global model is the same learning rate tuned on FedAvg, and it is set to be 0.001 on both datasets. The learning rate in standalone training is also 0.001. The batch size is 32 under all settings. We further tune the personalized learning rates and the hyperparameter λ in Ditto for updating personalized models. All the experiments were run three times, and the average performances with standard deviations were reported. The evaluation metrics used in our analysis are:

- **Performance** Metrics: We use the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) as metrics to measure prediction performance because the dataset is highly imbalanced.
- **Fairness** Metrics: We select three metrics EOPP, EOD, and DP to measure the fairness gaps defined in Section 3.3.

To measure the impact of personalized FL, we define the benefits of personalized FL in terms of fairness and accuracy using two baselines, i.e., standalone training and standard FL. The benefits of personalization in FL for each client are defined as the difference between the baseline model θ_{baseline} and the personalized model θ_k in terms of accuracy and fairness gaps. Specifically, for a client k, we define the accuracy benefit of personalized FL as $Acc(\theta_k) - Acc(\theta_{\text{baseline}})$, using AUROC and AUPRC as the metrics. Similarly, the fairness benefit of personalized FL is defined as $\Delta(\theta_{\text{baseline}}) - \Delta(\theta_k)$, where the fairness metrics can be EOPP, EOD, and DP. A positive benefit indicates that personalized FL improves fairness and accuracy, while a negative benefit implies that it has a detrimental effect.

4.2 Personalization Does Not Ensure Fairness Improvements in FL

Tables 1 and 2 present the average performance and fairness gaps of all methods for eICU dataset. For 24-h mortality prediction on eICU, as indicated in Table 1, personalized FL significantly enhances (p-value$<$0.0001) prediction performance in terms of AUROC (by 4.1%) and AUPRC (by 10.4%) in comparison to standalone training. Furthermore, it also significantly reduces (p-value$<$0.0001) fairness gaps across all groups. When compared with standard FL, personalized FL demonstrates a significant improvement of 0.3% in AUPRC (p-value$<$0.0001), although it does

Table 1 Average (standard deviation) performance and fairness gaps of standalone training, standard FL (FedAvg), and personalized FL (Ditto) in terms of **24-h in-hospital mortality** prediction across selected hospitals — eICU

Group	Methods	AUROC	AUPRC	Δ_{EOPP}	Δ_{EOD}	Δ_{DP}
Race	Standalone	.779 (.009)	.293 (.016)	.688 (.1)	.694 (.1)	.274 (.056)
	FedAvg	.817 (.004)	.394 (.01)	**.257 (.04)**	**.262 (.041)**	.061 (.007)
	Ditto	**.820 (.006)**	**.397 (.009)**	.271 (.032)	.274 (.035)	**.059 (.011)**
Gender	Standalone	.779 (.009)	.293 (.016)	.182 (.009)	.189 (.01)	.049 (.009)
	FedAvg	.817 (.004)	.394 (.01)	.117 (.021)	.119 (.022)	.015 (.003)
	Ditto	**.820 (.006)**	**.397 (.009)**	**.115 (.02)**	**.116 (.020)**	**.014 (.002)**
Age	Standalone	.779 (.009)	.293 (.016)	.259 (.008)	.267 (.012)	.070 (.017)
	FedAvg	.817 (.004)	.394 (.01)	**.149 (.004)**	**.150 (.003)**	**.015 (.002)**
	Ditto	**.820 (.006)**	**.397 (.009)**	.158 (.009)	.159 (.009)	.017 (.002)

The standard deviation across three runs is indicated between parentheses. Area under the curve (AUC) is measured for two prediction metrics (AUROC and AURPC) respectively. The highest prediction performance (AUROC and AUPRC) and lowest fairness gaps (Δ_{EOPP}, Δ_{EOD} and Δ_{DP}) are bold

not show a significant enhancement in AUROC (p-value>0.05). It is worth noting that the fairness gaps in terms of all metrics observed in personalized FL are not significantly better or worse than (p-value>0.05) standard FL across all groups. In other words, although personalized FL can outperform standard FL in terms of model performance, fairness improvements are not ensured compared to standard FL. Similar results can also be observed in Table 2 for the task of 48-h mortality prediction. One plausible explanation is that the personalized model, obtained through learning, is an interpolation between the standalone model and the global model. This model improves the prediction performance by incorporating the local data distribution to a

Table 2 Average (standard deviation) performance and fairness gaps of standalone training, standard FL (FedAvg), and personalized FL (Ditto) in terms of **48-h in-hospital mortality** prediction across selected hospitals — eICU

Group	Methods	AUROC	AUPRC	Δ_{EOPP}	Δ_{EOD}	Δ_{DP}
Race	Standalone	.746 (.009)	.307 (.019)	.461 (.058)	.501 (.057)	.268 (.036)
	FedAvg	.792 (.005)	.392 (.006)	**.118 (.014)**	**.124 (.015)**	**.05 (.011)**
	Ditto	**.794 (.003)**	**.396 (.006)**	.16 (.05)	.169 (.046)	.064 (.016)
Gender	Standalone	.746 (.009)	.307 (.019)	.195 (.011)	.198 (.011)	.05 (.007)
	FedAvg	.792 (.005)	.392 (.006)	**.084 (.003)**	**.088 (.004)**	**.016 (.002)**
	Ditto	**.794 (.003)**	**.396 (.006)**	.091 (.029)	.095 (.028)	.018 (.003)
Age	Standalone	.746 (.009)	.307 (.019)	.228 (.028)	.234 (.029)	.056 (.012)
	FedAvg	.792 (.005)	.392 (.006)	**.114 (.021)**	**.117 (.021)**	**.015 (.003)**
	Ditto	**.794 (.003)**	**.396 (.006)**	.136 (.048)	.14 (.046)	.019 (.003)

The standard deviation across three runs is indicated between parentheses. The highest prediction performance (AUROC and AUPRC) and lowest fairness gaps (Δ_{EOPP}, Δ_{EOD} and Δ_{DP}) are bold

larger extent, while at the same time, it also inherits certain biases from the standalone model.

The average results for the MIMIC-IV dataset are presented in Tables 3 and 4. Specifically, personalized FL significantly enhances (p-value$<$0.05) predictive accuracy for 24-h mortality prediction, increasing AUROC by 7.6% and AUPRC by 14.4%, compared to standalone training. It also significantly reduces (p-value$<$0.001) fairness gaps in terms of DP but the improvements are not statistically significant (p-value$>$0.05) for EOPP and EOD, with race as the sensitive attribute. In comparison with standard FL, there are no significant differences (p-value$>$0.05) in all prediction and fairness metrics, indicating that both methods offer comparable performance in predictive accuracy and fairness on the MIMIC-IV dataset. The results from both datasets indicate that personalization in FL does not necessarily ensure improvements in fairness.

4.3 Disparate Impact of Personalization on Group Fairness

To explore the accuracy and fairness benefits obtained through personalization, we show the average benefits of personalization across hospitals in Fig. 6. When compared with standalone training, we can observe that personalized FL can improve prediction performance and the variance in accuracy improvement across hospitals is small. Note that no significant difference has been observed in the AUROC. AUROC can be misleading for mortality prediction because the data is highly imbalanced, with the positive class being the minority class. AUROC can be easily influenced by the large number of true negatives, and this can result in a high AUROC score even if the model's performance on the minority class is poor. On the other hand, AUPRC considers both precision and recall, which are metrics that are sensitive to imbalanced data, especially when the positive class is the minority class. We also find that the

Table 3 Average (standard deviation) performance and fairness gaps of standalone training, standard FL (FedAvg), and personalized FL (Ditto) in terms of **24-h in-hospital mortality** prediction across clients — MIMIC-IV

Group	Methods	AUROC	AUPRC	Δ_{EOPP}	Δ_{EOD}	Δ_{DP}
Race	Standalone	.753 (.016)	.306 (.022)	.547 (.019)	.547 (.019)	.129 (.005)
	FedAvg	**.837 (.008)**	.433 (.013)	**.419 (.056)**	**.420 (.056)**	**.070 (.006)**
	Ditto	.829 (.022)	**.450 (.030)**	.478 (.042)	.478 (.042)	.097 (.005)
Gender	Standalone	.753 (.016)	.306 (.022)	.107 (.015)	.110 (.014)	.021 (.002)
	FedAvg	**.837 (.008)**	.433 (.013)	**.094 (.021)**	**.095 (.021)**	**.012 (.002)**
	Ditto	.829 (.022)	**.450 (.030)**	.124 (.013)	.125 (.013)	.020 (.003)
Age	Standalone	.753 (.016)	.306 (.022)	.131 (.020)	.135 (.021)	.040 (.004)
	FedAvg	**.837 (.008)**	.433 (.013)	**.107 (.009)**	**.109 (.010)**	**.017 (.003)**
	Ditto	.829 (.022)	**.450 (.030)**	.120 (.012)	.120 (.012)	.029 (.007)

The standard deviation across three runs is indicated between parentheses. The highest prediction performance (AUROC and AUPRC) and lowest fairness gaps (Δ_{EOPP}, Δ_{EOD} and Δ_{DP}) are bold

Table 4 Average (standard deviation) performance and fairness gaps of standalone training, standard FL (FedAvg), and personalized FL (Ditto) in terms of **48-h in-hospital mortality** prediction across clients — MIMIC-IV

Group	Methods	AUROC	AUPRC	Δ_{EOPP}	Δ_{EOD}	Δ_{DP}
Race	Standalone	.724 (.009)	.324 (.015)	.542 (.015)	.545 (.016)	.165 (.007)
	FedAvg	.775 (.004)	.434 (.021)	**.442 (.035)**	**.442 (.035)**	**.089 (.008)**
	Ditto	**.787 (.013)**	**.434 (.020)**	.525 (.029)	.525 (.029)	.132 (.025)
Gender	Standalone	.724 (.009)	.324 (.015)	.107 (.007)	.112 (.009)	.033 (.007)
	FedAvg	.775 (.004)	.434 (.021)	**.108 (.020)**	**.108 (.019)**	**.017 (.001)**
	Ditto	**.787 (.013)**	**.434 (.020)**	.077 (.017)	.079 (.016)	.020 (.002)
Age	Standalone	.724 (.009)	.324 (.015)	.145 (.004)	.151 (.002)	.048 (.007)
	FedAvg	.775 (.004)	.434 (.021)	**.109 (.019)**	**.110 (.019)**	**.017 (.001)**
	Ditto	**.787 (.013)**	**.434 (.020)**	.125 (.015)	.126 (.014)	.036 (.005)

The standard deviation across three runs is indicated between parentheses. The highest prediction performance (AUROC and AUPRC) and lowest fairness gaps (Δ_{EOPP}, Δ_{EOD} and Δ_{DP}) are bold

personalized model can provide more fairness benefits in terms of EOPP, EOD, and DP. It implies that most hospitals obtain a fairer model in personalized FL compared to standalone training. Furthermore, we notice that the variance in the fairness benefits across hospitals is large, suggesting that hospitals do not benefit from personalization in FL equally. Additionally, the personalized model does not yield substantial benefits in comparison to the global model learned from standard FL. While personalization can indeed enhance both predictive accuracy and fairness for certain hospitals, it leads to a more biased model for other hospitals. We also notice that there are variances in the fairness benefits across hospitals, further demonstrating that hospitals do not benefit equally from personalization.

To further explore the impact of personalization on fairness at the client (hospital) level, Fig. 7 shows the correlations between the fairness benefits across race groups that a hospital obtains in personalized FL and the fairness gaps of the standalone model for the hospital. Note that the fairness gap of the standalone model represents

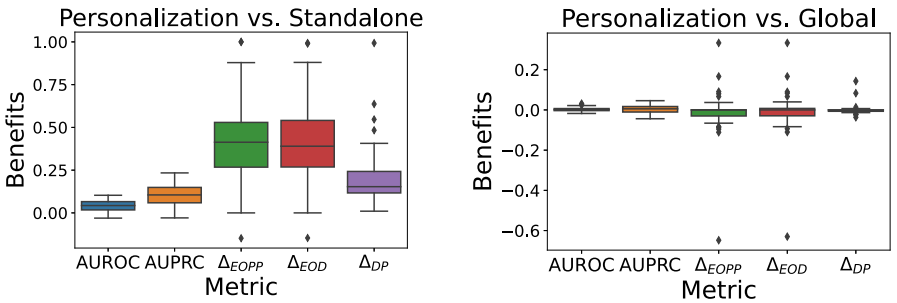

Fig. 6 Average performance and fairness benefits of personalized FL versus standalone training and standard FL for 24-h in-hospital mortality prediction — eICU (Race)

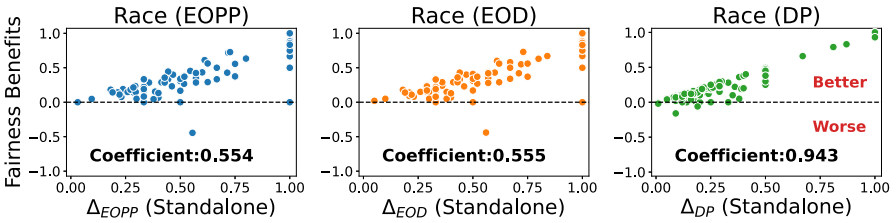

Fig. 7 Fairness benefit of personalized FL with regard to standalone training — eICU (Race)

the hospital's bias. It shows strong positive correlations between the fairness benefits obtained in personalized FL and the hospital's bias. This result indicates the benefits of personalized FL on fairness compared to the standalone training. This finding also highlights the disparate impact of personalization on fairness: it can improve fairness for more biased clients but at the cost of worsening the issue for less biased clients. Similar results can also be observed in Fig. 8 for gender and age groups.

However, when comparing the personalized model and the global model learned in standard FL, there is no strong correlation. Figure 9 shows the correlations between the fairness benefits across race groups a hospital obtains in personalized FL and the fairness gaps of the global model for the hospital. We can observe that personalized models and the global model can provide comparable fairness benefits for most hospitals. We can still find that while personalization can increase fairness for more biased clients in standard FL, it unfortunately tends to worsen the problem for those clients who are less biased. Similar results can also be observed in Fig. 10 for gender and age groups.

4.4 Personalized Model Learns Similar Patterns as Global Model

Figures 11 and 12 show the distribution of attribution values for the sensitive attribute "race" over individual test points from different race groups within the most biased hospital. We utilize Integrated Gradients [69] to quantify the contribution of each input feature to the models' predictions with respect to the positive class. Notably, the race attribute exhibits a higher attribution value in the standalone model's predictions compared to both the global and personalized models. This trend indicates a pronounced dependency of the standalone model's predictions on the race attribute of the test data, which could lead to biased outcomes, especially if the local dataset itself contains biases. On the other hand, the predictions of the global and personalized models

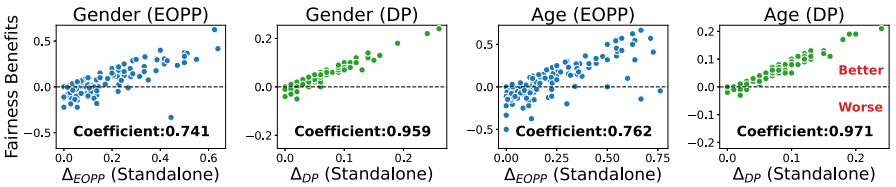

Fig. 8 Fairness benefit of personalized FL with regard to standalone training — eICU (Gender, Age)

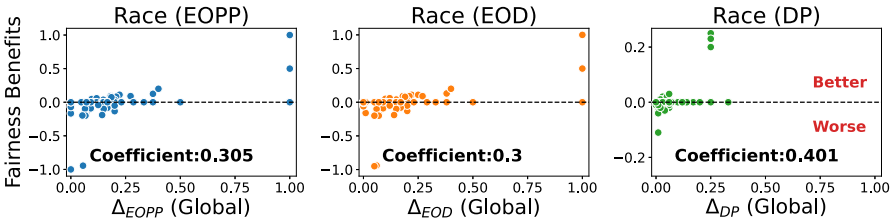

Fig. 9 Fairness benefit of personalized FL with regard to standard FL — eICU (Race)

depend less heavily on the sensitive attribute compared to those of the standalone model. The difference in attribution to the race attribute implies that standalone models tend to inadvertently learn more biased patterns compared to what could be learned in the FL setting. Specifically, standalone models, without the benefits of collaboration and aggregation provided by FL, have the risk of overfitting to biased patterns present in localized datasets.

Furthermore, the comparison between global and personalized models further enriches our understanding of bias in FL. Despite the expectation that personalized models would adapt more closely to individual data characteristics, our findings indicate that they tend to learn similar patterns compared to what could be learned from the global model. The similarity in pattern learning between personalized and global models raises critical questions about the effectiveness of personalization in advancing bias mitigation efforts. It appears that personalization alone is not enough to mitigate bias in FL. This finding calls for a deeper exploration of personalization strategies in FL, perhaps by integrating additional fairness-aware and debiasing mechanisms that can more effectively mitigate bias.

5 Discussion

In this section, we suggest some potential solutions to mitigate the bias in personalized FL and discuss some implications that our work can offer for different stakeholders. Within the realm of healthcare, privacy is an exceptionally significant concern. This is primarily because patient-sensitive information is subject to stringent privacy regulations that strictly prohibit its sharing. Therefore, FL emerges as a highly promising solution as it empowers multiple parties or clients to collaboratively train a ML model without the necessity of sharing their individual training data. Nonetheless, one

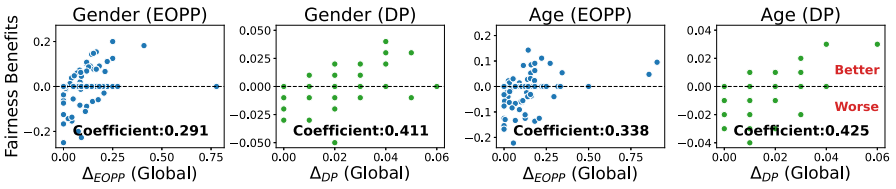

Fig. 10 Fairness benefit of personalized FL with regard to standard FL — eICU (Gender, Age)

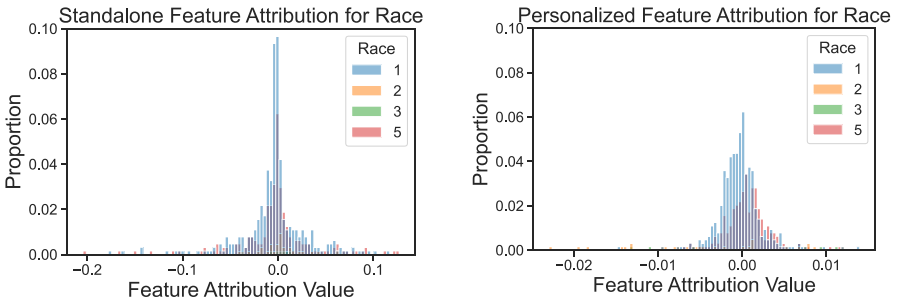

Fig. 11 Feature attribution value for the sensitive attribute of personalized FL versus standalone training for 24-h in-hospital mortality prediction — eICU (Race)

prominent issue is the possibility of the models trained via FL becoming biased against certain groups. Once there are significant differences in data distribution across parties, as well as across socio-demographic groups, the learned global model may have a disparate impact on group fairness across parties. Our paper takes the first step in this direction by providing a comprehensive analysis of the impact of personalization in FL on local group fairness for parties. By demonstrating that personalization in FL may exacerbate the issue of fairness for certain parties, we call for auditing group fairness in the personalized FL and designing fairness-aware learning algorithms that can mitigate biases. Specifically, a promising direction for future work is to identify which model parameters contribute to the bias and conduct a thorough audit of how bias propagates in this context. Another important direction is to design personalized FL algorithms capable of mitigating clients' local biases, such as by imposing extra constraints during training, and incorporating local reweighing techniques into the local training process, etc. Given that personalization in FL can achieve superior predictive performance, we encourage future research to explore the development of debiasing techniques, both at the local and global sides, to enhance fairness. This will

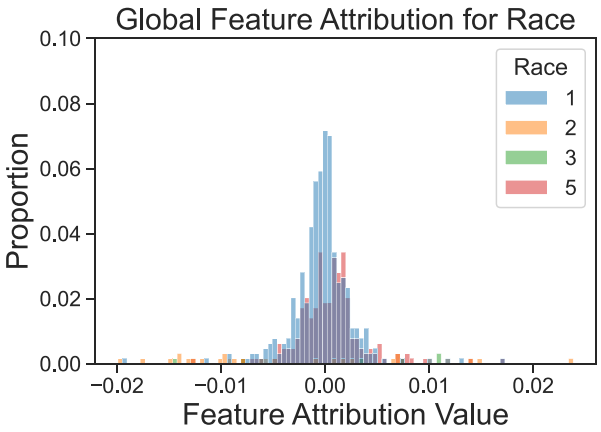

Fig. 12 Feature attribution value for the sensitive attribute of standard FL for 24-h in-hospital mortality prediction — eICU (Race)

enable clients in FL to acquire less biased models through a combination of fair local training and fair global aggregation.

Our study offers important practical implications for different stakeholders by providing insights into addressing bias within the personalized FL powered healthcare systems. Patients and healthcare recipients belonging to marginalized or underrepresented groups can enjoy more equitable access to healthcare services and personalized predictive models tailored to their unique needs while safeguarding their privacy through the adoption of such personalized FL healthcare systems. Specifically, beyond socio-demographic factors, research has underscored the substantial influence of socioeconomic, geographic, and environmental determinants on the health and well-being of individuals and communities [6]. These factors notably impact access to quality healthcare facilities, potentially exacerbating health disparities. In this context, harnessing a personalized FL-based healthcare system that incorporates these factors holds the potential to transform healthcare fundamentally. Such a system can extend its capabilities beyond mere prediction and become a powerful tool for the active characterization and correction of health disparities. Hospitals and healthcare providers can enhance the quality of care through collaborative efforts in personalized FL without concerns about becoming data donors. Additionally, they can boost fairness through the implementation of fairness-aware algorithms, potentially resulting in improved patient outcomes and contributing to achieving health equity. Ultimately, for society as a whole, these personalized FL systems have the potential to foster healthcare equity, ultimately resulting in enhanced overall healthcare outcomes and a more equitable society.

6 Conclusion

In this work, we have investigated how personalization affects fairness in FL through an empirical analysis on two real-world EHR datasets. Our findings have shown that, on average, models trained using personalized FL can achieve better fairness compared to standalone training. Additionally, we have found that personalized models and the global model can provide comparable fairness benefits for most hospitals but the benefits vary across hospitals. Specifically, personalization enhances fairness for more biased hospitals but at the cost of worsening the fairness issues for less biased hospitals. Our work suggests that a combination of personalized FL with fairness-aware design may have the potential to simultaneously improve prediction performance and decrease fairness gaps. Therefore, we encourage future research to further audit group fairness within this context and develop personalized FL algorithms that are capable of addressing group fairness issues.

Author Contributions Tongnian Wang: conception, implementation, analysis, and writing. Kai Zhang: writing support, and cross-reading. Jiannan Cai: writing support, and cross-reading. Yanmin Gong: conception, writing support, and cross-reading. Kim-Kwang Raymond Choo: Conception and working as co-supervisor. Yuanxiong Guo: providing ideas and working as supervisor. All authors contributed to the manuscript and reviewed it.

Funding The work of Y. Guo was partially supported by NSF CNS-2106761, CMMI-2222670, and UTSA Office of the Vice President for Research, Economic Development, and Knowledge Enterprise. The work of Y. Gong was partially supported by NSF CNS-2047761, CNS-2106761, and Cisco Research Award. The work of J. Cai was partially supported by NSF CMMI-2222670.

Availability of Data and Materials Data used in this study are openly available and free for research [59, 64].

Code Availability Code will be made available upon request.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

References

1. Purushotham S, Meng C, Che Z, Liu Y (2018) Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 83:112–134
2. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A (2019) Multitask learning and benchmarking with clinical time series data. Sci Data 6(1):96
3. Wang S, McDermott MB, Chauhan G, Ghassemi M, Hughes MC, Naumann T (2020) MIMIC-extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In: Proceedings of the ACM conference on health, inference, and learning, pp 222–235
4. Bhatt P, Liu J, Gong Y, Wang J, Guo Y (2022) Emerging artificial intelligence-empowered mhealth: scoping review. JMIR mHealth and uHealth 10(6):35053
5. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K et al (2020) The future of digital health with federated learning. NPJ Digit Med 3(1):119
6. Chen IY, Szolovits P, Ghassemi M (2019) Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 21(2):167–179
7. Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A (2021) Does AI stand for augmenting inequality in the era of covid-19 healthcare? BMJ 372
8. Braveman P (2006) Health disparities and health equity: concepts and measurement. Annu Rev Public Health 27:167–194
9. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R (2020) A review of challenges and opportunities in machine learning for health. AMIA Summits Transl Sci Proc 2020:191
10. Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M (2020) Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM conference on health, inference, and learning, pp 110–120
11. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178(11):1544–1547
12. Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, Koenig B, Ramos EM, Sorokin EP, Wand H et al (2018) The clinical imperative for inclusivity: race, ethnicity, and ancestry (rea) in genomics. Hum Mutat 39(11):1713–1720
13. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH (2018) Ensuring fairness in machine learning to advance health equity. Ann Intern Med 169(12):866–872
14. Voigt P, Bussche A (2017) The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing. 10(3152676):10–5555

15. Health UD, Services H (2013) Others: Modifications to the hipaa privacy, security, enforcement, and breach notification rules under the health information technology for economic and clinical health act and the genetic information nondiscrimination act; other modifications to the hipaa rules. Fed Regist 78(17):5566–5702
16. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp 214–226
17. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 259–268
18. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Adv Neural Inf Process 29
19. Agarwal A, Dudík M, Wu ZS (2019) Fair regression: Quantitative definitions and reduction-based algorithms. In: International conference on machine learning. PMLR, pp 120–129
20. Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A reductions approach to fair classification. In: International conference on machine learning. PMLR, pp 60–69
21. Roh Y, Lee K, Whang SE, Suh C (2021) Fairbatch: batch selection for model fairness. In: 9th International conference on learning representations
22. Chai J, Wang X (2022) Fairness with adaptive weights. In: International conference on machine learning. PMLR, pp 2853–2866
23. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. PMLR, pp 1273–1282
24. Wu X, Huang F, Hu Z, Huang H (2023) Faster adaptive federated learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 37, pp 10379–10387
25. Guo Y, Sun Y, Hu R, Gong Y (2022) Hybrid local sgd for federated learning with heterogeneous communications. In: International conference on learning representations
26. Hu R, Gong Y, Guo Y (2021) Federated learning with sparsification-amplified privacy and adaptive optimization. In: Proceedings of the thirtieth international joint conference on artificial intelligence
27. Wang T, Du Y, Gong Y, Choo K-KR, Guo Y (2023) Applications of federated learning in mobile health: scoping review. J Med Internet Res 25:43006
28. Wang T, Guo Y, Choo K-KR (2023) Enabling privacy-preserving prediction for length of stay in ICU-a multimodal federated-learning-based approach. In: European conference on information systems (ECIS)
29. Cui S, Pan W, Liang J, Zhang C, Wang F (2021) Addressing algorithmic disparity and performance inconsistency in federated learning. Adv Neural Inf Process Syst 34:26091–26102
30. Du W, Xu D, Wu X, Tong H (2021) Fairness-aware agnostic federated learning. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). SIAM, pp 181–189
31. Papadaki A, Martinez N, Bertran M, Sapiro G, Rodrigues M (2022) Minimax demographic group fairness in federated learning. In: 2022 ACM Conference on fairness, accountability, and transparency, pp 142–159
32. Chang H, Shokri R (2023) Bias propagation in federated learning. In: The Eleventh international conference on learning representations. https://openreview.net/forum?id=V7CYzdruWdm
33. Smith V, Chiang C-K, Sanjabi M, Talwalkar AS (2017) Federated multi-task learning. Adv Neural Inf Process Syst 30
34. Li T, Hu S, Beirami A, Smith V (2021) Ditto: fair and robust federated learning through personalization. In: International conference on machine learning. PMLR, pp 6357–6368
35. Collins L, Hassani H, Mokhtari A, Shakkottai S (2021) Exploiting shared representations for personalized federated learning. In: International conference on machine learning. PMLR, pp 2089–2099
36. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V (2018) Federated learning with non-iid data. Preprint at arXiv:1806.00582
37. Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, pp 329–338
38. Blum A, Stangl K (2020) Recovering from biased data: can fairness constraints improve accuracy? In: 1st Symposium on foundations of responsible computing
39. Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp 335–340

40. Kim MP, Ghorbani A, Zou J (2019) Multiaccuracy: Black-box post-processing for fairness in classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 247–254
41. Pfohl S, Marafino B, Coulet A, Rodriguez F, Palaniappan L, Shah NH (2019) Creating fair models of atherosclerotic cardiovascular disease risk. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 271–278
42. Pfohl SR, Duan T, Ding DY, Shah NH (2019) Counterfactual reasoning for fair clinical risk prediction. In: Machine learning for healthcare conference. PMLR, pp 325–358
43. Marcinkevics R, Ozkan E, Vogt JE (2022) Debiasing deep chest x-ray classifiers using intra-and post-processing methods. In: Machine Learning for Healthcare Conference. PMLR, pp 504–536
44. Ezzeldin YH, Yan S, He C, Ferrara E, Avestimehr AS (2023) Fairfed: Enabling group fairness in federated learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 37, pp 7494–7502
45. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. PMLR, pp 1126–1135
46. Khodak M, Balcan M-FF, Talwalkar AS (2019) Adaptive gradient-based meta-learning methods. Adv Neural Inf Process Syst 32
47. Hu R, Guo Y, Li H, Pei Q, Gong Y (2020) Personalized federated learning with differential privacy. IEEE Internet Things J 7(10):9530–9539
48. Dinh CT, Tran N, Nguyen J (2020) Personalized federated learning with moreau envelopes. Adv Neural Inf Process Syst 33:21394–21405
49. Li D, Wang J (2019) Fedmd: Heterogenous federated learning via model distillation. Preprint at arXiv:1910.03581
50. Deng Y, Kamani MM, Mahdavi M (2020) Adaptive personalized federated learning. Preprint at arXiv:2003.13461
51. Liang PP, Liu T, Ziyin L, Allen NB, Auerbach RP, Brent D, Salakhutdinov R, Morency L-P (2020) Think locally, act globally: Federated learning with local and global representations. Preprint atarXiv:2001.01523
52. Qin Z, Yao L, Chen D, Li Y, Ding B, Cheng M (2023) Revisiting personalized federated learning: Robustness against backdoor attacks. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '23, Association for Computing Machinery, New York, USA, pp 4743–4755
53. Li X, Jiang M, Zhang X, Kamp M, Dou Q (2021) FedBN: Federated learning on non-IID features via local batch normalization. In: International conference on learning representations. https://openreview.net/forum?id=6YEQUn0QICG
54. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2020) Federated optimization in heterogeneous networks. Proc Mach Learn Syst 2:429–450
55. Chen H-Y, Chao W-L (2022) On bridging generic and personalized federated learning for image classification. In: International conference on learning representations. https://openreview.net/forum?id=I1hQbx10Kxn
56. Fallah A, Mokhtari A, Ozdaglar A (2020) Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. Adv Neural Inf Process Syst 33:3557–3568
57. Li C, Niu D, Jiang B, Zuo X, Yang J (2021) Meta-har: Federated representation learning for human activity recognition. In: Proceedings of the web conference 2021, pp 912–922
58. Wu Q, Chen X, Zhou Z, Zhang J (2020) Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring. IEEE Trans Mob Comput 21(8):2818–2832
59. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O (2018) The eicu collaborative research database, a freely available multi-center database for critical care research. Sci Data 5(1):1–13
60. Rocheteau E, Liò P, Hyland S (2021) Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In: Proceedings of the conference on health, inference, and learning, pp 58–68
61. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453
62. Mauvais-Jarvis F, Merz NB, Barnes PJ, Brinton RD, Carrero J-J, DeMeo DL, De Vries GJ, Epperson CN, Govindan R, Klein SL et al (2020) Sex and gender: modifiers of health, disease, and medicine. Lancet 396(10250):565–582
63. Che Z, Purushotham S, Cho K, Sontag D, Liu Y (2018) Recurrent neural networks for multivariate time series with missing values. Sci Rep 8(1):1–12

64. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R (2020) Mimic-iv (version 0.4). PhysioNet. Available online at: https://physionet.org/content/mimiciv/0.4/. Accessed 13 Aug 2020
65. Hsu T-MH, Qi H, Brown M (2019) Measuring the effects of non-identical data distribution for federated visual classification. Preprint arXiv:1909.06335
66. Poulain R, Bin Tarek MF, 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
67. Kalchbrenner N, Espeholt L, Simonyan K, Oord Avd, Graves A, Kavukcuoglu K (2016) Neural machine translation in linear time. Preprint at arXiv:1610.10099
68. Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: A generative model for raw audio. Preprint arXiv:1609.03499
69. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International conference on machine learning. PMLR, pp 3319–3328

Authors and Affiliations

Tongnian Wang[1] · Kai Zhang[2] · Jiannan Cai[3] · Yanmin Gong[4] · Kim-Kwang Raymond Choo[1] · Yuanxiong Guo[1]

Tongnian Wang
tongnian.wang@utsa.edu

Kai Zhang
kai.zhang.1@uth.tmc.edu

Jiannan Cai
jiannan.cai@utsa.edu

Yanmin Gong
yanmin.gong@utsa.edu

Kim-Kwang Raymond Choo
raymond.choo@fulbrightmail.org

1 Department of Information Systems and Cyber Security, The University of Texas at San Antonio, San Antonio 78249, TX, USA

2 McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston 77030, TX, USA

3 School of Civil and Environmental Engineering, and Construction Management, The University of Texas at San Antonio, San Antonio 78249, TX, USA

4 Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio 78249, TX, USA

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com