CHiP-FL: A Federated Learning Framework for Equitable ICU Risk Modeling

Rhea Zhou
Cary Academy rheazhou2026@gmail.com

Abstract In healthcare, machine learning (ML) model development is constrained by strict privacy regulations that limit cross-hospital data sharing. Federated Learning (FL) allows institutions to train shared models by updating local parameters and aggregating them through a central coordinator. However, hospitals produce non-IID training environments that destabilize optimization and introduce performance disparities between large and small hospitals. To address these issues, we propose **CHiP-FL** (Client Clustering, Hierarchical Aggregation, and Personalization Federated Learning), a federated optimization algorithm balancing predictive performance with cross-site equity. Each hospital is represented by a data signature vector capturing dataset size, label prevalence, missingness structure, and feature statistics, which are then clustered via k-means. Training proceeds through a multi-level regularized objective that jointly optimizes local institutional loss and divergence between hospital, cluster, and global models. We evaluate CHiP using the eICU Collaborative Research Database and compare it against established federated baselines. While optimization-centric methods such as FedProx achieve the highest global AUROC (0.933), this gain is accompanied by substantial inter-site inequality, including the largest standard deviation (0.062) and high size bias (-0.032). In contrast, CHiP achieves competitive global performance (0.903 global; 0.897 personalized) while reducing institutional performance bias by over 90% (-0.002), nearly eliminating the relationship between hospital size and predictive benefit. CHiP also maintains a low Gini (0.028 global; 0.027 personalized) and moderate worst-decile AUROC (0.839 global; 0.840 personalized), demonstrating balanced protection across sites without sacrificing generalization. Multi-objective frontier analysis reveals that no single FL method simultaneously optimizes performance, robustness, and structural equity, illustrating the fundamental fairness-efficiency tension in healthcare FL. These findings position equity as a core optimization objective in federated clinical AI.

Keywords: Federated Learning, Healthcare AI, Fairness, ICU Mortality Prediction, Non-IID Data, Hierarchical Aggregation, Personalization

1. Introduction

In 1996, HIPAA was enacted to grant individuals rights over their data and to balance privacy protections with the need for information flow to improve quality of care [Centers for Disease Control and Prevention, 2024]. These constraints slow the deployment of AI in healthcare. ML models trained within a single institution are limited to that institution's patient population, preventing them from learning generalizable patterns across diverse clinical environments.

Federated Learning

Federated Learning (FL) is a distributed ML framework gaining traction in the healthcare space. Introduced by Google in 2016, it enables multi-party collaboration while preserving data privacy [Teo et al., 2024]. A centralized server broadcasts initial weights of global model parameters to selected

participating sites, who each train a local model using their local data. The updated model parameters are then sent back to the centralized server, which aggregates them to update the global model. The updated model is broadcast to a new set of sites for another round of local training, and the cycle repeats [Liu, 2022]. FL allows various sites to train ML models in a way where raw data never leaves the local device [Fauzi et al., 2022]. The number of FL papers in healthcare increased from 1 in 2018 to over 250 by 2023, with over half of real-life FL applications involving international or regional collaboration [Teo et al., 2024].

Generalizable Insights about Machine Learning in the Context of Healthcare

This paper yields several insights relevant to the broader ML-in-health community:

- **Equity is a multi-dimensional construct.** Healthcare fairness cannot be reduced to a single scalar metric. Tail robustness (worst-decile AUROC), distributional spread (Gini coefficient), and structural bias (size-dependent performance slope) are orthogonal objectives; optimizing one does not imply progress on the others.
- **Federated algorithms occupy distinct regions of a performance-equity phase space.** Standard global averaging (FedAvg) minimizes dispersion but preserves size bias; proximal regularization (FedProx) maximizes accuracy but amplifies inequality; hierarchical clustering improves tail performance but sacrifices global optimality. No single convex objective dominates all three axes simultaneously.
- **Hierarchical regularization and local personalization are complementary, not redundant.** CHiP's cluster-level constraints suppress scale-driven domination during training, while post-hoc personalization recovers institution-specific signal without reintroducing institutional bias–a combination unavailable in single-level FL frameworks.
- **Fairness must be specified before optimization.** Defining "fair" FL requires normative choices about whose outcomes matter most. Healthcare FL designers should articulate distributional objectives explicitly rather than relying on global accuracy as a proxy.

2. Related Work

The FedAvg algorithm enables decentralized model training by iteratively aggregating client updates without sharing raw data [McMahan et al., 2017]. Clients perform several local training steps before sending model updates to a central server, which averages the parameters to update the global model. This framework allows collaborative training while preserving data privacy, making it particularly attractive for healthcare. However, FedAvg assumes relatively homogeneous client data and struggles with non-IID environments.

To address this, several extensions have been proposed. FedProx introduces a proximal term that constrains local updates to remain close to the global model parameters, improving stability when client objectives diverge [Li et al., 2020]. Another line of work explores clustered FL, which groups clients with similar data distributions and trains cluster-specific models. The IFCA framework alternates between assigning clients to clusters and training cluster-level models, enabling improved personalization for heterogeneous client populations [Ghosh et al., 2020]. However, clustered approaches typically require careful initialization and can struggle when client distributions evolve during training.

Other methods attempt to correct for client drift during training. SCAFFOLD introduces control variates that reduce variance in client updates and mitigate drift caused by heterogeneous local training

objectives [Karimireddy et al., 2021]. More recent work focuses on representation-level alignment: the MOON framework incorporates contrastive learning to align client model representations with the global model while discouraging divergence from previous local models [Li et al., 2021]. This method improves generalization but primarily addresses representation drift rather than structural heterogeneity between clients.

The challenge of balancing global performance with fairness across clients of varying sizes remains open. Existing approaches address one aspect of data heterogeneity but do not jointly address macro-level institutional differences, meso-level group heterogeneity, and micro-level client personalization. Furthermore, standard FL aggregation tends to amplify bias toward larger institutions whose local updates have greater influence on the global model, preventing smaller hospitals from effectively adopting these models.

3. Methods: The CHiP-FL Algorithm

We propose the CHiP-FL (Client Clustering, Hierarchical Aggregation, and Personalization Federated Learning) algorithm to address non-IID data, negative transfer from global averaging, and the marginalization of minority hospitals. It uses clustering, proximal terms, and personalization to handle macro, meso, and micro heterogeneity, respectively. Full pseudocode can be found in Appendix A.

3.1 Notation

Table 1: Notation for CHiP-FL Variables

Variable	**Notation**
Hospitals/clients	$i \in \{1,\ldots,n\}$
Cluster index	$k \in \{1,\ldots,K\}$
Cluster assignment	$C(i) = k$
Global model	θ_g
Cluster model	θ_k
Client model	θ_i
Local loss	$\mathscr{L}_i(\theta)$
Hyperparameters	λ, μ, α
Learning rate	η
Local steps	E

3.2 Data Signature Vectors

We first compute a signature vector for each client containing information about sample size, label prevalence, missingness rates, and means/standard deviations of key vitals and labs. The hospitals are then clustered using k-means, and these clusters remain fixed throughout the experiment. Fixed cluster assignments prevent instability from frequent reassignment during training, allowing cluster models to converge to stable representations of institution-level data distributions.

3.3 Formal Objective Function

Each hospital (client i) optimizes a dual-regularized objective balancing three competing goals: accurate local learning, cluster-level consistency, and global stability:

$$F_i(\theta) = \mathscr{L}_i(\theta) + \lambda \|\theta - \theta^t_{C(i)}\|^2 + \mu \|\theta - \theta^t_g\|^2 \tag{1}$$

The first term is the log loss on the local dataset. The second term introduces cluster regularization, promoting similarity among hospitals with comparable data distributions. The third term applies global regularization, preventing excessive divergence across clusters. The hyperparameters λ and μ control cluster cohesion and global alignment, respectively. We set $\lambda = 0.5$ and $\mu = 0.05$, placing stronger emphasis on cluster-level regularization, reflecting the goal of allowing hospitals to learn primarily from similar peers.

3.4 Training Procedure

At each communication round t, clients receive both the current cluster model and the global model, then perform local updates:

$$\theta \leftarrow \theta - \eta \nabla F_i(\theta) \tag{2}$$

where

$$\nabla F_i(\theta) = \nabla \mathscr{L}_i(\theta) + 2\lambda(\theta - \theta^t_{C(i)}) + 2\mu(\theta - \theta^t_g) \tag{3}$$

The proximal correction terms act as stabilizing forces that constrain model drift while allowing local specialization. After E local steps, each client communicates the update difference:

$$\Delta_i = \theta_i^{(t+1)} - \theta^t_k \tag{4}$$

Transmitting update differences rather than full parameters emphasizes incremental refinement and improves aggregation stability.

3.5 Hierarchical Aggregation

CHiP performs two-level aggregation to preserve heterogeneity while enabling global knowledge sharing. For each cluster k, client updates are combined using sample-size-weighted averaging over participating clients S_k:

$$\theta_k^{(t+1)} = \theta^t_k + \sum_{i \in S_k} \frac{n_i}{\sum_{j \in S_k} n_j} \Delta_i \tag{5}$$

Once all clusters are updated, the global model is constructed through hierarchical averaging:

$$\theta_g^{(t+1)} = \sum_{k=1}^{K} \frac{N_k}{N} \theta_k^{(t+1)} \tag{6}$$

where N_k denotes the total samples in cluster k and $N = \sum_k N_k$. To prevent excessive cluster drift, a stabilization step blends global and cluster representations:

$$\theta_k^{(t+1)} \leftarrow \alpha\, \theta_k^{(t+1)} + (1-\alpha)\, \theta_g^{(t+1)} \tag{7}$$

The blending coefficient $\alpha \in [0,1]$ controls the tradeoff between cluster specialization and global consistency.

3.6 Personalization

After federated training converges, CHiP performs an optional personalization stage. Each client i initializes from its cluster model $\theta^T_{C(i)}$ and fine-tunes using only its private dataset:

$$\theta \leftarrow \theta - \eta_p \nabla \mathscr{L}_i(\theta) \tag{8}$$

No regularization toward cluster or global models is applied, allowing full adaptation to local data. Personalization occurs entirely locally without additional communication, preserving privacy. By combining clustering, hierarchical aggregation, and personalization, CHiP addresses heterogeneity at macro (global), meso (cluster), and micro (client) levels.

4. Cohort

4.1 Dataset and Task

We use the eICU Collaborative Research Database [Johnson et al., 2018], a publicly available multi-center critical care dataset comprising 200,000+ de-identified ICU admissions from 208 hospital units across the US, collected between 2014 and 2015. The dataset spans 139,000+ patients across 335 ICU units, encompassing substantial variation in hospital size, patient acuity, and clinical practice. The prediction task is in-hospital mortality, formulated as binary classification over individual ICU stays. This task is clinically relevant to FL evaluation, as mortality risk profiles vary systematically across institutions due to differences in case mix, admission thresholds, and resource availability.

4.2 Data Extraction

Features are drawn from vital signs (heart rate, blood pressure, oxygen saturation, respiratory rate), lab results (sodium, potassium, creatinine, lactate), medications (antibiotics, sedatives, vasopressors), diagnoses (sepsis, cardiac arrest, pneumonia, trauma), and care plans (intubation, dialysis, ventilation, surgery). Raw data was accessed via Google BigQuery. A structured SQL pipeline was developed to extract a model-ready feature table from eICU's relational schema. The cohort was restricted to adult patients (age $\geq$ 18). Discharge status strings were excluded to prevent label leakage. Missing values were retained and encoded with explicit missingness indicator flags, preserving the clinical signal embedded in data absence patterns.

4.3 Feature Choices

The final feature table was constructed using a series of CTEs targeting distinct clinical domains: BMI derived from height and weight, first-day GCS scores, periodic and aperiodic vital sign aggregates, comorbidity flags, ventilation status, and vasopressor exposure. A global preprocessing pipeline was fitted on a stratified sample of 30,943 training observations. Continuous features were standardized using z-score normalization, and missing values were imputed using per-feature training medians. Features with complete missingness across all training instances were excluded. This pipeline was applied identically across all clients and FL methods to ensure fair comparison.

4.4 FL Setting and Implementation

The dataset was partitioned into 208 federated clients corresponding to individual hospital units, simulating a realistic cross-silo healthcare federation. All methods were implemented using Flower [Beutel et al., 2020] and trained with federated logistic regression. Logistic regression was selected for

its interpretability, stable convergence under federated gradient aggregation, and tractability for controlled multi-method comparison.

At each communication round, clients were sampled using a 10% participation rate (minimum 10 clients per round). Sampling used inverse-square-root weighting by local training set size, oversampling smaller hospitals. Training proceeded for 20 rounds with mini-batch SGD ($\eta = 0.05$, batch size 512, 1 local epoch per round). CHiP hyperparameters were $\lambda = 0.5$, $\mu = 0.05$, and $\alpha = 0.9$. These values were manually set rather than chosen via formal hyperparameter search. After federated training, personalization was performed for 1 additional local epoch with $\eta_p = 0.03$. All results reflect a single run with random seed 42; multi-seed evaluation is left for future work.

5. Results on Real Data

5.1 Evaluation Approach

We compare CHiP against five baseline FL strategies: FedAvg, FedProx, hierarchical aggregation, clustered training without global aggregation, and post-hoc personalization. Performance was evaluated using per-client AUROC at the final training round, weighted by site validation size. Cross-site equity was assessed using worst-decile AUROC, Gini coefficient, standard deviation, and size bias (slope of AUROC vs. $\log(n)$, where n is ICU stays per institution).

5.2 Comparison of FL Methods

Table 2: Performance and Equity Metrics Across Federated Learning Methods

Method	Mean AUROC	Worst-Decile AUROC	Gini	Std Dev	Size Bias
FedAvg	0.9106	0.8507	0.0228	0.0382	−0.0127
FedProx (μ=0.01)	0.9333	0.8318	0.0421	0.0623	−0.0322
Personalization (μ=0.01)	0.9122	0.8466	0.0246	0.0428	−0.0103
Hierarchical (k=5)	0.8938	0.8536	0.0290	0.0479	+0.0339
Clustered (No global)	0.9126	0.8421	0.0527	0.0531	−0.0343
CHiP (Global)	0.9035	0.8385	0.0282	0.0450	−0.0023
CHiP (Personalized)	0.8971	0.8397	0.0268	0.0429	−0.0019

FedProx achieved the highest global predictive performance (AUROC = 0.933), but this came with substantial inter-site disparity: the lowest worst-decile AUROC (0.832) and highest standard deviation (0.062) indicate improvements were concentrated at larger hospitals. Hierarchical aggregation achieved the strongest worst-decile AUROC (0.854) despite the lowest global performance (0.894), and was the only method with a positive size-bias slope (+0.034), specifically benefiting smaller institutions at the cost of global accuracy. Clustered training without global aggregation produced strong global performance (0.913) but the greatest Gini (0.053), reflecting divergence across clusters.

CHiP was designed to balance predictive performance with cross-site equity. Although it does not achieve the highest global AUROC (0.903 global; 0.897 personalized), it demonstrates several important structural advantages. First, while FedProx exhibits a strong negative size bias slope of −0.032, CHiP nearly eliminates this relationship (−0.0023 global; −0.0019 personalized), indicating that large hospitals no longer dominate optimization and small hospitals are not marginalized. Second, CHiP successfully constrains divergence while allowing structured heterogeneity–in contrast with clustered training (Gini = 0.053) and FedProx (Std Dev = 0.062). Third, CHiP's worst-decile AUROC ($\approx$ 0.84) is competitive: while hierarchical aggregation achieves the highest worst-site

protection, it sacrifices global performance (0.894). CHiP places in a structurally efficient region of the performance-equity space, with higher global performance than hierarchical aggregation and near-zero size bias.

CHiP with personalization outperforms its global counterpart across dispersion and size-bias metrics while incurring only marginal loss in mean AUROC. This suggests that hierarchical structural regularization and local adaptation act as complementary mechanisms: global constraints prevent scale-driven domination while personalization recovers site-specific decision boundaries without reintroducing institutional bias.

5.3 Pareto Frontier Analysis

We evaluated FL methods across global performance (mean AUROC), distributional equity (Gini and worst-decile AUROC), and structural fairness (size-dependent bias). This multi-axis evaluation reveals tradeoffs invisible under conventional mean benchmarking.

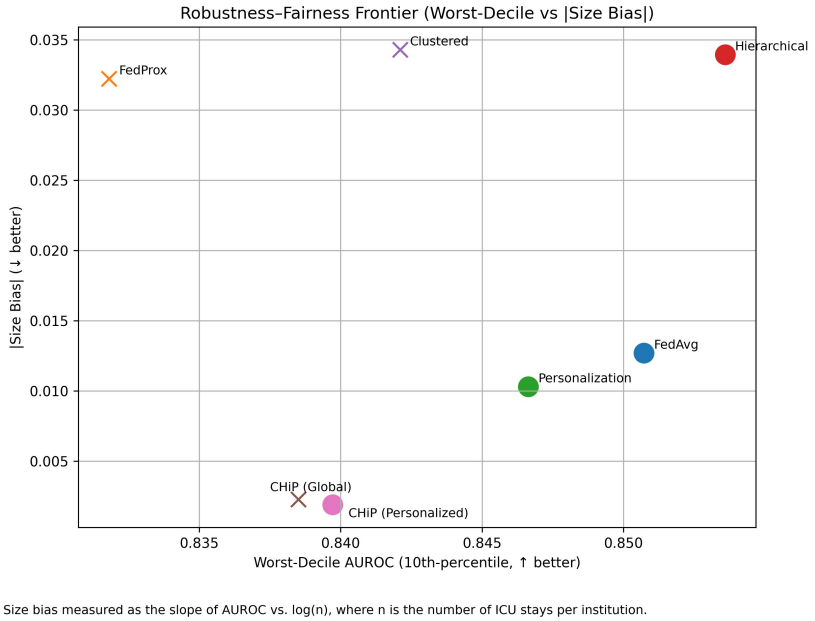

Figure 1: Robustness-Fairness Frontier (Worst-Decile AUROC vs. |Size Bias|).

Figure 1 shows the Pareto frontier between site-size performance bias and worst-decile AUROC. There is no clear correlation between the two axes, confirming that healthcare inequality is multi-dimensional. For example, hierarchical FL has the highest worst-decile AUROC and the second-highest |size bias|–a tradeoff between tail robustness and structural parity. CHiP becomes Pareto-optimal when structural bias is emphasized, underscoring that equity is a normative design choice rather than a purely technical optimization problem.

Figure 2 reveals distinct tradeoff regions in the 3D objective space. FedProx represents the performance extreme: large hospitals dominate due to lower gradient noise, stabler updates, and aggregate weights that scale with dataset size, producing high AUROC alongside high Gini and high size bias. FedAvg represents the dispersion-minimizing extreme: simple global averaging produces similar performance across hospitals (lowest Gini), but this apparent equality masks size bias since large hospitals align better with the mean global model. CHiP represents the fairness extreme, shifting optimization from minimizing global loss to minimizing structured multi-level

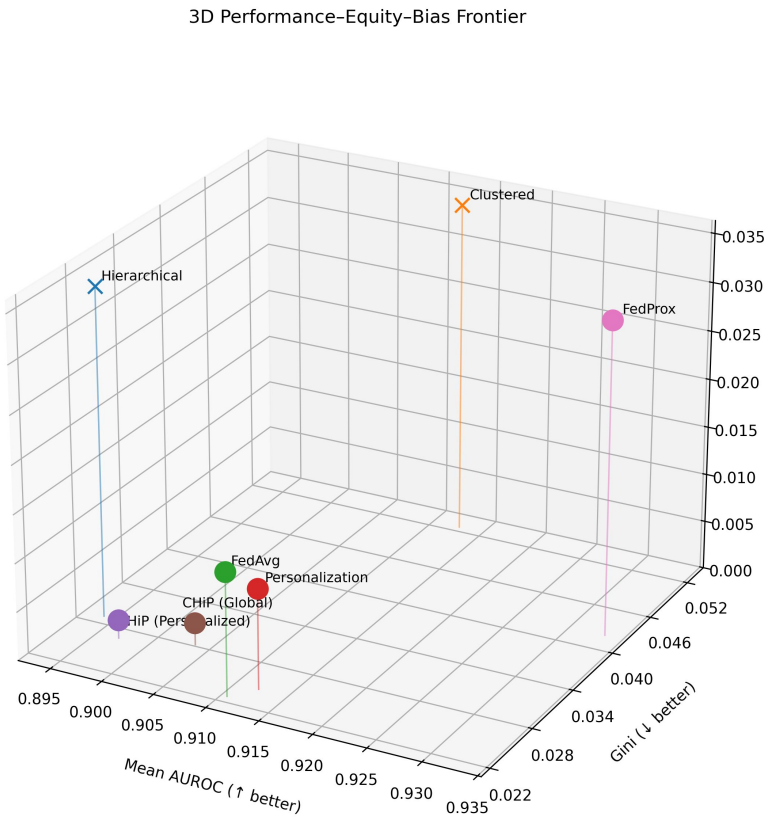

Figure 2: 3D Performance-Equity Bias Frontier (Mean AUROC vs. Gini vs. |Size Bias|).

loss–eliminating size bias because small hospitals learn within similar clusters rather than being forced toward a global optimum dominated by large sites.

Overall, both plots reveal performance, dispersion, and fairness as orthogonal optimization priorities. FL algorithms occupy different regions of a performance-equity phase space, shaped by competing forces: statistical efficiency (large data wins), distributional heterogeneity, and institutional fairness. No single convex objective optimizes all simultaneously. CHiP prioritizes equity structure at the cost of global optimality, representing a structural regularization framework rather than a pure performance optimizer.

6. Discussion

6.1 Real-World Implications

The results suggest that FL in healthcare should not be evaluated solely through global predictive accuracy, but through its distributional consequences across participating institutions. When federated optimization favors large medical centers, smaller community hospitals may receive systematically weaker predictive tools–reinforcing existing inequities in healthcare delivery despite the intention to democratize access to data-driven medicine.

CHiP reframes FL as a structural coordination problem rather than a purely statistical aggregation task. By introducing hierarchical regularization across global, cluster, and institutional levels, the

framework enables knowledge sharing while preserving clinically meaningful variation between hospitals. This structure reflects true healthcare ecosystems, where institutions operate within regional, demographic, and infrastructural contexts. Healthcare AI systems may therefore evolve toward multi-objective training paradigms that balance accuracy, robustness, and institutional fairness–improving trust and increasing adoption among hospitals historically excluded from large-scale ML initiatives.

6.2 Limitations

Several limitations should be considered. First, experiments used federated logistic regression; further work is required to evaluate CHiP's scalability under deep neural architectures and non-convex optimization. Second, clustering used static data signature vectors computed prior to training; adaptive clustering mechanisms may better capture temporal heterogeneity in real-world federations. Third, equity is evaluated through statistical proxies (size bias, dispersion) that do not directly measure downstream clinical outcomes or patient-level fairness. Fourth, communication and system-level constraints were simulated rather than deployed in a live federated environment, which introduces additional challenges such as asynchronous participation and heterogeneous computational resources. Fifth, all results reflect a single run with fixed random seed 42; multi-seed evaluation is left for future work. Finally, the definition of fairness remains normative–different healthcare stakeholders may prioritize tail robustness, variance minimization, or system efficiency differently. CHiP represents one point within a broader design space.

7. Conclusion

This work introduces CHiP-FL, a hierarchical FL framework designed to address institutional heterogeneity and structural inequity in multi-hospital clinical ML. By integrating client clustering, hierarchical aggregation, and post-training personalization within a unified optimization objective, CHiP balances predictive performance with cross-site equity. Evaluation on the eICU Collaborative Research Database demonstrates that conventional FL approaches occupy distinct regions of a performance-equity tradeoff space: methods that maximize global accuracy tend to amplify institutional disparities, and vice versa. CHiP achieves competitive accuracy while reducing site-size performance bias by over 90%, nearly eliminating the relationship between institutional scale and predictive benefit.

These findings highlight a broader shift in FL design: healthcare AI systems must optimize for both statistical efficiency and equitable participation across heterogeneous institutions. Hierarchical FL modeling provides a principled mechanism for achieving this balance. Future work will explore adaptive clustering, deep-model extensions, and real-world clinical deployment scenarios to further validate hierarchical federated optimization as a foundation for equitable medical AI.

References

Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. Flower: A friendly federated learning research framework, 2020. arXiv preprint arXiv:2007.14390.

Centers for Disease Control and Prevention. Health insurance portability and accountability act of 1996 (HIPAA). U.S. Department of Health and Human Services, 2024.

Retrieved March 2, 2026 from https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html.

Muhammad Ali Fauzi, Bian Yang, and Bernd Blobel. Comparative analysis between individual, centralized, and federated learning for smartwatch based stress detection. *Journal of Personalized Medicine*, 12(10):1584, 2022. doi: 10.3390/jpm12101584.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning (IFCA). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018. doi: 10.1038/sdata.2018.178.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning (MOON). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks (FedProx). In *Proceedings of the 3rd MLSys Conference*, 2020.

Yi Liu. Federated learning for smart healthcare: Challenges, methods, and future directions, 2022. ResearchGate preprint, September 2022.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *PMLR*, pages 1273–1282, 2017.

Zhen Ling Teo, Liyuan Jin, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, Ting Fang Tan, Deborah Meixuan Lee, Kai Jie Chua, John Heng, Yong Liu, Rick Siow Mong Goh, and Daniel Shu Wei Ting. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*, 5(2):101419, 2024. doi: 10.1016/j.xcrm.2024.101419.

A. CHiP-FL Algorithm Pseudocode

Input: Clients $i = 1 \dots n$ with local datasets D_i, sizes n_i; number of clusters K; hyperparameters λ, μ, α; total rounds T; learning rate η; local steps E.

Output: Global model θ_g; cluster models $\{\theta_k\}$; (optional) personalized models $\{\theta_i^*\}$.

Pre-Clustering:

1. For each client i, compute signature vector s_i from D_i: $\big(\log(n_i)$, label prevalence, missingness, feature stats$\big)$.
2. Run k-means on $\{s_i\}$; assign each client to cluster $C(i) \in \{1 \dots K\}$.
3. For each cluster k: $\mathscr{C}_k \leftarrow \{i : C(i) = k\}$, $N_k \leftarrow \sum_{i \in \mathscr{C}_k} n_i$. Set $N \leftarrow \sum_k N_k$.

Initialization: Initialize global model θ_g^0. For each cluster k, set $\theta_k^0 \leftarrow \theta_g^0$.

Federated Training (for $t = 0$ to $T-1$):

1. *Client updates:* For each cluster k, select subset $S_k \subseteq \mathscr{C}_k$. For each $i \in S_k$ in parallel:
   (a) Receive θ_k^t, θ_g^t; initialize $\theta \leftarrow \theta_k^t$.
   (b) For $e = 1$ to E: compute $g \leftarrow \nabla \mathscr{L}_i(\theta) + 2\lambda(\theta - \theta_k^t) + 2\mu(\theta - \theta_g^t)$; update $\theta \leftarrow \theta - \eta g$.
   (c) Set $\theta_i^{(t+1)} \leftarrow \theta$ and $\Delta_i \leftarrow \theta_i^{(t+1)} - \theta_k^t$.
2. *Cluster aggregation:* For each cluster k: $\theta_k^{(t+1)} \leftarrow \theta_k^t + \sum_{i \in S_k} \dfrac{n_i}{\sum_{j \in S_k} n_j} \Delta_i$
3. *Global aggregation:* $\theta_g^{(t+1)} = \sum_{k=1}^{K} \dfrac{N_k}{N} \theta_k^{(t+1)}$
4. *Cluster stabilization:* For each k: $\theta_k^{(t+1)} \leftarrow \alpha\, \theta_k^{(t+1)} + (1-\alpha)\, \theta_g^{(t+1)}$

Personalization: For each client i: set $k \leftarrow C(i)$, initialize $\theta \leftarrow \theta_k^T$. For $p = 1$ to P: $\theta \leftarrow \theta - \eta_p \nabla \mathscr{L}_i(\theta)$. Set $\theta_i^* \leftarrow \theta$.

Return θ_g^T, $\{\theta_k^T\}$, $\{\theta_i^*\}$.