

Federated Learning and Privacy-Preserving ML

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

*Chapter 16 (enrichment reading) of "Healthcare Analytics and AI: Building Systems That Actually Work"
The essentials: what to know, why it matters at the bedside, and how models fail.*

One-sentence thesis. Federated learning moves the model to the data instead of the data to the model, but the algorithm is the easy part: governance, fairness, and adversarial robustness decide whether it actually works.

PART 1: The Silo Problem and FedAvg

Why Hospitals Can't Share Data

The obstacle to a great healthcare model is almost never the algorithm; it is that the data lives in vaults that law and economics forbid you to open. Data silos are a legal and economic problem, not a technical one. HIPAA blocks research use of protected health information without patient authorization or an IRB waiver; the EU's GDPR treats health data as a special category and restricts cross-border transfer; and hospitals treat their data as a competitive asset worth guarding.

ANALOGY

A paradox of locked libraries: the patients who would benefit most from a model trained on diverse, multi-institution data are exactly the patients whose privacy laws keep that data from ever being assembled in one room.

The cost of the silos is poor generalization. A model trained only at Massachusetts General sees a patient population nothing like a rural critical-access hospital in Mississippi, and neither model transfers to the other's patients. Even MIMIC, the most-used open clinical dataset, is one institution in one city, so the field's models "work well on MIMIC patients because that is all they have ever seen."

Federated Averaging (FedAvg)

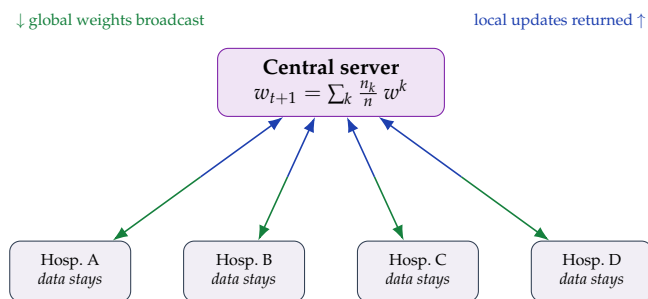
So how do five hospitals train one model without sharing a single patient record? FedAvg (McMahan

et al., 2017) answers this by running in communication rounds. The server broadcasts the current global weights; each client trains locally on its own data for several epochs; each client uploads only its updated weights; the server takes a weighted average, weighting each client by its number of examples; then it repeats. The data never moves, the way a shared study guide is written by many hospitals at once, each studying its own cases, revising its local copy, and sending back only the revised pages for a coordinator to merge.

In symbols, each round the server takes a size-weighted average of the clients' new weights:

$$w_{t+1} = \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$$

The insight that makes this practical is that clients take *multiple* local steps before communicating, cutting communication rounds 10- to 100-fold over slow hospital links. The payoff is that well-designed federated models routinely reach **95 to 98 percent** of the performance of a model trained on fully pooled data, a modest accuracy cost that buys enormous privacy protection. **Watch out.** FedAvg quietly assumes each client's data is drawn from the same distribution (IID) and that every client is honest. In healthcare both assumptions fail, and the next two parts are about what happens when they do.



Federated averaging. Each round the server broadcasts the global weights (green); each hospital trains locally and uploads only its updated weights (blue); the server takes a size-weighted average. The patient data never leaves the hospital.

Why Hospital Data Is Never IID

Take FedAvg's tidy assumption that every hospital's data looks alike and it collapses the moment you visit two hospitals. Institutions differ in disease prevalence, demographics, coding conventions, and even EHR vendor, so averaging their model updates produces a global model that fits none of them well,

the way averaging three regional accents produces a voice that sounds natural to nobody. This is the non-IID problem, the single largest technical challenge in healthcare federated learning. Concretely, when FedAvg blends the updates of three very different hospitals, a sepsis model tuned to the majority's physiology and documentation can quietly misfire on the minority's patients.

Hospital	Setting	Patients	Coding	EHR
A	Urban academic	Young, 40% Black, 25% Hispanic	Aggressive (CDI)	Epic
B	Suburban community	Older (median 67), mostly White	Conservative	Epic
C	Rural critical access (25 beds)	Low volume, high-acuity	Variable	Meditech

Fixing Drift: FedProx, SCAFFOLD, Personalization

The cure for drift is to stop each hospital from dragging the shared model toward its own corner of the data. Three families of fixes do this. **FedProx** adds a proximal term that penalizes any local model for straying too far from the global one, a leash that stops any single hospital from yanking the shared model toward its own distribution. **SCAFFOLD** uses control variates to correct the client drift that accumulates over multiple local steps, like a correction signal

handed to each driver in a convoy crossing a foggy mountain road so everyone stays pointed the same way. **Personalized FL** trains a global model, then lets each hospital fine-tune it locally before deployment.

The tradeoffs are real. In highly heterogeneous healthcare data FedProx has shown a **22 percent** absolute accuracy gain over FedAvg, and SCAFFOLD actually improves with more local computation, a gift for bandwidth-limited hospitals. Personalization fits each site best but produces divergent models, which complicates regulatory approval and accountability.

Algorithm	Non-IID mechanism	vs FedAvg	Tradeoff
FedAvg	None (baseline)	—	Diverges under heterogeneity
FedProx	Proximal regularization	~+22% in heterogeneous data	Slower convergence
SCAFFOLD	Control variates	Improves with more local steps	Higher per-round compute
Personalized FL	Local fine-tuning	Best institutional fit	Models diverge; governance

How Averaging Amplifies Bias

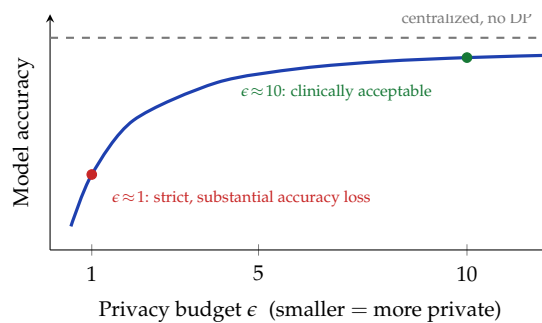
Here is the uncomfortable surprise: a model that never touches a single patient record can still be systematically unfair. FedAvg weights each client by its dataset size, so if four large hospitals serve mostly White patients and one smaller hospital serves mostly Black patients, the global model optimizes for the majority and marginalizes the minority through nothing more sinister than the arithmetic of averaging. It is majority rule by headcount, and the smaller hospital's hard-won lessons about its own patients are outvoted before the model ever reaches a bedside.

This mirrors the Optum algorithm failure (Chapter 4), where a cost-prediction model underallocated care to Black patients. FairFedMed (Harvard, 2025), the first federated-fairness benchmark for medical imaging, delivered a blunt finding: privacy preservation does not automatically produce fair models. FairLoRA addresses this by customizing per-demographic components while sharing the rest. **Watch out.** Deploy a federated model without measuring demographic performance across every participating site and you may build a system that works on average and fails

the patients who need it most.

Federated Is Not the Same as Private

A colleague will tell you federation makes the data private, so you can skip everything else. They are half right. Federation is strictly safer than centralizing data, since an attacker must breach every hospital rather than one repository, but it is not perfectly private: gradient-inversion attacks can reconstruct training samples from the shared parameters alone. The strongest deployments therefore add differential privacy, calibrated noise sprinkled on the shared gradients and governed by a privacy budget, epsilon. Federation spreads the valuables across many safes instead of one vault; differential privacy smudges the ledger so no single entry can be read back, and the harder you smudge, the safer and the less useful it becomes. A 2025 *npj Digital Medicine* study found a moderate budget (epsilon around 10) keeps clinically acceptable imaging performance, while a strict budget (epsilon around 1) often costs substantial accuracy, worst on small or heterogeneous datasets. Federated learning is not a substitute for differential privacy; the two are combined.



The privacy–utility tradeoff. A moderate budget ($\epsilon \approx 10$) keeps clinically acceptable accuracy; a strict budget ($\epsilon \approx 1$) often costs substantial accuracy, worst on small or heterogeneous datasets. Federation and differential privacy are combined, not substituted. The shape is schematic: the study reports the two budgets qualitatively, so the accuracy axis carries no scale.

Communication Cost and Compression

Even when only parameters travel, the bandwidth bill is brutal. A ResNet-50 update is about 100 MB, so 100 rounds with 20 clients exceeds 200 GB, and a 7-billion-parameter LLM costs roughly 14 GB per round per client, over the same slow hospital links. Three compression families help, and the idea behind all of them is to mail only the sentences that changed, in shorthand: gradient sparsification sends only the largest gradients, quantization drops precision from 32-bit down to 8-bit or even ternary, and knowledge distillation shares soft predictions on a public set instead of weights. MedHE (2025) pairs sparsification with homomorphic encryption for a **97.5 percent** cut in communication volume with negligible accuracy loss, and parameter-efficient LoRA transmits only 0.1 to 1 percent of a model’s size, which is what makes federated LLM fine-tuning on sensitive text, such as mental-health notes, practical at all.

Model Poisoning and Byzantine Tolerance

A distributed architecture buys an attack surface that centralized training never has: what if one client lies? A malicious or ransomware-compromised hospital can submit poisoned updates, either by flipping its training labels (data poisoning) or by scaling its uploaded gradient to dominate the average (model poi-

soning). The defense is to stop trusting the plain mean.

ANALOGY

One bad witness cannot move the coordinate-wise median the way it can drag the average, so the median simply ignores the liar.

That is the whole idea behind Byzantine-robust aggregation, and the tolerable number of malicious clients has a clean bound:

$$f \leq \frac{n-1}{3} \text{ malicious clients tolerable}$$

Practical defenses include median aggregation, trimmed mean (drop the highest and lowest per parameter), Krum, and consistency scoring against a validation set. In the drill, a single hospital flipping all of its labels measurably degrades naive FedAvg, and median aggregation recovers most of the loss.

Watch out. There is also a free-rider problem: a hospital can collect the global model each round while contributing noise or the unchanged model. Incentive mechanisms such as Shapley-value contribution scoring exist, but computing them exactly is exponentially expensive.



Why robust aggregation matters. One poisoned client (red) drags the coordinate-wise mean (amber) away from the honest cluster, but barely moves the median (green). Median, trimmed-mean, and Krum replace the vulnerable average in Byzantine-robust FedAvg.

Case Study: FeTS (Federated Tumor Segmentation)

The clearest proof that this works at scale is a real deployment. FeTS moved the model to the data across **32 institutions on six continents** (1,251 glioblastoma cases) to segment brain-tumor sub-regions from MRI. Published in *Nature Communications* (2025), the federated models generalized as well as centralized ones; an earlier 71-institution study improved tumor detection **17.2 percent** over single-site training. A hospital that sees mostly adult glioblastomas learns from the gradient contributions of one that sees mostly pedi-

atric medulloblastomas, without either ever seeing the other’s scans.

Imaging is the natural first fit because model parameters are orders of magnitude smaller than the volumetric scans they learn from, and the data-scarce hospitals benefit most: in a 21-hospital Brazilian mortality study, 16 of 21 sites improved, and the smallest gained the most. **Watch out.** FeTS spent years coordinating 32 institutional IRBs before a single gradient was exchanged. Federation lowers the technical barrier to multi-site research, not the governance barrier,

and deploying institutions must still validate a federated model on their own local data before clinical use.

Should You Build This?

Build it when several institutions hold data they legally cannot pool and the smaller participants see too few cases to train anything robust alone. That is the FeTS setup, and it is why the smallest sites in the Brazilian mortality study gained the most from joining. A federated model within a few points of centralized performance is a good trade when the alternative is not a centralized model but no model. **Think twice when** local distributions diverge far enough

that FedAvg drifts and you are reaching for FedProx or SCAFFOLD to hold the global model together, or when size-weighted averaging quietly marginalizes the one hospital serving a different patient population. Think twice again before selling federation as a privacy guarantee: gradient inversion is demonstrated, differential privacy is still required, and strict budgets cost real accuracy. And price the governance honestly. **Who bears the risk:** the deploying institution owns the wrong prediction while controlling almost none of the training that produced it, which is why local validation is non-negotiable; patients at underweighted sites absorb the fairness cost; and regulators are still writing the framework that would assign the rest.

Quick Self-Test

1. In one sentence, what does FedAvg move, and what does it deliberately keep in place?
2. Why does the IID assumption almost always fail across hospitals, and which algorithm adds a “leash” against the resulting drift?
3. Four large White-majority hospitals federate with one smaller hospital serving Black patients. Why can standard FedAvg make the global model less fair?
4. A colleague says federated learning makes patient data private, so differential privacy is unnecessary. Why are they wrong?
5. A single hospital flips all its training labels. What is this attack called, and what aggregation change defends against it?
6. Which participants benefit most from federation, and what evidence from FeTS or the Brazilian study supports that?

If You Remember One Thing

Federated learning is a technique; making it work is an organizational achievement. It is most valuable when local datasets are small, skewed, or unrepresentative, but its privacy is not absolute, its averaging can amplify bias, and a single poisoned client can corrupt it. The algorithm is the easy part. Governance, fairness auditing, and local validation are what turn it into a system a clinician can trust.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552