Fair to Federate? Two-Sided Prediction of Performance and Fairness Outcomes in Healthcare Federated Learning

Marina Ceccon
marina.ceccon@phd.unipd.it
University of Padova
Padova, Italy

Alessandro Fabris
alessandro.fabris@units.it
University of Trieste
Trieste, Italy

Ornella Irrera
ornella.irrera@unipd.it
University of Padova
Padova, Italy

Gianmaria Silvello
gianmaria.silvello@unipd.it
University of Padova
Padova, Italy

Gian Antonio Susto
gianantonio.susto@unipd.it
University of Padova
Padova, Italy

Abstract

Federated learning (FL) has emerged as a promising paradigm for collaborative healthcare modeling under privacy constraints, yet institutions currently lack tools to anticipate how joining a federation will impact predictive performance and fairness. Existing work has primarily focused on post-hoc fairness mitigation, while overlooking the pre-collaboration decision itself. In this work, we introduce the Federation Assessment Problem, a two-sided prediction problem concerned with anticipating how federation would affect both a candidate institution seeking to join a federation and the institutions already participating in it, before collaboration occurs. To address this problem, we propose a two-stage meta-learning framework that predicts changes in performance and fairness metrics from institution-level metadata. We evaluate the framework on two healthcare imaging domains, across federations of varying sizes. Results show that federation-induced changes in both predictive performance and fairness can be estimated with high accuracy. The analysis further identifies label bias as one of the strongest predictors of fairness outcomes across both domains, often rivaling or exceeding demographic composition in importance. These findings demonstrate that informed federation assessment is possible without raw data sharing and highlight the need to explicitly measure multiple sources of bias when reasoning about fairness in healthcare FL.

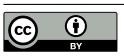
CCS Concepts

• **Computing methodologies** → **Learning paradigms**; **Distributed artificial intelligence**.

Keywords

Federated Learning, Meta-Learning, Fairness

ACM Reference Format:
Marina Ceccon, Alessandro Fabris, Ornella Irrera, Gianmaria Silvello, and Gian Antonio Susto. 2026. Fair to Federate? Two-Sided Prediction of Performance and Fairness Outcomes in Healthcare Federated Learning. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799682.3840981



CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840981

1 Introduction

Federated Learning (FL) [27], enables machine learning models to be trained across distributed data sources without centralizing raw data. Rather than transferring datasets to a single server, FL distributes the training process: clients perform local updates on their private data and share only model parameters or gradients, which are then aggregated into a global model by the server [25, 52].

Although originally motivated by large-scale mobile and edge-device scenarios with many small, intermittent participants (cross-device FL) [27, 52], FL extends to cross-silo settings, where a set of institutions collaborates on shared tasks while retaining custody of sensitive data [52]. Healthcare is the canonical cross-silo case where patient-level clinical data are subject to legal and ethical constraints that preclude centralized aggregation: regulatory frameworks such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) restrict the transfer and centralized storage of patient records [19, 34]. Moreover, single-site data are typically insufficient for the task. For instance, rare-disease cohorts at any one institution are too small to support robust deep learning [25, 41]; site-level case mix reflects regional demographics, referral patterns, and disease prevalence, biasing the empirical distribution and limiting external validity [50]; and clinical specialisation typically over-represents a hospital's focus pathologies while under-representing others [25]. Models trained under these conditions overfit to site-specific signals and may generalize poorly across centers [50], with the cost of poor generalization concentrated on the demographic and clinical subpopulations that are under-represented at the training site. FL responds to these structural conditions by enabling collaborative training while preserving data locality [25, 41] and by exposing the global model to heterogeneity unavailable at any single site.

Empirical work has demonstrated the effectiveness of FL across a broad range of clinical tasks [2, 4, 13, 24, 33, 36], with federated models reaching performance comparable to centralized training while preserving data privacy [38, 41]. Yet these successes mask a structural difficulty: the same site-level heterogeneity that makes FL valuable in healthcare also makes its outcomes hard to anticipate. In the FL literature, this heterogeneity is studied as the non-IID

condition [1, 28]. Beyond the per-site distributional framing, this condition has aggregation-level consequences for fairness: local biases can persist or amplify across communication rounds, and bias originating from a single client can propagate through the global model [6, 22, 26].

Distinct sources of bias can shape fairness outcomes in such federations. Demographic imbalance across clients has received substantial attention in the FL fairness literature, both as a driver of disparate performance and as a target for mitigation [14, 32, 53]. Instead, *label bias* – the systematic mislabeling of training samples in a way that correlates with a sensitive attribute – has received no comparable attention in FL. In the healthcare context, a dominant pattern is the under-diagnosis of pathologies in minority subgroups, which translates at the data level into a higher rate of false-negative labels for those subgroups [9, 30]. Evidence from centralized fairness studies suggests that label bias drives unfairness more strongly than demographic underrepresentation alone and that unexpected fairness phenomena can occur in its presence [5, 21], and existing FL work on client-level label noise [18, 43] does not address the group-targeted form of mislabeling that is most consequential in healthcare.

Together, these results raise a question that the federation literature has not addressed directly: *given the asymmetric and sometimes counterintuitive effects of label bias and demographic composition, how can an institution decide whether to join a federation, and how can a federation decide whether to admit it, before any training and without sharing patient-level data?* The decision is two-sided, since the candidate and the current members may have different (and occasionally opposed) interests [6, 8], and neither party can inspect the other's raw data or model weights.

To address this gap, we propose a meta-learning task that formalizes the prediction, for each performance and fairness metric of interest, of the per-client and per-federation gain (or loss) that would follow if a candidate client joined an existing federation, using only the high-level summary statistics that the two parties are willing to exchange.

The main contributions of this work are the following:

(1) We introduce the *Federation Assessment Problem* in FL, defined as predicting metric-change vectors symmetric between the candidate client and current federation members.
(2) We propose a two-stage estimation procedure that learns these metric-change vectors from controlled FL simulations and a compact set of size, demographic, and label-bias descriptors, without federated training at inference time.
(3) We evaluate the framework in two clinical imaging domains, multi-disease chest X-ray diagnosis and dermatology image classification, across federations of two, three, and five clients, and show that performance and fairness changes can be reliably predicted from both perspectives.
(4) We show that label bias is one of the strongest predictors, on par with or exceeding demographic composition for fairness outcomes in both domains.
(5) We show that the candidate and current federation members can have conflicting interests and that the framework makes such cases predictable from metadata alone.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the Federation Assessment Problem and presents the proposed framework. Section 4 describes the experimental setup and Section 5 presents the experimental results. Section 6 discusses the implications and limitations of the proposed approach. Finally, Section 7 concludes the paper and outlines future work. Code and supplementary material are available at `https://github.com/marinaceccon1/fair_to_federate/`.

2 Related Work

Federated learning. FL is a distributed machine learning paradigm that enables multiple parties to collaboratively train a shared model without directly exchanging raw data [20, 25, 52]. In a typical FL setup, participating entities, referred to as *clients*, optimize model parameters locally using their private datasets, while a coordinating *server* aggregates the resulting model updates into a global model [20, 27]. Existing FL deployments are typically categorized into *cross-device* settings, with many intermittently available edge devices, and *cross-silo* settings, with fewer stable institutions such as hospitals, banks, or research centers [20, 48]. Federated learning can also be categorized according to how data are partitioned across participants [20, 25]. In horizontal FL, clients share the same feature space while holding different samples. In vertical FL, clients share overlapping samples but operate on different feature spaces [19, 25]. We focus on horizontal FL, which is the most common scenario in healthcare.

Fairness in FL. Recent work has increasingly focused on fairness in FL, particularly in cross-silo environments where participating institutions often serve demographically distinct populations [6, 14]. Existing approaches primarily address fairness through modifications to aggregation procedures, optimization objectives, or training constraints. Early foundational work by Mohri et al. [29] introduced *Agnostic Federated Learning* (AFL), which formulates FL as a minimax optimization problem over mixtures of client distributions, thereby promoting robustness and fairness toward worst-case clients. q-FFL [23] proposed a flexible fairness objective that reweights clients according to their local losses, enabling explicit trade-offs between average accuracy and fairness. Recent work by Cui et al. [10] proposed FCFL, which jointly addresses per-client demographic fairness and cross-client performance consistency via constrained multi-objective optimization with Pareto guarantees.

Complementary approaches focus on fairness-aware aggregation and reweighting strategies such as FairFed [14] and FairFL [53]. Other approaches address fairness through alternative optimization formulations as FedMinMax [32] and mFairFL [42], which enforces group fairness via gradient conflict mitigation during aggregation.

Moreover, a growing body of work has examined how bias emerges and propagates in federated systems. Chang and Shokri [6] demonstrated that bias from a small number of biased clients can propagate through aggregation to all parties, improving fairness for heavily biased clients while worsening it for less biased ones. Heterogeneous client distributions can degrade fairness [31], with even a small number of highly skewed clients. Similar phenomena have been observed in healthcare federations [12] and in settings where local fairness objectives conflict with federation-wide optimization goals [8]. Collectively, these studies suggest that fairness in FL is

shaped not only by local data properties, but also by interactions between client heterogeneity and the aggregation process itself.

Unlike these approaches, we predict how fairness and performance will change for each party *before* federation, from institution-level metadata and without any federated training.

Label bias. Existing federated fairness research has largely focused on demographic imbalance and distributional heterogeneity while implicitly assuming that labels are reliable. In contrast, the centralized fairness literature has identified *label bias* as a major source of unfairness [5, 17, 21]. Recent studies in centralized settings show that label bias can have a stronger effect on fairness than demographic underrepresentation alone and that standard balancing interventions may become ineffective or even counterproductive when labels are systematically biased [5, 21, 45]. Despite the importance of these findings, the role of group-targeted label bias in federated settings remains largely unexplored.

Healthcare FL. Healthcare provides a particularly important context in which to study these issues [19, 35]. Prior work has demonstrated the effectiveness of FL for a wide range of medical applications, including medical imaging, disease diagnosis, and risk prediction [13, 36, 41]. More recent large-scale deployments further illustrate the feasibility of privacy-preserving collaborative learning in real-world clinical environments. Relevant examples are Swarm Learning [47], a decentralized framework combining federated optimization and blockchain coordination, which demonstrated improved diagnostic performance across distributed transcriptomic and imaging datasets and in computational pathology [37].

Algorithmic bias in clinical AI systems has been extensively documented [11, 30, 39], and several studies trace these disparities to historical patterns of underdiagnosis, undertreatment, and unequal access to care affecting disadvantaged populations [9, 30]. Obermeyer et al. [30] showed that a widely deployed commercial risk-scoring algorithm encoded racial bias because of the proxy used as the training label (annual healthcare spending). Analogous patterns are reported in radiology and dermatology, where chest pathologies in female patients and malignancies in darker skin tones have been historically under-detected and consequently under-labeled in clinical records [11, 40]. While conventional mitigation strategies often focus on adjusting dataset proportions, recent work demonstrates that standard balancing interventions can become ineffective when evaluated under real-world distribution shifts. Yang et al. [49] showed that, although algorithmic corrections can satisfy group fairness constraints within a local training distribution, these guarantees often fail to generalize to out-of-distribution external clinical datasets. This limitation shows that demographic composition alone cannot anticipate post-federation fairness outcomes and highlights the need to explicitly predict performance and fairness variations before collaboration.

Meta-learning. Meta-learning methods learn from the outcomes of previous learning experiments to predict future ones [44]. A prominent line of work applies this principle to algorithm selection, using compact dataset descriptors such as size, class balance and statistical moments as meta-features to predict which model or configuration will perform best on an unseen task [3]. More recent work has extended these ideas to the federated setting, treating each client as a distinct task [7, 15] and using meta-learning as a training algorithm to improve model quality within an existing federation. However, they do not address the question of whether federation should occur in the first place, nor do they consider the fairness implications of joining for any of the parties involved.

In sum, existing studies offer limited guidance on how institutions can anticipate the fairness and performance effects of joining a federation before collaboration, which is the focus of this work.

3 Methodology

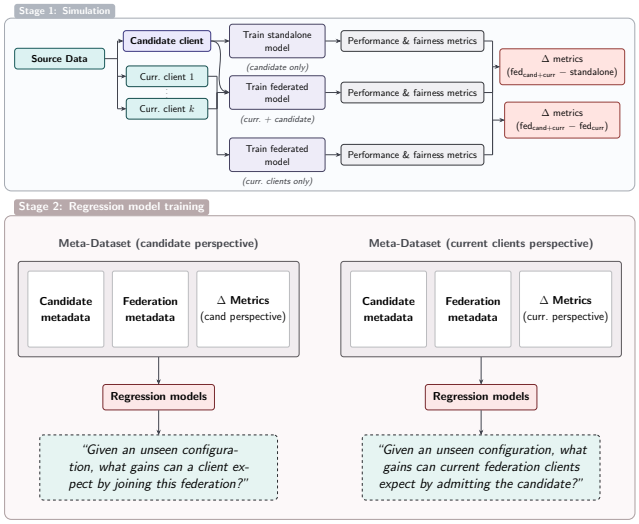


Figure 1: Overview of the proposed framework. Stage 1: for each configuration, three models are trained and evaluated for performance and fairness: (i) a standalone model on the candidate client, (ii) a federated model on the candidate and the current federation clients, and (iii) a federated model on the current federation clients alone. Stage 2: the outcomes form two parallel meta-datasets, one from the candidate client's perspective and one from the federations clients' perspective, are used to train regression models that predict federation gains for unseen configurations.

Let $C = \{c_1, \ldots, c_N\}$ be a set of clients, each holding a private local dataset $\mathcal{D}_i$. A *federation* $\mathcal{F} \subseteq C$ is a group of clients that train a shared model by exchanging only model parameters. Formally, the federation solves the weighted empirical risk minimization problem

$$\min_{\theta} \sum_{i \in \mathcal{F}} \frac{|\mathcal{D}_i|}{|\mathcal{D}_{\mathcal{F}}|} \mathcal{L}(f_\theta; \mathcal{D}_i), \tag{1}$$

where $|\mathcal{D}_{\mathcal{F}}| = \sum_{i \in \mathcal{F}} |\mathcal{D}_i|$ and $\mathcal{L}$ is the local loss. Equation (1) can be solved via FedAvg [27], which alternates between local gradient steps and server-side weighted averaging of client parameters. We denote by $M(\mathcal{S})$ the model obtained by training on a federation $\mathcal{S} \subseteq C$, and write $m_k(M(\mathcal{S}))$ for the value of metric m_k achieved by $M(\mathcal{S})$ on a shared held-out test set.

Let $c \notin \mathcal{F}$ be a *candidate client* that must decide whether to join the current federation $\mathcal{F}$, while $\mathcal{F}$ simultaneously decides whether to admit c. Both decisions rest on the expected changes in

a set of performance and fairness metrics $(m_1, \ldots, m_K)$ on which the federation is evaluated. Crucially, neither party can inspect the other's raw data or model weights; only high-level summary statistics, referred to as *client metadata*, may be exchanged.

Definition 1 (Federation Assessment Problem). *Given a candidate client c represented by metadata $\mathbf{x}_c \in \mathcal{X}_c$ and a federation $\mathcal{F}$ represented by metadata $\mathbf{x}_f \in \mathcal{X}_f$, the* Federation Assessment Problem *consists of estimating, from $(\mathbf{x}_c, \mathbf{x}_f)$ alone and without executing any federated training, the metric-change vectors*

$$\Delta^{\text{cand}} = \big(\Delta^{\text{cand}} m_1,\ \ldots,\ \Delta^{\text{cand}} m_K\big), \tag{2}$$

$$\Delta^{\text{curr}} = \big(\Delta^{\text{curr}} m_1,\ \ldots,\ \Delta^{\text{curr}} m_K\big), \tag{3}$$

where the component-wise deltas are defined as

$$\Delta^{\text{cand}} m_k = m_k\big(M(\mathcal{F} \cup \{c\})\big) - m_k\big(M(\{c\})\big), \tag{4}$$

$$\Delta^{\text{curr}} m_k = m_k\big(M(\mathcal{F} \cup \{c\})\big) - m_k\big(M(\mathcal{F})\big). \tag{5}$$

Δ^{cand} measures the per-metric gain (or loss) for c relative to training alone; Δ^{curr} measures the corresponding change for the current federation clients relative to operating without c.

The federated meta-learning task requires predicting all components of Δ^{cand} and Δ^{curr}. Based on this estimate, candidate and federation clients can take informed decisions, based on their priorities in metric trade-offs.

We represent each candidate client through three scalar attributes:

- $\text{size}_c = |\mathcal{D}_c|$: number of training samples;
- $\text{comp}_c = |\{x \in \mathcal{D}_c : a(x) = \text{maj}\}| \,/\, |\mathcal{D}_c|$: fraction of majority-group samples, where $a : \mathcal{D}_c \to \{\text{maj}, \text{mnr}\}$ is the sensitive attribute;
- $\text{bias}_c \in [0, 1]$: label bias rate, the fraction of minority-group positive samples whose labels are flipped to negative, modeling systematic underdiagnosis of minority populations [9, 30].

The candidate's feature vector is thus $\mathbf{x}_c = (\text{size}_c, \text{comp}_c, \text{bias}_c)$.

For a federation $\mathcal{F} = \{c_1, \ldots, c_n\}$, let $\text{maj}_i = |\mathcal{D}_i| \cdot \text{comp}_i$ and $\text{mnr}_i = |\mathcal{D}_i| \cdot (1 - \text{comp}_i)$ be the majority- and minority-group counts of client i. The federation meta-features are

$$\text{comp}_f = \sum_i \text{maj}_i\,, \tag{6}$$

$$\text{bias}_f = \sum_i \text{mnr}_i \cdot \text{bias}_i\,, \tag{7}$$

where comp_f is the total number of majority-group samples pooled across the federation and bias_f sums the per-client bias rates with weights equal to each member's minority-group size mnr_i. Because label bias acts only on the minority group, this weighting ties a member's influence on bias_f to its share of the federation's minority population. The feature vector is $\mathbf{x}_f = (\text{size}_f, \text{comp}_f, \text{bias}_f) \in \mathcal{X}_f$.

We seek two estimators

$$f^{\text{cand}} : \mathcal{X}_c \times \mathcal{X}_f \to \mathbb{R}^K, \qquad f^{\text{curr}} : \mathcal{X}_c \times \mathcal{X}_f \to \mathbb{R}^K, \tag{8}$$

approximating Δ^{cand} and Δ^{curr} respectively. Our approach to constructing and benchmarking these estimators, illustrated in Figure 1, proceeds in two stages.

Stage 1: Meta-dataset construction. We construct the training data for f^{cand} and f^{curr} through controlled simulation. For a broad range of candidate and federation configurations with varying (size, comp, bias) attributes, we train three model types per configuration:

(1) $M(\{c\})$: a standalone model on the candidate client alone;
(2) $M(\mathcal{F})$: a federated model on the federation clients alone;
(3) $M(\mathcal{F} \cup \{c\})$: a federated model on both the federation clients and the candidate jointly.

Models (1) and (3) yield Δ^{cand} via Equation (4); models (2) and (3) yield Δ^{curr} via Equation (5). The resulting pairs $\big((\mathbf{x}_c, \mathbf{x}_f),\ \Delta^{\text{cand}}\big)$ and $\big((\mathbf{x}_c, \mathbf{x}_f),\ \Delta^{\text{curr}}\big)$ populate two parallel tabular datasets sharing the same input meta-features but differing in their targets. For standalone models, we enumerate all valid combinations of client parameters. For federated configurations, we enumerate exhaustively when feasible; when the configuration space is too large, we resort to stratified sampling with the candidate configuration as the stratification variable, ensuring balanced representation across all candidate types. The specific parameter grids and sampling budgets are detailed in Section 4.

Stage 2: Regression model training. Since Δ^{cand} and Δ^{curr} are continuous-valued, the estimators f^{cand} and f^{curr} are instantiated as meta-regression models. We train one regressor per target component k, independently for the candidate and federation perspectives. Before training, the raw meta-feature vectors $\mathbf{x}_c$ and $\mathbf{x}_f$ are transformed into an augmented input representation. First, the candidate's composition and bias features are re-expressed as absolute counts to reflect their effective contribution to the minority-group annotation noise when joining the federation:

$$\widetilde{\text{comp}}_c = \text{size}_c \cdot \text{comp}_c, \tag{9}$$

$$\widetilde{\text{bias}}_c = \text{size}_c \cdot (1 - \text{comp}_c) \cdot \text{bias}_c. \tag{10}$$

The federation's aggregate meta-features size_f, comp_f, and bias_f are defined as size-weighted quantities above and are used directly without further transformation. Second, three interaction meta-features are appended to capture relative differences between candidate and federation:

$$r_{\text{size}} = \log \frac{\text{size}_f}{\text{size}_c}, \tag{11}$$

$$\delta_{\text{comp}} = \text{comp}_f - \widetilde{\text{comp}}_c, \tag{12}$$

$$\delta_{\text{bias}} = \text{bias}_f - \widetilde{\text{bias}}_c. \tag{13}$$

Third, a categorical meta-feature $n_{\mathcal{F}}$ encoding the total number of federation participants is included to account for scale effects across federation sizes. Table 1 summarizes the meta-features used by the estimators. Once fitted, the regressors provide estimates of Δ^{cand} and Δ^{curr} for any new configuration $(\mathbf{x}_c, \mathbf{x}_f)$ without requiring any additional federated training at inference time.

4 Experimental Setup

We instantiate the proposed framework across two clinical imaging domains, namely multi-disease chest X-ray diagnosis and dermatology image classification. For chest X-ray classification we use the multi-label dataset ChestX-ray14 (NIH) [46]. We adopt patient sex as the sensitive attribute and consider women as the minority

Table 1: Name, category, and meaning of each meta-feature used in the regression models.

Feature name	Category	Meaning
size_c	Size	Dataset size of the candidate client
size_f	Size	Aggregate dataset size of the federation
r_{size}	Size	Log-ratio of federation size to candidate size
$\widetilde{\text{comp}}_c$	Demography	Absolute number of majority-group samples in the candidate client
comp_f	Demography	Total number of majority-group samples pooled across federation members
δ_{comp}	Demography	Difference in majority-group sample count between federation and candidate client
$\widetilde{\text{bias}}_c$	Label bias	Effective number of flipped minority labels in the candidate client
bias_f	Label bias	Minority-count-weighted sum of per-client flip rates across federation members
δ_{bias}	Label bias	Difference between the federation's and candidate's weighted flip quantities
$n_{\mathcal{F}}$	Scale	Number of clients in the current federation

group, as they have been historically disadvantaged in the medical domain [39] and following previous fairness analyses on this dataset [5, 50]. For dermatology we use Fitzpatrick17k [16], where we binarize the coarsest-level skin condition taxonomy into tumoral and non-tumoral conditions, to obtain a binary classification task. We adopt skin tone as the sensitive attribute, binarized into light- and dark-skin groups for a more tractable fairness analysis. We consider the dark-skin group as the minority group, as darker skin tones have been historically underrepresented and disadvantaged in clinical practice [11]. We refer the reader to the online supplementary material for dataset composition and group prevalence statistics. For both datasets, data are partitioned into training (70%), validation (15%), and test (15%) sets.

In the chest X-ray setting, the training set size size_c is 3%, 6%, or 12% of the full dataset, and comp_c, defined as the fraction of majority-group (male) samples, is in $\{0.0, 0.5, 1.0\}$. The label bias fraction bias_c is drawn from $\{0.0, 0.15, 0.30, 0.45\}$, with clients having $\text{comp}_c = 1.0$ fixed at $\text{bias}_c = 0.0$. Each client in the federation is drawn from the same parameter grid as the candidate, i.e. taking values of size_c, comp_c, and bias_c from the same discrete sets defined above. The federation's aggregate meta-feature vector $\mathbf{x}_f = (\text{size}_f, \text{comp}_f, \text{bias}_f)$ is then computed from the individual client attributes.

In the dermatology setting, comp_c, defined as the fraction of light-skin samples, takes values in $\{1.0, 0.75, 0.50\}$, and size_c is set to $\{0.07, 0.08, 0.1\}$ of the initial dataset.[1] The flip fractions bias_c are the same as before, with the constraint that clients having $\text{comp}_c = 1.0$ have $\text{bias}_c = 0.0$.

The candidate grid yields 27 valid single-client configurations: 3 sizes $\times$ 3 compositions $\times$ 4 bias rates, minus 9 invalid cases where $\text{comp}_c = 1.0$ (no minority samples, forcing $\text{bias}_c = 0$). A two-client federation pairs a candidate with one current member, each from these 27 configurations, so we enumerate all $27 \times 27 = 729$ pairs per domain and train the federated model $M(\mathcal{F} \cup \{c\})$ for each, plus the 27 standalone models $M(\{c\})$ as single-client baselines. For three- and five-client federations, the number of member combinations grows combinatorially (about 10^4 and 10^6), making exhaustive enumeration infeasible, so we sample 800 configurations per federation size and domain using the candidate-stratified procedure of Section 3. This budget balances representation across federation sizes while retaining enough observations per candidate stratum for stable regression fitting. Federated training uses FedAvg [27] for 30 communication rounds with full client participation and one local epoch per round. The final model is selected based on the lowest weighted validation loss across clients, with weights proportional to local validation set sizes. All the details of model architectures, training hyperparameters, and data augmentation, along with additional details on the budget selection, are provided in the supplementary material to favor reproducibility of the results.

Both standalone and federated models are evaluated on a shared held-out test set using the following performance and fairness metrics. As performance metrics we use the Area Under the ROC Curve (AUC) and balanced accuracy. AUC is defined as $P(\hat{s}^+ > \hat{s}^-)$, where $\hat{s}^+$ and $\hat{s}^-$ denote the predicted scores of a positive and a negative sample, respectively. Letting TP, TN, FP, FN denote the standard confusion matrix counts, balanced accuracy is:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \qquad \text{BalAcc} = \frac{\text{TPR} + \text{TNR}}{2}.$$

As fairness metrics we use the TPR gap and minimum TPR across demographic groups:

$$\text{TPR}_{\text{gap}} = \text{TPR}_{\text{maj}} - \text{TPR}_{\text{mnr}}, \tag{14}$$

$$\text{minTPR} = \min\left(\text{TPR}_{\text{maj}}, \text{TPR}_{\text{mnr}}\right). \tag{15}$$

where a smaller TPR gap and a larger minTPR both indicate fairer behavior. These four metrics define the components of $\boldsymbol{\Delta}^{\text{cand}}$ and $\boldsymbol{\Delta}^{\text{curr}}$. Following the convention introduced in Section 3, the TPR gap target is negated so that a positive Δm_k uniformly corresponds to an improvement across all metrics. In the multi-label chest X-ray setting, AUC and balanced accuracy are macro-averaged across all pathological conditions, so that each pathology contributes equally to the final performance estimate regardless of its prevalence, as is common in multi-label medical imaging benchmarks [46]. Fairness metrics instead use micro-averaged TPR values, aggregating predictions across conditions before computing group-level rates. We adopt this formulation because it captures the overall disparity in correct predictions between demographic groups independently of any specific pathology, while also reducing the instability that fairness estimates may exhibit for rare conditions with very few positive samples. This is particularly relevant in chest X-ray datasets, where pathology prevalence is highly imbalanced and

[1] Larger client sizes or compositions with a higher fraction of dark-skin samples were not feasible given the limited availability of such samples in the dataset (only 14%).

fairness disparities can vary substantially across conditions and subpopulations [39, 40].

In the Fitzpatrick17k setting, the dark-skin group has relatively few positive samples in the test set, making the standard TPR a coarse and unstable regression target. We therefore replace the TPR with a soft variant defined as the average predicted probability over positive instances within each group:

$$\mathrm{TPR}_{\mathrm{soft}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \hat{p}_i, \tag{16}$$

where $\mathcal{P}$ is the set of positive samples in the group and $\hat{p}_i$ the predicted probability for sample i. Unlike the standard TPR, $\mathrm{TPR}_{\mathrm{soft}}$ varies continuously with model confidence, yielding a more stable regression target. Accordingly, the TPR-based components of Δ^{cand} and Δ^{curr} are computed using $\mathrm{TPR}_{\mathrm{soft}}$ in place of the standard TPR throughout the dermatology experiments. For brevity, all subsequent references to TPR in Fitzpatrick17k refer to $\mathrm{TPR}_{\mathrm{soft}}$.

Model selection and evaluation are performed via 5-fold cross-validation with shuffled splits, and hyperparameters are tuned through randomized search within the cross-validation loop. The suite of regression models comprises Linear Regression, Ridge Regression, Random Forest, Gradient Boosting, and XGBoost.

Regression performance is assessed using the coefficient of determination R^2:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{17}$$

where y_i is the observed federation gain, $\hat{y}_i$ the predicted value, and $\bar{y}$ the mean of observed values. $R^2 = 1$ indicates perfect prediction, $R^2 = 0$ corresponds to predicting the mean, and negative values indicate worse-than-mean predictions.

5 Results

5.1 Impact on the Candidate Client

Predicting federation gains. Figure 2 shows the actual versus predicted changes in performance and fairness metrics experienced by the candidate client upon joining a federation, for the ΔAUC and $\Delta\mathrm{TPR}_{\mathrm{gap}}$ targets, using the best-performing regression models – i.e., tree-based ensembles – trained on the combined dataset across federation sizes, for the chest X-ray setting (top) and the dermatology setting (bottom). Each point corresponds to a single federation configuration, with its predicted value obtained from the held-out fold in cross-validation, ensuring that no configuration is predicted by a model that has seen it during training. Consistent results for the remaining two targets (BalAcc and minTPR) are reported in the online supplementary material.

The results show that both performance and fairness gains from federation can be reliably predicted from client and federation metadata alone, though predictive success varies substantially across model families. Tree-based ensemble methods consistently achieve the best predictive accuracy across all targets, with predicted values close to the perfect-prediction diagonal in all four panels. For the NIH dataset, the best models attain $R^2 = 0.961$ for the ΔAUC and $R^2 = 0.941$ for the ΔTPRgap. Predictive accuracy remains high for Fitzpatrick17k, with $R^2 = 0.855$ for the ΔAUC and $R^2 = 0.944$ for the ΔTPRgap. In contrast, linear models achieve substantially lower accuracy and fail to capture the structure of the prediction problem: on NIH, they reach $R^2 = 0.776$ for the ΔAUC but yield $R^2 = 0.059$ for the $\Delta\mathrm{TPR}_{\mathrm{gap}}$, while on Fitzpatrick17k the corresponding values are $R^2 = 0.493$ and $R^2 = 0.620$. This gap shows that the link between federation metadata and predicted gains—especially for fairness targets—is inherently nonlinear and requires expressive ensemble methods to capture reliably.

A qualitative inspection of the configurations with the highest prediction errors shows that they systematically correspond to cases where some or most participating clients for NIH predominantly hold female samples, especially when these samples are also subject to considerable label noise. This is encouraging: stronger failures are confined to rare edge cases, while the vast majority of realistic configurations are predicted with high accuracy. Relative to Fitzpatrick17k, no systematic pattern of configurations for which the regression models consistently underperform is observed.

Analyzing the most predictive meta-features. Table 2 reports the mean drop in R^2 across five cross-validation folds when each group of meta-features is removed from the dataset, providing a measure of how much each group contributes to the predictions (feature-to-group assignments are detailed in Table 1). In Figure 3, we illustrate the SHAP beeswarms relative to the ΔAUC and the $\Delta\mathrm{TPR}_{\mathrm{gap}}$, relative to NIH and Fitzpatrick. Each dot in the plots represents one federation configuration; its horizontal position indicates the magnitude and direction of the corresponding feature's contribution to the predicted gain relative to the baseline average, and its color encodes the feature value from low (blue) to high (red).

For ΔAUC, size-related features are the main predictors in the NIH setting, consistent with the intuition that smaller clients benefit more from joining a large federation. In the Fitzpatrick17k setting, demographic composition is the primary driver: the beeswarm plot in Figure 3c shows that federating with a demographically balanced federation strongly improves the AUC, especially for candidates with imbalanced compositions. Label bias features also play an important role for the ΔAUC in Fitzpatrick17k, reflecting the benefit noisy clients gain from joining clean federations.

For the $\Delta\mathrm{TPR}_{\mathrm{gap}}$, demographic composition and label bias features play a comparably important role across both datasets. As shown in Figures 3b and 3d, the largest fairness gains for NIH arise when the candidate is demographically balanced and the federation contains a higher proportion of minority-group samples, whereas for Fitzpatrick17k the most favorable configurations correspond to a balanced candidate paired with a majority-group-heavy federation. In both cases, label bias features exhibit a consistent directional effect: candidates with higher label noise benefit most from joining federations with lower noise levels.

Scalability and generalization. Figure 4 reports the R^2 achieved by the best-performing model for each target on the NIH dataset, computed by restricting the held-out cross-validation folds to configurations corresponding to two-, three-, and five-client federations, respectively. Results for Fitzpatrick17k are analogous and reported in the supplementary material.

For the ΔAUC, R^2 decreases slightly from 0.97 at two clients to 0.94 at five clients. For the remaining targets, performance improves slightly with federation size, with R^2 rising from 0.79 to 0.90 for the ΔBalAcc, and from 0.90 to 0.97 for the $\Delta\mathrm{TPR}_{\mathrm{gap}}$ and the ΔminTPR. This consistency suggests that federation gains can

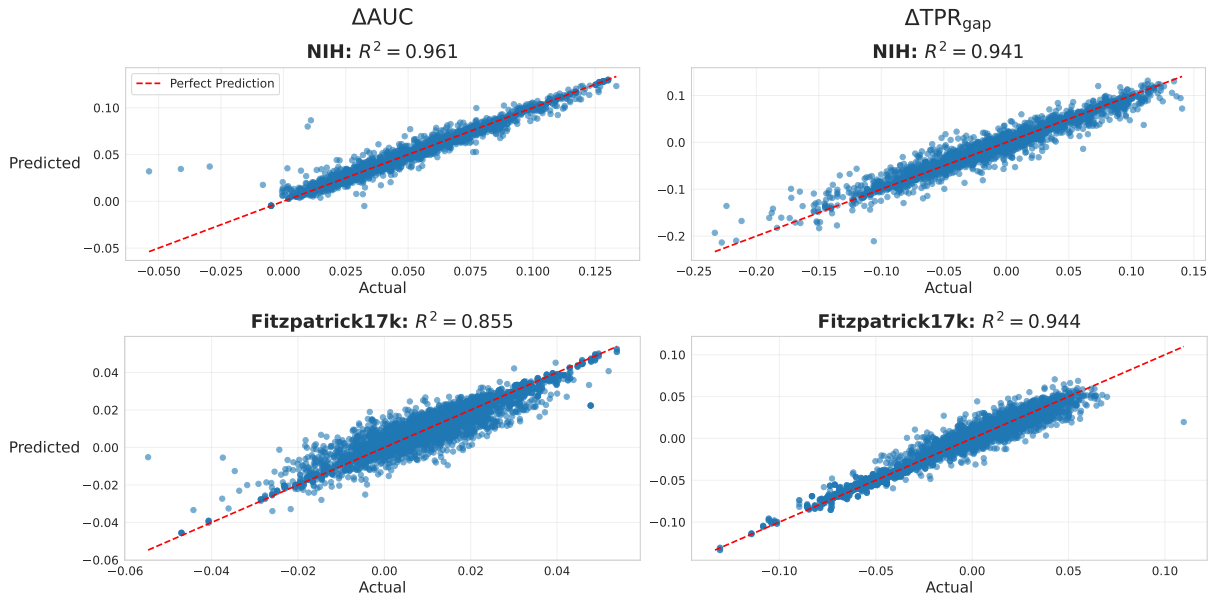


Figure 2: Actual vs. Predicted changes in ΔAUC and ΔTPR$_{gap}$ experienced by the candidate client upon joining a federation, for NIH (top) and Fitzpatrick17k (bottom) datasets.

Table 2: Mean R^2 drop (± standard deviation across five cross-validation folds) for each feature group and target, relative to NIH and Fitzpatrick17k. Results refer to the prediction of gains experienced by the candidate client upon joining a federation.

	NIH		Fitzpatrick17k	
Feature group	ΔAUC	ΔTPR$_{gap}$	ΔAUC	ΔTPR$_{gap}$
Size	0.188 ± 0.002	0.069 ± 0.012	0.037 ± 0.005	0.002 ± 0.002
Demographic composition	0.071 ± 0.008	0.330 ± 0.030	0.416 ± 0.021	0.176 ± 0.009
Label bias	0.097 ± 0.008	0.312 ± 0.043	0.190 ± 0.023	0.177 ± 0.010
Number of clients	0.008 ± 0.001	0.054 ± 0.012	0.000 ± 0.002	0.000 ± 0.001

be reliably predicted across federation sizes, and that representing the federation through aggregate descriptors rather than individual client characteristics does not degrade predictive performance. In contrast, linear models perform substantially worse across all targets and federation sizes. Their R^2 remains low for performance targets (0.68–0.78 for the ΔAUC and 0.27–0.55 for the ΔBalAcc) and near zero or negative for fairness targets (0.03–0.10 for the ΔTPR$_{gap}$ and as low as −0.28 for the ΔminTPR), confirming that the relationship between federation metadata and predicted gains is strongly nonlinear.

To test our framework's robustness, we assess whether models trained on smaller federations can predict outcomes for a larger, unseen federation size. We run 150 independent experiments with 7 clients on the NIH dataset and compute the true Δ targets as before. Models trained on the combined 2-, 3-, and 5-client configurations are then evaluated on this 7-client dataset without further training.

These experiments show that the models retain strong predictive capacity at larger scales. For the predictive targets, the Random Forest models achieve $R^2 = 0.674$ for the ΔAUC (vs. a training cross-validation score of 0.962) and $R^2 = 0.649$ for the ΔBalAcc (vs. a baseline R^2 of 0.852). Despite the performance gap relative to in-distribution results, these scores indicate that the aggregate metadata features remain highly informative beyond the federation sizes seen during training. The gap is narrow for fairness metrics, which remain highly consistent across sizes. The Gradient Boosting models achieve $R^2 = 0.901$ for the ΔTPR$_{gap}$ target (vs. 0.942 in cross-validation) and $R^2 = 0.865$ for the ΔminTPR target (vs. 0.973 on the training configurations). This accuracy shows that the relationships between group disparities and institutional metadata are stable, so federation-induced shifts can be reliably predicted even for larger federations.

5.2 Impact on the Federation Members

Figure 5 shows the analogous actual-versus-predicted results for the current federation members upon admission of the candidate: two- and three-client outcomes are compared against, respectively, the external standalone model and the two-client baseline; five-client configurations are excluded, as they admit no natural paired comparison under this formulation.

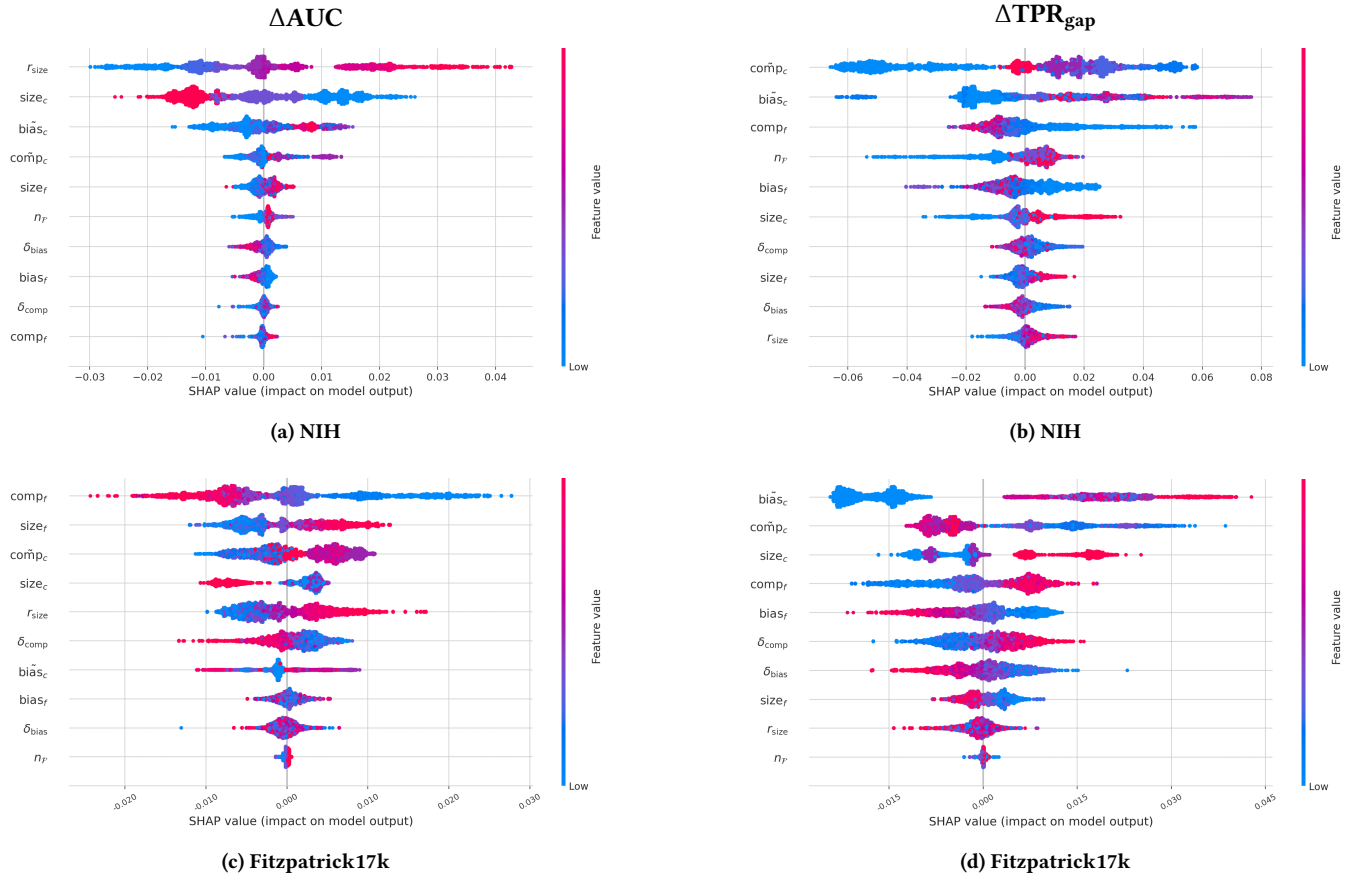


Figure 3: SHAP beeswarm plots illustrating meta-feature contributions to the predicted ΔAUC (left) and $\Delta\text{TPR}_{\text{gap}}$ (right) gains experienced by the candidate client upon joining a federation, for the NIH (top) and Fitzpatrick17k (bottom) datasets.

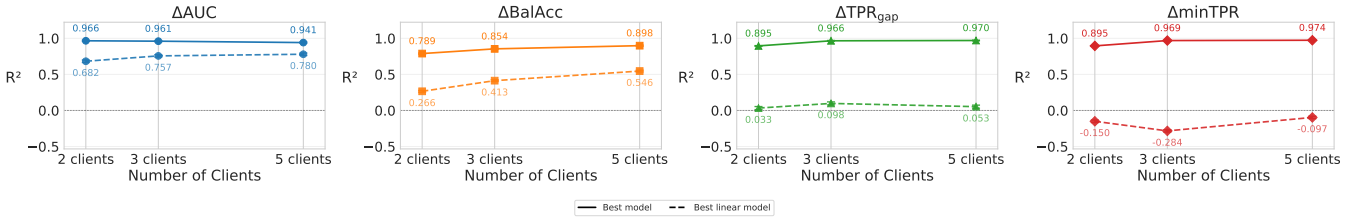


Figure 4: R^2 scores for each prediction target, evaluated separately on test points from two-, three-, and five-client federation configurations, relative to NIH. Results refer to the prediction of gains experienced by the candidate client upon joining a federation.

Tree-based ensembles (Random Forest and Gradient Boosting) perform best, with R^2 of 0.957 and 0.789 for ΔAUC and $\Delta\text{TPR}_{\text{gap}}$ on NIH, and 0.890 and 0.915 on Fitzpatrick17k. Linear models perform substantially worse, with $R^2 = 0.722$ and 0.046 on NIH, and 0.610 and 0.592 on Fitzpatrick17k. The near-zero R^2 of linear models on $\Delta\text{TPR}_{\text{gap}}$ relative to NIH aligns with the candidate-level results, confirming that the link between federation metadata and fairness outcomes is strongly nonlinear.

Table 3 shows the same pattern of feature importance as the candidate's perspective: size-related features drive performance only for NIH, especially for ΔAUC, while demographic composition and label bias dominate for $\Delta\text{TPR}_{\text{gap}}$ on NIH and for both metrics on Fitzpatrick17k.

The SHAP analysis reveals directional effects that mirror the candidate's perspective: current members benefit most from a large incoming candidate for size, and from a clean candidate when the federation itself is noisy for label bias. The demographic pattern is similarly symmetric, with the greatest gains arising from a balanced federation admitting a predominantly minority-group candidate

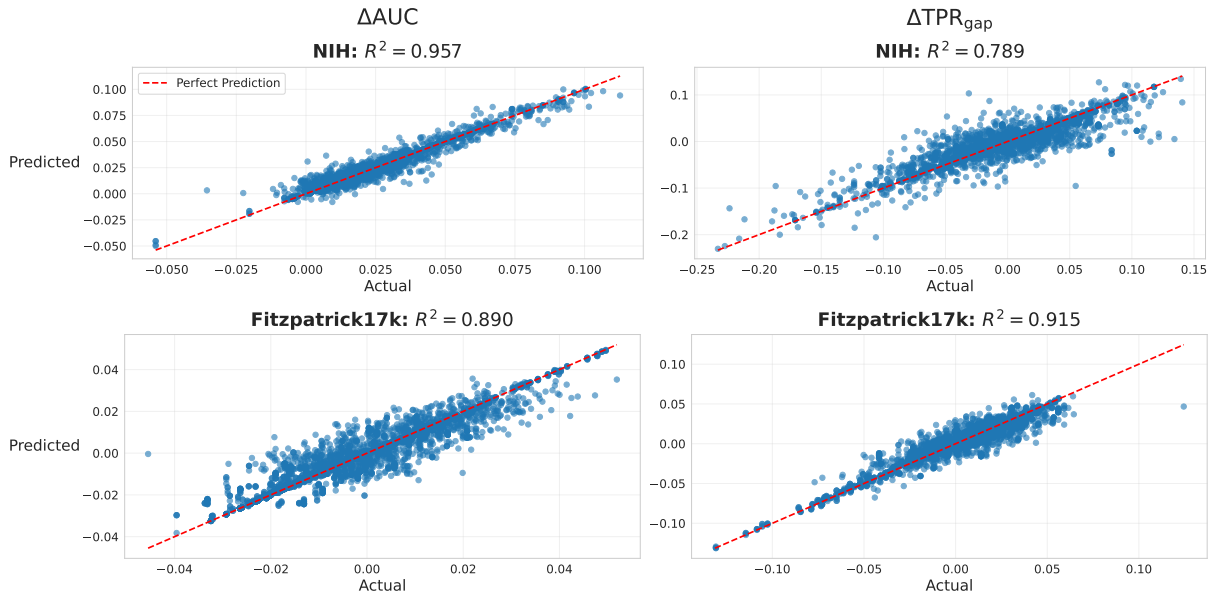


Figure 5: Actual vs. Predicted changes in ΔAUC and ΔTPR$_{gap}$ experienced by the current federation members upon admission of a new client, for NIH (top) and Fitzpatrick17k (bottom) datasets.

Table 3: Mean R^2 drop (± standard deviation across five cross-validation folds) for each feature group and target, relative to NIH and Fitzpatrick17k. Results refer to the prediction of gains experienced by the current federation members upon admission of a new client.

	NIH		Fitzpatrick17k	
Feature group	ΔAUC	ΔTPR$_{gap}$	ΔAUC	ΔTPR$_{gap}$
Size	0.202 ± 0.013	0.117 ± 0.031	0.003 ± 0.003	0.009 ± 0.005
Demographic composition	0.094 ± 0.021	0.337 ± 0.018	0.279 ± 0.028	0.208 ± 0.011
Label bias	0.117 ± 0.007	0.294 ± 0.030	0.153 ± 0.019	0.224 ± 0.016
Number of clients	0.019 ± 0.005	0.031 ± 0.010	0.000 ± 0.001	0.000 ± 0.001

on NIH, or a majority-group candidate on Fitzpatrick17k. This symmetry reflects a structural tension whereby the same configuration that benefits one party often disadvantages the other.

To quantify this, we measure the fraction of joint configurations yielding conflicting outcomes, i.e., where one party benefits at the expense of the other. On NIH, conflicts are rare for the ΔAUC (2%) but affect 53% of configurations for the ΔTPRgap, of which 24% benefit the candidate only and 29% the federation only. On Fitzpatrick17k, the conflict rate is more uniformly high, reaching 36% for the ΔAUC (19% candidate only, 17% federation only) and 51% for ΔTPRgap (31% candidate only, 20% federation only). These findings suggest that the tension between individual and collective incentives is dataset-dependent but consistently most pronounced for fairness metrics.

Finally, Figure 6 reports R^2 scores computed separately on two- and three-client held-out configurations for NIH (analogous Fitzpatrick17k results are in the supplementary material). Predictive accuracy decreases slightly for three-client configurations, suggesting that the impact on current members grows somewhat harder to predict in larger federations. Nevertheless, the best models retain strong accuracy across both sizes, with R^2 ranging from 0.75 to 0.97 at two clients and from 0.67 to 0.85 at three clients, suggesting that the aggregate metadata features capture stable relationships between institutional characteristics and federation outcomes, supporting the potential of our framework to generalize as federation size grows, within the same-source-dataset setting studied here. Linear models again lag far behind, with R^2 remaining near zero for fairness targets, further corroborating the inadequacy of linear estimators for capturing federation dynamics.

6 Discussion

Although earlier studies have examined how federation influences predictive accuracy and fairness [6, 14], none offers a principled framework for forecasting these effects before collaboration begins. Our results close this gap, showing that privacy and informed decision-making are not conflicting objectives. This predictive capability is especially valuable given that candidate and federation interests are not always aligned, consistent with prior work on bias

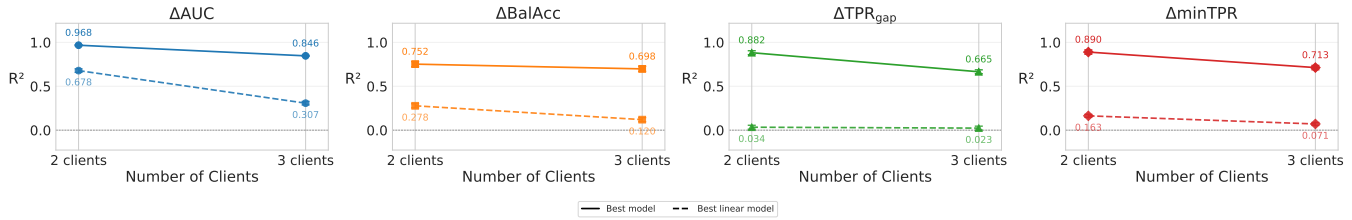


Figure 6: R^2 scores for each prediction target, evaluated separately on test points from two- and three-client federation configurations, relative to NIH. Results refer to the prediction of gains experienced by the current federation members upon admission of a new client.

propagation in FL [6] and the misalignment between client-level and federation-level fairness objectives [8].

In practice, this predictive utility makes the framework a promising tool for a pre-collaboration audit module: individual hospitals would not need to build their own Stage 1 meta-dataset, as a research consortium or public health body could instead build it once per task, using publicly available data for that task (e.g., existing chest X-ray collections for pathology diagnosis), and release the resulting Stage 2 predictor. Any institution considering federation for that task could then query this predictor directly, without running any simulation or training of its own. Large-scale collaborations such as the FeTS challenge [51] have documented severe post-hoc performance drops driven by site-specific variation that are difficult to anticipate from limited metadata; while our current evaluation does not directly test this form of cross-institutional heterogeneity, the underlying prediction task is structurally similar, and extending our framework to such settings is a natural direction for future validation.

Label bias plays a central role in shaping fairness outcomes: in the dermatology setup, candidate clients benefit in terms of TPR gap reduction from joining federations with higher majority-group prevalence, a counterintuitive finding explained by label bias, since minority-group data is frequently paired with higher annotation noise, hence compensating for demographic imbalance does not trivially improve fairness. Moreover, our feature importance analysis consistently highlights label bias as one of the most influential determinants of fairness outcomes in both datasets, consistent with and extending prior evidence from centralized settings [5, 21] to the federated context. As with any simulator-based analysis, this evidence is obtained under controlled conditions, in which label bias is injected explicitly and its rate is known by construction; in real deployments, where label bias cannot be observed directly and must instead be estimated, its precise impact remains to be quantified.

Limitations and future work. While the results are promising, some limitations warrant acknowledgment. First, although we evaluate our framework across two distinct medical imaging domains, extending it to other clinical settings remains future work. Second, in our experimental setup, all clients within a federation are constructed as disjoint samples from the same source dataset. This isolates the effect of demographic composition and label bias in a controlled manner, but omits heterogeneity present in real cross-institutional federations, where sites also differ in acquisition equipment and imaging protocols. Our reported predictive accuracy should therefore be read as a result relative to controlled bias conditions, not a guarantee under real-world heterogeneity. Third, our framework assumes access to accurate label bias estimates, which in practice must be inferred from data. A concrete path towards this is offered by Ceccon et al. [5], whose estimation method requires no unbiased ground truth and is directly compatible with federated deployment, since each institution can compute it on-site and share only a scalar summary. Discrepancies between estimated and true values nonetheless remain possible, and their effect on prediction quality deserves further investigation. Fourth, our current evaluation is restricted to a single binary sensitive attribute per dataset and to fairness notions built on the true positive rate, namely the TPR gap and the minimum TPR across groups. This keeps the demographic and fairness state space tractable and isolates the effect of label bias and demographic composition, but leaves out intersectional subgroups and clinically relevant notions beyond recall. We expect our framework to extend naturally along both directions. Fifth, our federation-level meta-features (size$_f$, comp$_f$, bias$_f$) summarize each federation through simple aggregates, which can in principle coincide for federations composed of very different individual clients. Future work could explore richer, permutation-invariant representations of the client set to capture such heterogeneity more faithfully. Finally, our framework does not account for client dropouts, a phenomenon that, while more common in cross-device settings [20], can also arise in cross-silo scenarios and represents a natural extension of this work.

7 Conclusions

In this work, we presented a framework for estimating how federation will affect predictive performance and fairness for all participating parties, using only high-level metadata, and showed that these effects can be anticipated with high accuracy across two medical imaging domains.

Acknowledgments

This work is partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

GenAI Usage Disclosure

The authors used generative AI tools, specifically Anthropic Claude and OpenAI ChatGPT, solely for checking grammar and spelling and for occasional rewriting of author-written text. These tools were not used for research ideation, study design, implementation, data analysis, or the generation of results. The authors reviewed all edited text and take full responsibility for the content of the paper.

References

[1] Navya Annapareddy, Yingzheng Liu, and Judy Fox. 2023. Towards fair and privacy preserving federated learning for the healthcare domain. *arXiv preprint arXiv:2308.01529* (2023).

[2] Sourasekhar Banerjee, Rajiv Misra, Mukesh Prasad, Erik Elmroth, and Monowar H. Bhuyan. 2020. Multi-diseases Classification from Chest-X-ray: A Federated Deep Learning Approach. In *AI 2020: Advances in Artificial Intelligence*, Marcus Gallagher, Nour Moustafa, and Erandi Lakshika (Eds.). Springer International Publishing, Cham, 3–15. doi:10.1007/978-3-030-64984-5_1

[3] Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, and Ricardo Vilalta. 2022. *Metalearning: Applications to Algorithm Selection and Hyperparameter Optimization*. Springer. doi:10.1007/978-3-030-67024-5

[4] Eoin Brophy, Maarten De Vos, Geraldine Boylan, and Tomas Ward. 2021. Estimation of continuous blood pressure from PPG via a federated learning approach. *Sensors* 21, 18 (2021), 6311. doi:10.3390/s21186311

[5] Marina Ceccon, Giandomenico Cornacchia, Davide Dalle Pezze, Alessandro Fabris, and Gian Antonio Susto. 2025. Underrepresentation, label bias, and proxies: Towards Data Bias Profiles for the EU AI act and beyond. *Expert Systems with Applications* 292 (2025), 128266. doi:10.1016/j.eswa.2025.128266

[6] Hongyan Chang and Reza Shokri. 2023. Bias Propagation in Federated Learning. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=V7CYzdruWdm

[7] Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. 2018. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876* (2018).

[8] Luca Corbucci, Xenia Heilmann, and Mattia Cerrato. 2025. Benefits of the Federation? Analyzing the Impact of Fair Federated Learning at the Client Level. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2232–2248. doi:10.1145/3715275.3732152

[9] James L Cross, Michael A Choma, and John A Onofrey. 2024. Bias in medical AI: implications for clinical decision-making. *PLOS digital health* 3, 11 (2024), e0000651. doi:10.1371/journal.pdig.0000651

[10] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1998, 12 pages.

[11] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* 8, 32 (2022), eabq6147. arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abq6147 doi:10.1126/sciadv.abq6147

[12] Êrica Peters Do Carmo, Caetano Traina-Jr, and Agma J. M. Traina. 2025. FairMedFL: Federated Learning for Fair and Unbiased Deep Learning in Medical Imaging. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*. 1–6. doi:10.1109/CBMS65348.2025.00158

[13] Q. Dou, T. Y. So, M. Jiang, Q. Li, L. Yu, L. Zhao, J. Qin, D. Wang, H. Chen, H. Chen, and P. A. Heng. 2021. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *npj Digital Medicine* 4 (2021), 60. doi:10.1038/s41746-021-00431-6

[14] Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. 2023. FairFed: enabling group fairness in federated learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 842, 9 pages. doi:10.1609/aaai.v37i6.25911

[15] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 300, 12 pages.

[16] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1820–1828.

[17] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics*. PMLR, 702–712.

[18] Xuefeng Jiang, Sheng Sun, Jia Li, Jingjing Xue, Runhan Li, Zhiyuan Wu, Gang Xu, Yuwei Wang, and Min Liu. 2024. Tackling Noisy Clients in Federated Learning with End-to-end Label Correction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 1015–1026. doi:10.1145/3627673.3679550

[19] Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. 2022. Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges. *ACM Trans. Comput. Healthcare* 3, 4, Article 40 (Nov. 2022), 36 pages. doi:10.1145/3533708

[20] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.* 14, 1–2 (June 2021), 1–210. doi:10.1561/2200000083

[21] Magali Legast, Toon Calders, and François Fouss. 2026. No evaluation without fair representation: Impact of label and selection bias on the evaluation, performance and mitigation of classification models. *arXiv preprint arXiv:2603.09662* (2026).

[22] Siqi Li, Qiming Wu, Doudou Zhou, Xin Li, Di Miao, Chuan Hong, Wenjun Gu, Yuqing Shang, Yohei Okada, Michael Hao Chen, et al. 2025. FairFML: fair federated machine learning with a case study on reducing gender disparities in cardiac arrest outcome prediction. *npj Health Systems* 2, 1 (2025), 29. doi:10.1038/s44401-025-00035-2

[23] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*.

[24] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. 2019. Privacy-preserving federated brain tumour segmentation. In *International workshop on machine learning in medical imaging*. Springer, 133–141.

[25] Bingyan Liu, Nuoyan Lv, Yuanchun Guo, and Yawen Li. 2024. Recent advances on federated learning: A systematic survey. *Neurocomputing* 597 (2024), 128019. doi:10.1016/j.neucom.2024.128019

[26] Disha Makhija, Xing Han, Joydeep Ghosh, and Yejin Kim. 2024. Achieving fairness across local and global models in federated learning. *arXiv preprint arXiv:2406.17102* (2024).

[27] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. Pmlr, 1273–1282.

[28] Usevalad Milasheuski, Luca Barbieri, B Camajori Tedeschini, Monica Nicoli, and Stefano Savazzi. 2024. On the impact of data heterogeneity in federated learning environments with application to healthcare networks. In *2024 IEEE conference on artificial intelligence (CAI)*. IEEE, 1017–1023.

[29] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. 2019. Agnostic federated learning. In *International conference on machine learning*. PMLR, 4615–4625.

[30] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. doi:10.1126/science.aax2342

[31] Mustafa Safa Ozdayi and Murat Kantarcioglu. 2021. The Impact of Data Distribution on Fairness and Robustness in Federated Learning. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. 191–196. doi:10.1109/TPSISA52974.2021.00022

[32] Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. 2022. Minimax demographic group fairness in federated learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 142–159. doi:10.1145/3531146.3533081

[33] Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G. Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. 2022. Federated learning enables big data for rare cancer boundary detection. *Nature Communications* 13, 1 (5 December 2022), 7346.

doi:10.1038/s41467-022-33407-5 Open access.

[34] Feng Qian and Andrew Zhang. 2021. The value of federated learning during and post-COVID-19. *International Journal for Quality in Health Care* 33, 1 (02 2021), mzab010. arXiv:https://academic.oup.com/intqhc/article-pdf/33/1/mzab010/42971424/mzab010.pdf doi:10.1093/intqhc/mzab010

[35] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine* 3, 1 (2020), 119. doi:10.1038/s41746-020-00323-1

[36] Holger R. Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C. Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B. Min Yun, Elshaimaa Sharaf, Katharina V. Hoebel, Jay B. Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta D. Pisano, Laura Coombs, Daguang Xu, Keith J. Dreyer, Ittai Dayan, Ram C. Naidu, Mona Flores, Daniel Rubin, and Jayashree Kalpathy-Cramer. 2020. Federated Learning for Breast Density Classification: A Real-World Implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M. Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, Nicola Rieke, Holger Roth, Daguang Xu, and Ziyue Xu (Eds.). Springer International Publishing, Cham, 181–191. doi:10.1007/978-3-030-60548-3_18

[37] Oliver Lester Saldanha, Philip Quirke, Nicholas P. West, Jacqueline A. James, Maurice B. Loughrey, Heike I. Grabsch, Manuel Salto-Tellez, Elizabeth Alwers, Didem Cifci, Narmin Ghaffari Laleh, Tobias Seibel, Richard Gray, Gordon G. A. Hutchins, Hermann Brenner, Marko van Treeck, Tanwei Yuan, Titus J. Brinker, Jenny Chang-Claude, Firas Khader, Andreas Schuppert, Tom Luedde, Christian Trautwein, Hannah Sophie Muti, Sebastian Foersch, Michael Hoffmeister, Daniel Truhn, and Jakob Nikolas Kather. 2022. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nature Medicine* 28, 6 (June 2022), 1232–1239. doi:10.1038/s41591-022-01768-5 Open access.

[38] Karthik V Sarma, Stephanie Harmon, Thomas Sanford, Holger R Roth, Ziyue Xu, Jesse Tetreault, Daguang Xu, Mona G Flores, Alex G Raman, Rushikesh Kulkarni, et al. 2021. Federated learning improves site performance in multicenter deep learning without data sharing. *Journal of the American Medical Informatics Association* 28, 6 (2021), 1259–1264. doi:10.1093/jamia/ocaa341

[39] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *Biocomputing 2021: Proceedings of the Pacific Symposium*. World Scientific, 232–243. doi:10.1142/9789811232701_0022

[40] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-Served Patient Populations. *Nature Medicine* 27, 12 (2021), 2176–2182. doi:10.1038/s41591-021-01595-0

[41] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports* 10, 1 (2020), 12598. doi:10.1038/s41598-020-69250-1

[42] Cong Su, Guoxian Yu, Jun Wang, Hui Li, Qingzhong Li, and Han Yu. 2024. Multi-dimensional fair federated learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Article 1682, 8 pages. doi:10.1609/aaai.v38i13.29430

[43] Vasileios Tsouvalas, Aaqib Saeed, Tanir Ozcelebi, and Nirvana Meratnia. 2024. Labeling chaos to learning harmony: Federated learning with noisy labels. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–26. doi:10.1145/3626242

[44] Joaquin Vanschoren. 2018. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548* (2018).

[45] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair Classification with Group-Dependent Label Noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 1–12. doi:10.1145/3442188.3445915

[46] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyuan Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2097–2106.

[47] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. 2021. Swarm learning for decentralized and confidential clinical machine learning. *Nature* 594, 7862 (2021), 265–270. doi:10.1038/s41586-021-03583-3

[48] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (2019). doi:10.1145/3298981

[49] Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. 2024. The limits of fair medical imaging AI in real-world generalization. *Nature medicine* 30, 10 (2024), 2838–2848. doi:10.1038/s41591-024-03113-4

[50] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* 15, 11 (2018), e1002683. doi:10.1371/journal.pmed.1002683

[51] Maximilian Zenk, Ujjwal Baid, Sarthak Pati, Akis Linardos, Brandon Edwards, Micah Sheller, Patrick Foley, Alejandro Aristizabal, David Zimmerer, Alexey Gruzdev, et al. 2025. Towards fair decentralized benchmarking of healthcare AI algorithms with the Federated Tumor Segmentation (FeTS) challenge. *Nature communications* 16, 1 (2025), 6274. doi:10.1038/s41467-025-60466-1

[52] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems* 216 (2021), 106775. doi:10.1016/j.knosys.2021.106775

[53] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. 2020. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 1051–1060. doi:10.1109/BigData50022.2020.9378043