HAVING IT BOTH WAYS: PERSONALIZATION AND GENERALIZATION IN FEDERATED EHR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Patient privacy regulations restrict the sharing of electronic health records (EHRs), limiting their use for clinical machine learning. While sharing realistic synthetic data can bypass these legal restrictions, EHR generation approaches trained on a single hospital's data generalize poorly to patient populations elsewhere, and their quality is often limited by the quantity of patient data available. To address these issues of data sparsity, federated learning enables training on data across many hospitals by only sharing model updates, allowing hospitals to learn statistical features of other hospital populations without leaking protected health information (PHI). In this paper, we first identify a trade-off inherent to this setting between generalization and personalization: joint training improves transfer across hospitals but degrades performance on each hospital's local distribution. We characterize this trade-off across a range of federated methods and propose FEDLIBRA, an approach that directly pushes the frontier of personalization–generalization, enabling the generation of high-quality patient data for any hospital.

1 INTRODUCTION

Electronic health records (EHR) encode a patient's trajectory across conditions, diagnoses, medications, and lab results, making them a critical and complex data type for clinical research. The potential for AI systems to inform chronic disease prediction, treatment dosing, and other clinical decisions cannot be overstated (Xiao et al., 2018).

As a form of highly sensitive protected health information (PHI), EHR carry extensive restrictions on their distribution under privacy regulations such as HIPAA (Gostin, 2001) in the United States and GDPR (European Parliament & Council of the European Union, 2016) in Europe. Generative models have been widely explored as a way around these restrictions, producing synthetic records that mirror real patient populations without leaking information about any individual (Pezoulas et al., 2024). Their development, however, is impeded by the very restrictions that motivated them. High-fidelity generative modeling (Oussidi & Elhassouny, 2018) requires both a sufficiently large dataset and a training distribution that mirrors the distribution at deployment (Manduchi et al., 2025). Individual hospitals typically do not hold enough data on their own, or are restricted by the same privacy laws that synthetic EHR aims to overcome (Pezoulas et al., 2024). The longitudinal nature of EHR compounds this difficulty (Xu et al., 2021; Ren et al., 2025; Xiao et al., 2018).

By never sharing patient data and communicating only model gradients, federated learning (FL) (McMahan et al., 2017) has been systematically developed in recent years to overcome these limitations (Xu et al., 2021), including for generative EHR (Bai et al., 2026), both to bypass data-sharing procedures and to increase accessibility of data that already exists across institutions. As sharing model weights and gradients can come with exposure risks (Stadler et al., 2022; Zhu et al., 2019), we investigate whether federated approaches leak information in practice through a nearest neighbor analysis (Yale et al., 2020).

However, challenges arise from naively federating EHR generation. Patient populations are highly diverse, meaning each client hospital distribution is complex and multimodal. Furthermore, patient populations can be effectively disjoint, making generalization and personalization difficult at once (Wu et al., 2023; Yang et al., 2023). In the multi-hospital eICU dataset (Pollard et al., 2018), training EHR models with FL leads to a trade-off (Figure 1) between personalization (hospitals fitting their own local distributions) and generalization (models that can represent patients from the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

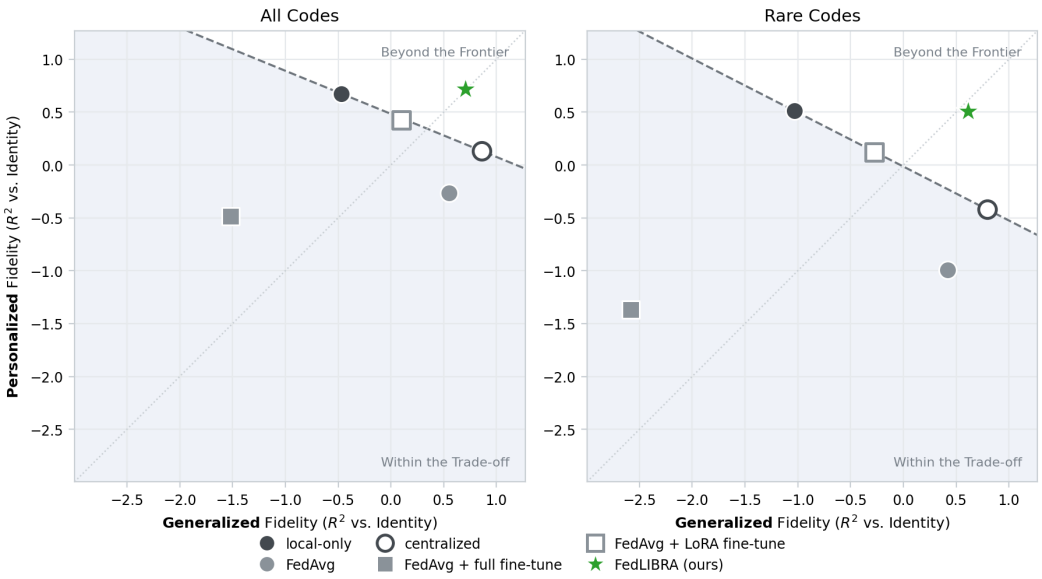


Figure 1: **FEDLIBRA advances the personalization–generalization frontier across patient distributions.** We evaluate each method on two axes: *personalization*, the average fidelity of each hospital's model on its own test set, and *generalization*, its average fidelity on the union of all hospitals' test sets. The frontier is bounded by two extremes: (1) a local-only baseline that trains a separate model for each of the 8 hospital clients, and (2) a centralized model trained on data pooled across all hospitals. Local models fit their own hospital's distribution but fail to generalize, whereas the centralized model performs well on the pooled test set but poorly on each hospital individually. A federated method should ideally perform well on both axes. However, standard federated averaging underperforms on both, and its personalized variant with local fine-tuning does not move beyond the frontier. This indicates that (1) federated training poses a distinct challenge for generalization, and (2) standard fine-tuning alone does not resolve the trade-off.

whole system of hospitals). We then ask: *How do we train generative models that fit local hospital distributions well, rather than overfitting to the average distribution across hospitals?*

Personalized federated learning directly aims to address this challenge (Li et al., 2020; 2021; Dinh et al., 2022). These methods live on a spectrum between two boundary points: isolated local training (`local-only`), which fits a hospital's own patients best and transfers worst; and centralized pooling (`centralized`), which represents the whole system best but fits each individual hospital worst (Figure 1). Cheng et al. (2023); Yu et al. (2020) show that locally fine-tuning the global federated model is the standard to beat for personalized federation. Strikingly, federated averaging with fine-tuning (FedAvg+FT) does not escape the trade-off: it lands on the straight line joining `local-only` and `centralized` in Figure 1. We call this line the personalization–generalization *frontier*: moving along it buys fit to one's own hospital only by giving up fit to the whole system.

This motivates a sharper question: can the frontier be broken rather than traversed? We show that it can, and to that end we propose FEDLIBRA, a federated training procedure for EHR generators that acts at two levels. *Across* clients, it aims to learn features invariant to the data source (Arjovsky et al., 2020), which are properties of realistic EHR sequences that hold at every hospital rather than on average, combating hospital heterogeneity. *Within* each client, adapter fine-tuning (Rebuffi et al., 2017; Hu et al., 2022) with best-of-K training (Gladstone et al., 2026; Guzmán-rivera et al., 2012; Yuan et al., 2023) lets local generators learn to cover multimodal client distributions. Together, rather than trading personalization for generalization, FEDLIBRA improves both.

Contributions.

C1 We characterize a clear trade-off between personalization and generalization for federated generative EHR models, supported by FL methods like FedAvg+FT.

C2 We push the frontier outward, improving generalization and personalization. Our method, FEDLIBRA, uses an IRM-style invariance penalty (Arjovsky et al., 2020) to learn common features despite heterogeneous clients, and best-of-K (Gladstone et al., 2026) adapter fine-tuning (Hu et al., 2022) to personalize without costing generalization performance, allowing for federated best-of-both world generators.

2 RELATED WORK

Our work draws on four lines of research: generative modeling of electronic health records (EHRs), personalized federated learning, invariant representation learning, and Best-of-K training objectives for diverse sample generation.

EHR generative modeling. Synthetic data generation has been studied broadly, both for health data (Rujas et al., 2025) and across other domains (Bauer et al., 2024). We focus on EHR generation, specifically the generation of discrete patient trajectories composed of medical codes (Slee, 1978; Organization et al., 2004), which are commonly explored in EHR modeling (Xiao et al., 2018; Ren et al., 2025). Approaches to this task fall mainly into two categories: generative adversarial networks (Choi et al., 2017; Torfi & Fox, 2020) and, more recently, autoregressive models (Wang & Sun, 2022; Theodorou et al., 2023; Vardhan et al., 2024; Pang et al., 2024), many of which build on language-model architectures such as GPT-2 (Radford et al., 2019). We adopt HALO (Theodorou et al., 2023) as our base architecture for its widespread use and accessible open-source implementation (Wu et al., 2026).

Personalized federated learning. FedAvg (McMahan et al., 2017) trains a single shared model by averaging the parameter updates of all clients. Personalized federated learning (perFL) instead adapts a model to each client's local distribution. Methods differ in how they constrain each personalized model's departure from the shared one: through a proximal penalty (Li et al., 2020; 2021), by decomposing the model into shared and private components (Collins et al., 2021; Liang et al., 2020), or through local fine-tuning of the shared model. Recent work finds that the simplest approach of FedAvg+FT remains a strong and consistent baseline across client distributions (Cheng et al., 2023; Yu et al., 2020). For synthetic EHR generation, however, we find that personalized methods, including FedAvg+FT, fail to extend the generalization–personalization frontier (Figure 1).

Invariant representation learning. Models trained across multiple distributions readily exploit spurious features, whose relationship with the target varies from one distribution to the next, limiting generalization regardless of architecture (Beery et al., 2018; Shah et al., 2020). Invariant representation learning addresses this by identifying features whose relationship with the target holds across all distributions (Peters et al., 2016). Invariant risk minimization (IRM) (Arjovsky et al., 2020) formalizes this idea as a training objective, adding a regularizer (IRMv1) that favors predictors that are optimal across all training distributions, or *environments*. In both centralized and federated settings, IRM has been studied almost exclusively for classification (Rosenfeld et al., 2021; Hao et al., 2025). To the best of our knowledge, we are the first to study invariance in generative modeling across patient distributions under the constraints of federated learning.

Best-of-K objectives. When a target distribution is composed of many distinct sub-distributions, standard reconstruction losses push a model toward the average of the modes, producing outputs that often match none of them. Best-of-K training addresses this by generating K candidates and back-propagating only through the one closest to the observed target, so that each candidate specializes to a different mode instead of collapsing to their mean. The idea originates in multiple choice learning (Guzmán-rivera et al., 2012), which updates only the lowest-loss member of a set of predictors, and has since been used to train diverse deep ensembles (Lee et al., 2016) and to predict multiple hypotheses in ambiguous vision tasks (Rupprecht et al., 2017). In generative modeling, best-of-K objectives have been used to learn diverse trajectories in motion forecasting (Bhattacharyya et al., 2018; Gupta et al., 2018), to select among samples via rejection sampling and best-of-N inference in language modeling (Yuan et al., 2023; Wu et al., 2025), and, more recently, to encourage exploration in diffusion models (Gladstone et al., 2026). These methods, however, assume that all modes are present in a single training set, so that the best-of-K selection can compare candidates against every kind of target. Federated EHR generation breaks this assumption: patient populations differ



Figure 2: Patient embeddings of (a) real patients taken from 2 eICU hospitals and (b), (c), (d), (e) synthetic patients of listed methods. We see in (a) how different patient distributions are. Notably, the centralized training method did not capture the patient distribution, with greater variation in synthetic patient embeddings. The two federated methods come closest to capturing the density of the real patients, with FEDLIBRA slightly closer to the density of the real data.

across hospitals, and the most dissimilar patients are, by construction, never observed together in the same local dataset. Our work asks whether best-of-K objectives can still recover these distinct modes when each client sees only its own.

Federated EHR generation. Federated learning on EHR data is well-established (Dasaradharami Reddy & Gadekallu, 2023; Kim et al., 2026), but federated training of EHR *generative* models is a recent development. To the best of our knowledge, Bai et al. (2026) is the only prior work in this setting, proposing the first federated generative model for EHR data. Our work differs in three ways: (1) we directly characterize patient heterogeneity across hospital clients and its effect on generation; (2) we provide a principled framing, through the generalization–personalization frontier, for why federated EHR generation is difficult; and (3) we publicly release our implementation in an open-source framework.

3 PRELIMINARIES

We consider a multi-hospital FL setting with K hospitals indexed $k \in 1 \dots K$. Each hospital k has its own private set of records $D_k = \{\boldsymbol{x}_k^{(1)}\}_{i=1}^{n_k}$, drawn from distribution P_k that can differ substantially between hospitals (Figure 2). We assume different hospitals may have very different n_k and denote $n = \sum_k n_k$. In addition, each hospital maintains a local generative model $h_k(\theta_k; \boldsymbol{x}_k)$ where θ_k are its local parameters.

Let $\boldsymbol{x}_k^{(i)}$ be the trajectory of patient i from hospital k. In this work, we treat a single trajectory as a patient's entire history in one hospital. Visits are encoded as multi-hot vector listings $\boldsymbol{v}_k^{(i)} \in \{0,1\}^m$, in which m is the size of the cross-hospital vocabulary. We use $L_k(\theta) = \frac{1}{n_k} \sum_{\boldsymbol{x}_k \in D_k} \ell(\theta; \boldsymbol{x}_k)$ to represent the empirical risk on hospital k, for a model with parameters θ; for example, in training, ℓ is binary cross-entropy for our autoregressive generator (Theodorou et al., 2023).

Federated Generation In standard FL a central server holds no data of its own and maintains a global model $h_g(\theta_g; \cdot)$. FedAvg (McMahan et al., 2017) minimizes a single pooled risk over the per-hospital risks of the whole system, $L_{\text{global}}(\theta_g) = \frac{1}{K}\sum_{k=1}^{K} \frac{n_k}{n} L_k(\theta_g)$. At each round t, hospitals perform local training and transmit updates to the server, $\theta_k^{t+1} - \theta_k^t$. Each hospital maintains a model h_k that can generate a synthetic dataset, S_k, which is not subject to the same privacy restrictions as $D_k$and can be shared with other hospitals.

To personalize, we decouple each hospital's parameters θ_k from the global parameters θ_g in a final local fine-tuning step. Cheng et al. (2023) and Yu et al. (2020) show that this simple approach is effective and requires no machinery beyond standard FedAvg. Subsequent work shows that fine-tuning only part of the model can further help by limiting overfitting, either to the hospital's distribution or to its finite training set. Existing personalized methods differ in how far each hospital's model may depart from the shared one: through a proximal penalty (Li et al., 2020; 2021), a split into shared and private parameters (Collins et al., 2021; Liang et al., 2020), or a budget of local fine-tuning epochs (Yu et al., 2020). In all of them, however, the shared model is still trained with the FedAvg objective. It therefore learns a weighted average of the hospital distributions rather than structure genuinely common to them. FEDLIBRA instead guides the shared model toward a representation that transfers across hospitals.

Personalized vs. Generalized performance From here, there is a tension between two ways of measuring performance: following standard FL practice, the *personalized* score where each hospital is evaluated against its own test set, and the *generalized* score which evaluates hospitals against the collective of all hospitals in the system. Assuming each hospital has a test split $T_k = \{\boldsymbol{x}_k^{(1)}\}_{i=1}^{n_{k,test}}$, we write $T_{\text{tot}} = \bigcup_k T_k$ for the system's global test set.

The personalized version of this is the average performance of each hospital across their personal datasets. That is, $L_{\text{per}}(\theta_{1:K}) := \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|T_k|}\sum_{\boldsymbol{x}\in T_k} \ell(\theta_k; \boldsymbol{x})$. The generalized version of this is the average performance of each hospital evaluated on the global test set, $L_{\text{gen}}(\theta_{1:K}) := \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|T_{\text{tot}}|}\sum_{\boldsymbol{x}\in T_{\text{tot}}} \ell(\theta_k; \boldsymbol{x})$. For those methods that produce only one global model θ_g at the end (namely centralized, FedAvg), replace every θ_k with θ_g.

4 FEDLIBRA

Here we introduce the objectives and algorithm for FEDLIBRA, and demonstrate how it addresses the challenges of heterogeneity across hospitals and multimodality within them. We propose FEDLIBRA Figure 3, which learns a shared generator that captures what hospitals hold in common, and adapts that cheaply to each individual hospital. Concretely, we (1) train the shared parameters θ_g with an invariance penalty and a best-of-K risk and (2) train a personalized adapter v_k on top of the globally broadcast θ_g.

Heterogeneity. Most perFL methods treat the global model as a shared starting point: FedAvg learns an average of the hospitals, and personalization decides how far each hospital may move away from it. FEDLIBRA instead changes what the global model learns. Some differences between hospitals come from how a site records rather than whom it treats. A teaching hospital, for instance, may document more secondary diagnoses per stay than a community hospital seeing similar patients, which we consider a hospital-dependent, or spurious, correlation. Instead of having an average model which absorbs these patterns, we want instead a shared generator that captures features that hold wherever a record was taken: for example, that heart failure co-occurs with hypertension and atrial fibrillation. These relationships should not depend on the hospital that recorded them. IRM (Peters et al., 2016; Arjovsky et al., 2020) targets exactly this: it treats each data source as an environment and favors what is equally useful in every one.

Concretely, we use the IRMv1 penalty. Let $L_k(\theta; w)$ be hospital k's risk when every code logit of the generator is multiplied by a scalar w, so that $L_k(\theta) = L_k(\theta; 1)$. Each hospital trains on $L_k^{\text{IRM}}(\theta) = L_k(\theta) + \rho \left\| \nabla_w L_k(\theta; w)\big|_{w=1} \right\|^2$, where ρ is the invariance penalty weight. Summed over hospitals, it asks for a single feature backbone that suits every hospital at once rather than the pooled average.

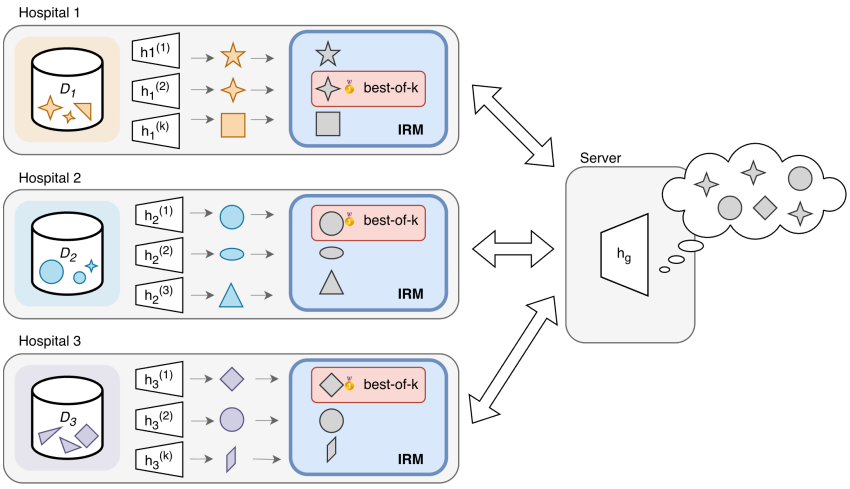


Figure 3: **Overview of FEDLIBRA.** Each hospital k trains a local generator h_k on its private data D_k. *Best-of-K*: for each real record, h_k draws K candidates, and only the candidate closest to the record (red, medal) receives a gradient, so the generator can commit to one patient phenotype instead of averaging many. *IRM* (blue): an invariance penalty downweights site-specific features (color) and keeps what holds across hospitals (shape). Only model updates reach the server, which aggregates them into a shared generator h_g.

The penalty is computed from each hospital's own data, so it changes only the local loss: FEDLIBRA needs no extra communication than FedAvg and no hospital's data or gradients leave the hospital, beyond the usual update. Following Arjovsky et al. (2020), we hold ρ at 1 for the first phase of training before raising it, to prevent training the trivial invariant solution of fitting nothing.

Multimodality. Hospital patients are not typically drawn from a single cluster. For example, an ICU treats respiratory, cardiac and septic patients, and their diagnosis codes overlap very little. Under maximum likelihood, a generator with limited capacity tends to spread probability mass across these groups, and its samples can fall between them rather than inside any one. In the federated setting the shared model has to cover the patient groups of all hospitals, which makes the problem larger. To address this, we train with a best-of-K objective from multiple-choice learning (Guzmán-rivera et al., 2012; Gladstone et al., 2026). We give HALO a latent input $z \sim \mathcal{N}(0, I_{16})$, linearly projected and added to the embedding at the position before the first visit, and drawn once per patient. For each training record, hospital k draws K independent latents $z_1, \ldots, z_K$ and trains only on the candidate that explains the record best, via $L_k^K(\theta) = \frac{1}{n_k} \sum_{\boldsymbol{x} \in D_k} \min_{j \in \{1,\ldots,K\}} \ell(\theta; \boldsymbol{x}, z_j)$, where ℓ is the per-record loss of Section 3. Because only the winning latent receives a gradient, different regions of latent space are free to specialize to different phenotypes instead of averaging over them. FEDLIBRA combines best-of-K with IRMv1, taking the minimum over candidates before computing the penalty.

Local adaptation. Fine-tuning the federated model on each hospital's own data is a strong personalization baseline (Cheng et al., 2023; Yu et al., 2020), but fine-tuning every parameter lets a small hospital overwrite the shared model and lose what federation contributed. FEDLIBRA instead keeps the shared parameters θ_g frozen and trains a low-rank adapter (Hu et al., 2022) on the output head only. We use rank 8, with roughly 38K trainable parameters in the adapter, representing , 0.6% of the model. Adaptation uses the same per-hospital objective as federated training, and takes 20 of the 100 epochs in the budget. Unlike methods which federate the adapters themselves (Xie et al., 2024), we federate the full model and use the adapter only for this final, local step.

5 RESULTS

Experimental Setup We simulate a federation of hospitals using eICU-CRD (Pollard et al., 2018), a multi-center critical-care database with records from 208 US hospitals. We select eight hospitals as clients, ranging from about 200 to over 2,600 patients, so that the federation is non-IID in both client size and patient population. Each client trains a HALO (Theodorou et al., 2023) generator on its own patients. We bound the frontier with local-only training and a centralized model trained on pooled data, and compare FEDLIBRA against FedAvg (McMahan et al., 2017), FedAvg with local fine-tuning (FedAvg+FT), and five personalized federated baselines: Ditto (Li et al., 2021), PerAda (Xie et al., 2024), FedProx (Li et al., 2020), FedRep (Collins et al., 2021), and LG-FedAvg (Liang et al., 2020). We compare methods under matched compute budget; that is, each gets 100 local epochs, interspersed with federated averaging rounds, usually every 2 epochs. Per-hospital statistics are in Appendix Table 5 and construction details in Appendix Section A.

Fidelity To measure fidelity, we consider two key prevalence metrics (Jiang et al., 2026; Theodorou et al., 2023). For each code, we compute its prevalence, the fraction of patients carrying it, in the real and synthetic data. We report the Pearson correlation r between the two prevalence vectors, which measures whether codes are ranked correctly, and R^2 against the identity line ($1 - \sum_c (p_c^{\text{syn}} - p_c^{\text{real}})^2 / \sum_c (p_c^{\text{real}} - \bar{p}^{\text{real}})^2$), which additionally requires the rates themselves to match and is negative when the synthetic rates are worse than a constant.

In Table 1, FEDLIBRA achieves Pearson correlations comparable to the strongest baselines on both axes, and is the only method that attains a high R^2 against the identity line on both axes. Because Pearson r is invariant to scale and offset, it measures only whether a method ranks code prevalences correctly, whereas R^2 against the identity also requires the absolute rates to match. FEDLIBRA therefore recovers both the relative frequency of codes and how often they actually occur. The federated baselines show the opposite pattern: FedProx, FedRep, and LG-FedAvg preserve the ranking of code frequencies ($r \geq 0.83$ on both axes) but produce codes at rates far from the true level, with R^2 well below zero on both axes. We hypothesize that these methods systematically miscalibrate code rates, generating codes too often or too rarely even when their relative ordering is correct.

Table 1: **Prevalence fidelity of FEDLIBRA and baselines for personalized federated learning.** Each method is scored against each hospital's own test set (pers.) and against the pooled test set (gen.). † If trained like every other row, centralized training fails to model code prevalence, and its R^2 falls far below zero. We upweight rare codes to have a better generalization reference, and this is the one present in visualization Figure 1.

	Pearson r ↑		R^2 vs identity ↑	
Approach	pers.	gen.	pers.	gen.
Reference bounds				
local-only	0.933	0.866	0.675	-0.469
centralized	0.578	0.686	-53.1	-44.3
centralized, rare-upweighted†	0.839	0.981	0.134	0.858
FedAvg	0.845	0.969	-0.262	0.552
Personalized FL baselines				
Ditto Li et al. (2021)	0.844	0.968	0.268	0.852
PerAda Xie et al. (2024)	0.841	0.968	0.284	0.851
FedProx Li et al. (2020)	0.829	0.964	-4.709	-3.021
FedRep Collins et al. (2021)	0.879	0.840	-1.704	-5.048
LG-FedAvg Liang et al. (2020)	0.854	0.898	-5.129	-17.7
FEDLIBRA	0.888	0.944	0.716	0.704

Utility. A central motivation for synthetic patient data is training downstream clinical models without access to the real records (Pezoulas et al., 2024). We evaluate utility with the train-on-synthetic, test-on-real (TSTR) protocol (Theodorou et al., 2023; Jiang et al., 2026; Bauer et al., 2024). For each method, we generate 8,000 synthetic patient trajectories, train a classifier on them, and evaluate it on the real test sets. We select 30 codes uniformly at random, and for each, remove it from

Table 2: **Downstream utility of synthetic data from each generator.** For each method and each hospital, a classifier is trained on that client's synthetic data and evaluated on the hospital's real held-out test set; metrics are averaged across hospitals. For reference, Local-Only trains the classifier on each hospital's own real training data, and All-Data* trains it on the real training data of all hospitals pooled, which is unavailable in the federated setting. FEDLIBRA (**bold**) outperforms most baselines and performs comparably to FedRep and LG-FedAvg (*italics*), despite their much lower prevalence fidelity (Table 1).

Approach	AUROC ↑	AP ↑	F1 ↑
Reference bounds			
Chance	0.500	0.029	0.056
Local-Only	0.420	0.031	0.068
All-Data*	0.785	0.153	0.232
FedAvg	0.617	0.061	0.114
Personalized FL baselines			
Ditto ($\lambda = 0.1$)	0.619	0.062	0.117
PerAda	0.610	0.062	0.119
FedProx	0.605	0.063	0.117
FedRep	*0.709*	*0.105*	*0.180*
LG-FedAvg	*0.700*	*0.100*	*0.168*
FEDLIBRA	**0.702**	**0.097**	**0.167**

every patient's trajectory and ask the classifier to predict whether the patient originally had it. We compute AUROC, average precision, and F1 on each hospital's test set, then average each metric across hospitals and over the 30 codes.

In Table 2, FEDLIBRA outperforms FedAvg, FedProx, Ditto, and PerAda on downstream AUROC (0.702 versus 0.605–0.619) and performs comparably to FedRep and LG-FedAvg (0.709 and 0.700), despite the severe prevalence miscalibration of these two methods noted above. This suggests that downstream utility does not require accurate absolute code rates: a classifier trained on synthetic data can learn which codes predict one another even when those codes appear far too often or too rarely. This points to a limitation of TSTR evaluation, which cannot distinguish a generator that matches real code rates from one that does not.

Table 3: **Privacy metrics across approaches.** An NNAAR near 0 indicates no leakage; a positive value indicates that synthetic records lie closer to the training data than to held-out data. For membership inference attack (MIA) accuracy, values near 0.5 are ideal. ‡FedProx generates about 24 codes per synthetic patient (real median: 2), so both adversarial accuracies saturate and its NNAAR of 0 reflects unrealistic trajectories rather than privacy. No method, federated or centralized, shows meaningful leakage.

Approach	NNAAR $\to 0$	MIA acc. $\to 0.5$
Reference bounds		
local-only	-0.0090 ± 0.0112	0.480
centralized	0.0200 ± 0.0282	0.513
FedAvg	0.0155 ± 0.0191	0.519
Personalized FL baselines		
fine-tune (full, 6 M)	-0.0070 ± 0.0306	0.501
fine-tune (LoRA head, 38 K)	0.0255 ± 0.0237	0.506
Ditto	0.0145 ± 0.0195	0.475
PerAda	0.0150 ± 0.0305	0.482
FedProx ‡	0.0000 ± 0.0000	0.510
FedRep	0.0060 ± 0.0085	0.512
LG-FedAvg	0.0040 ± 0.0068	0.500
FEDLIBRA	-0.0220 ± 0.0497	0.499

Privacy. We measure whether synthetic records leak their training data with two nearest-neighbor tests. NNAAR (Appendix Equation (5)) (Yale et al., 2020) compares how close synthetic records lie to the generator's training patients against how close they lie to held-out patients; a positive value indicates memorization, and values below 0.03 are conventionally considered safe. $\pm$ is the standard deviation of NNAAR over 5 independent draws of 200 training, held-out and synthetic patients each. Membership inference attack (MIA) accuracy (Shokri et al., 2017) is the success rate of a membership-inference attack that labels a patient a training member if their nearest synthetic record is closer than the median; 0.5 is chance. All approaches pass both tests. FEDLIBRA's NNAAR (-0.022 ± 0.050) is indistinguishable from zero, and a negative value means synthetic records are no closer to training patients than to held-out ones. MIA accuracy is within 0.02 of chance for every approach, including FEDLIBRA (0.499). Both tests use a simple distance-based attacker, so they show the absence of this kind of leakage rather than a formal privacy guarantee.

Table 4: Ablation. Personalized (pers.) and generalized (gen.) fidelity across approaches. Best value per column in **bold**.

	Pearson r ↑		R^2 vs identity ↑	
Approach	pers.	gen.	pers.	gen.
Reference approaches				
local-only	**0.933**	0.866	0.675	-0.469
centralized, rare-upweighted	0.839	**0.981**	0.134	0.858
FedAvg	0.845	0.969	-0.262	0.552
Changing the shared model (no fine-tuning)				
FedAvg + IRM	0.828	0.974	0.396	0.932
FedAvg + best-of-K (latent)	0.845	0.971	-1.693	-0.430
FedAvg + IRM + best-of-K (latent)	0.843	0.973	0.474	**0.940**
Fine-tuning				
FedAvg + full fine-tune	0.914	0.894	-0.486	-1.521
FedAvg + LoRA fine-tune	0.896	0.933	0.426	0.097
FEDLIBRA (IRM + best-of-K + LoRA)	0.888	0.944	**0.716**	0.704

Ablations. Table 4 separates the contribution of each component. Changing the shared model alone improves both axes over FedAvg: the IRM penalty raises R^2 from -0.262 to 0.396 on each hospital's own data and from 0.552 to 0.932 on the pooled data, and best-of-K has a similar effect. Their combination gives the strongest shared model (0.474 / 0.940). Fine-tuning behaves differently depending on its starting point. Applied to FedAvg, full fine-tuning overfits and loses both axes, and LoRA fine-tuning alone improves personalization at a large cost to generalization (0.426 / 0.097). Applied to the IRM and best-of-K model, the same LoRA fine-tuning gives the best personalization in the table (0.716) while keeping generalization at 0.704. FEDLIBRA therefore improves on FedAvg with fine-tuning on both axes, because it fine-tunes a better shared model.

6 CONCLUSION

Federated EHR generators face a trade-off: models fit to one hospital fail to transfer, and a model fit to all hospitals fits none of them well. Existing personalized federated learning methods move along this trade-off rather than past it. FEDLIBRA changes what the shared model learns. An invariance penalty keeps site-specific recording habits out of the shared model, and best-of-K training lets it represent distinct patient populations that no single hospital observes together. A small local adapter then fits each hospital. Across eight eICU hospitals, FEDLIBRA is the only method that accurately matches code prevalence on both each hospital's own data and the pooled data, performs on par with the best baselines in downstream utility, and shows privacy risk no higher than any baseline.