WISELEAF SCIENTIFIC VENTURES

SMART HEALTHCARE REVOLUTION: AI STRATEGIES FOR SUPERIOR GLOBAL HEALTH

CHAPTER 8- FEDERATED LEARNING ACROSS MULTI-HOSPITAL NETWORKS

Dr. Omkar Jayant Kulkarni

Assistant Professor, Department of Pharmacognosy, Dr. Kolpe Institute of Pharmacy, Kolpewadi, Maharashtra, India

Email: Komkar.dkipgmail.com@gmail.com

ABSTRACT

Clinical machine learning needs large and varied datasets, yet the records that would supply them sit in separate hospitals, each bound by privacy law, institutional policy and the ethical duty of confidentiality. Federated learning (FL) addresses this impasse by moving the model to the data: each participating site trains a shared model on its own records and returns only parameter updates, which a coordinator aggregates. This chapter sets out the method in technical detail for a health-sciences readership. It begins with the legal reasons why pooling is hard, covering the United States HIPAA Privacy Rule, the European Union's General Data Protection Regulation and India's Digital Personal Data Protection Act 2023 with the Rules notified in November 2025. It also covers the technical reasons why de-identification alone does not settle the matter. It then contrasts centralised, federated and swarm learning, derives the Federated Averaging (FedAvg) update, and explains why statistical heterogeneity across hospitals degrades it. Corrective algorithms such as FedProx, SCAFFOLD, adaptive server optimisers and personalised FL are discussed. Privacy is treated without overstatement: gradient inversion and membership inference attacks are described alongside differential privacy, secure aggregation, homomorphic encryption and trusted execution, with their costs. The evidence from landmark deployments is examined critically: brain-tumour segmentation across ten and then seventy-one sites, the EXAM COVID-19 oxygen-requirement model built across twenty institutes, swarm learning on transcriptomes and radiographs, and the MELLODDY consortium of ten pharmaceutical companies. Open-source frameworks are compared. The chapter closes with the governance, incentive and infrastructure questions that an Indian multi-hospital network would face, and with unresolved problems in fairness and evaluation. Federated learning lowers the barrier to collaboration; it does not remove the need for data curation, contracts or careful validation.

Keywords: federated learning; privacy-preserving machine learning; differential privacy; secure aggregation; non-IID data; Digital Personal Data Protection Act

1. Introduction

A deep-learning model is only as general as the data it has seen. That statement, banal in the abstract, has hard consequences in medicine. A chest radiograph classifier trained at one hospital learns that hospital's scanners, positioning habits, referral mix and labelling conventions, and it may learn them better than it learns the disease. Zech and colleagues showed exactly this with pneumonia detection: a convolutional network trained on radiographs from one health system generalised inconsistently when tested on images from other hospital systems, a sign that site-specific signals had entered its decision rule [1]. The cure for such shortcut learning is diverse training data drawn from many institutions. The obstacle is that those data cannot easily be moved.

The traditional answer has been the central repository: negotiate data-sharing agreements, de-identify records, transfer them to a single server and train there. This model built many public benchmarks, including the multimodal brain tumour segmentation (BraTS) collection [2]. It is slow, legally fraught and, for many datasets, simply unavailable. Hospitals are understandably reluctant to release records whose re-identification would breach patient trust, and some data, such as full-resolution imaging linked to longitudinal outcomes, are too large and too sensitive to copy freely [3,4].

Federated learning (FL) reverses the direction of travel. The model is sent to each data holder, trained locally, and only the resulting parameters, or changes to them, are returned for aggregation. The term and its now-standard algorithm, Federated Averaging, were introduced by McMahan and colleagues at Google in 2017, motivated by data held on mobile devices rather than in hospitals [5]. The healthcare community adopted the idea quickly because it maps onto the institutional reality of medicine: data are owned, governed and curated locally, while the clinical questions worth answering are shared. Rieke and co-authors argued in 2020 that FL could let digital health move beyond single-site proofs of concept toward models trained on populations large enough to matter [4].

This chapter treats FL as an engineering discipline with clinical consequences, not as a slogan. Section 2 explains why pooling is hard, legally and technically. Section 3 distinguishes centralised, federated and swarm architectures and the main FL variants. Section 4 derives FedAvg. Section 5 examines statistical heterogeneity, the central technical weakness of FL in hospital networks, and the algorithms proposed to correct it. Section 6 assesses privacy attacks and defences. Section 7 reviews landmark deployments in imaging, critical care and drug discovery, and Section 8 compares open-source frameworks. Sections 9 and 10 turn to governance and incentives, with particular attention to Indian hospital networks under the Digital Personal Data Protection Act 2023, and to unresolved problems in fairness and evaluation.

One clarification at the outset. Federated learning keeps raw data on site. It does not, by itself, make a model private, fair, accurate or legal. Each of those properties has to be engineered and demonstrated separately, and much of what follows is about the distance between the slogan "the data never leave the hospital" and the guarantees that a regulator, an ethics committee or a patient would actually want.

2. Why Medical Data Cannot Simply Be Pooled

2.1 The legal frame: HIPAA, GDPR and India's DPDP regime

In the United States, the Privacy Rule issued under the Health Insurance Portability and Accountability Act (HIPAA) governs how covered entities, chiefly providers, health plans and clearing-houses, may use and disclose protected health information. When the Rule took effect in 2003 it set national limits on disclosure for purposes other than treatment, payment and operations, and research uses generally require authorisation, a waiver from an institutional review board, or data that meet the Rule's de-identification standard [6]. HIPAA is a sectoral law: it binds particular entities rather than all holders of health data.

The European Union's General Data Protection Regulation (GDPR), applicable since May 2018, takes the opposite approach. It applies to any processing of personal data of people in the Union and classes data concerning health as a special category, processing of which is prohibited unless one of a closed list of conditions applies. Pseudonymised data remain personal data under the Regulation; only data rendered genuinely anonymous fall outside it [7]. Kaissis and colleagues note that both regimes impose strict rules on storing and exchanging identifiable information, and that GDPR also creates expectations of transparency in automated decision-making, which sits awkwardly with opaque models [3].

India's framework is newer. Parliament enacted the Digital Personal Data Protection (DPDP) Act on 11 August 2023. The Ministry of Electronics and Information Technology notified the implementing DPDP Rules through gazette notifications dated 13 November 2025 and announced them on 14 November, with an eighteen-month window for phased compliance [8,9]. The Rules create a four-member, digital-first Data Protection Board, with appeals to the Telecom Disputes Settlement and Appellate Tribunal, and require consent managers, who offer individuals a single platform to give, review and withdraw consent, to be companies based in India [8]. Professional analyses of the Rules describe a staged timetable. Provisions on the Board took effect immediately. Consent-manager provisions commence after one year, in November 2026, and most obligations on data fiduciaries, including the breach-notification requirements, commence after eighteen months, in May 2027 [9,10]. As of November 2026, then, the consent-manager phase is about to begin and hospitals have roughly six months before their core obligations bind.

Several features of the Rules bear directly on multi-hospital machine learning. A data fiduciary that suffers a personal data breach must inform affected individuals without delay, in plain language, and must file a detailed report with the Board within seventy-two hours [8,10]. Processing of children's data needs verifiable consent from a parent or guardian, although processing tied to essential services such as healthcare is exempted from that requirement [8]. Entities designated as Significant Data Fiduciaries must appoint a data protection officer, carry out annual data protection impact assessments and audits, and exercise due diligence over the algorithmic software they deploy [10]. Large hospital chains and national programmes could plausibly fall into that class. None of these duties disappears because training is federated, but the volume of personal data that crosses institutional boundaries, and so the surface exposed to a breach, falls sharply. Table 8.1 summarises the three regimes as they affect FL design.

Table 8.1: Data-protection regimes relevant to multi-hospital machine learning and their implications for federated designs

Regime	Scope and status	Treatment of health data	Implication for FL consortia
HIPAA Privacy Rule (United States)	Sectoral; binds covered entities and their business associates; in force since 2003 [6]	Protected health information; research use needs authorisation, waiver or de-identification [6]	FL reduces disclosures but model updates should still be covered by agreements; de-identification standards apply to any data that do move
GDPR (European Union)	Applies to all processing of personal data of people in the EU; applicable since 2018 [7]	Special category; pseudonymised data remain personal data [7]	Lawful basis and impact assessment still required; FL supports data minimisation but does not exempt processing [3]
DPDP Act 2023 and DPDP Rules 2025 (India)	Act enacted 11 August 2023; Rules notified November 2025 (gazette notifications dated 13 November) with 18-month phased compliance [8,9]	Consent-centred; breach report to Board within 72 hours; extra duties for Significant Data Fiduciaries [8,10]	Hospitals remain data fiduciaries for local training; consortium contracts must allocate breach, audit and algorithmic due-diligence duties [10]

2.2 Why de-identification is not enough

If de-identification worked perfectly, much of this would be moot. It does not. A systematic review of re-identification attacks on health data found that successful attacks had been demonstrated, although many involved small datasets or data not de-identified to existing standards [11]. The broader lesson from privacy research is that sparse, high-dimensional records are unusually distinctive: Narayanan and Shmatikov showed that a few known attributes could re-identify individuals in a large, nominally anonymised dataset of film ratings [12]. Medical imaging is no safer. Schwarz and colleagues reconstructed faces from research MRI scans and matched them to photographs of the participants using commercial face-recognition software [13]. Kaissis and co-authors describe naive de-identification as technically insufficient against identity inference and point out that removing facial features from head scans can degrade the images themselves [3].

The practical consequence is that institutions treat even de-identified clinical data as risky to release, and the governance overhead of release, which includes agreements, ethics review and audits, grows with the number of partners. A consortium of five hospitals may manage it. A consortium of seventy cannot negotiate bilateral transfers among every pair.

2.3 Why single-site models are not enough either

The scientific case for collaboration is just as strong as the legal case against pooling. Sheller and colleagues trained brain-tumour segmentation models on BraTS data partitioned by originating institution. Models trained at a single institution reached an average Dice coefficient of 0.732 on a shared test set, against 0.863 for a model trained on all data centrally [14]. Institutions with the least data produced the poorest models, and even the centrally trained model lost accuracy on data from institutions outside the original group, consistent with Zech's findings on limited diversity [1,14]. Rare conditions sharpen the point. Pati and colleagues note that glioblastoma, with an incidence of about three per 100,000, is rare by the usual epidemiological definition, so no single centre sees enough cases to train a general segmentation model [15].

3. Architectures for Collaborative Learning

3.1 Centralised, federated and swarm learning

Three architectures frame the design space (Figure 8.1). In **centralised learning** the raw data are copied to one location and a single party trains. This maximises statistical efficiency and simplifies debugging, because an engineer can inspect any record, but it concentrates legal and security risk.

In **federated learning** data stay with their custodians. A coordinating server distributes a model, each site trains on local data, and the server aggregates the returned parameters [5,16]. The coordinator never sees patient records, but it does see every update and it is a single point of trust and of failure.

Swarm learning, introduced by Warnat-Herresthal and colleagues in 2021, removes the coordinator. Edge nodes exchange parameters directly over a peer-to-peer network, and blockchain-based networking handles coordination, so no central party holds the model or orchestrates the merge [17]. The authors evaluated the approach on more than 16,400 blood transcriptomes from 127 clinical studies and more than 95,000 chest radiographs, across COVID-19, tuberculosis, leukaemia and lung pathologies, and found that swarm-trained classifiers outperformed those built at individual sites [17]. The blockchain is the distinctive element; the underlying parameter merging resembles federated averaging. Whether removing the coordinator is worth the additional infrastructure depends on whether any party in a

consortium can be trusted to run it. In a network led by a national agency it may be unnecessary; among commercial competitors it may be decisive.

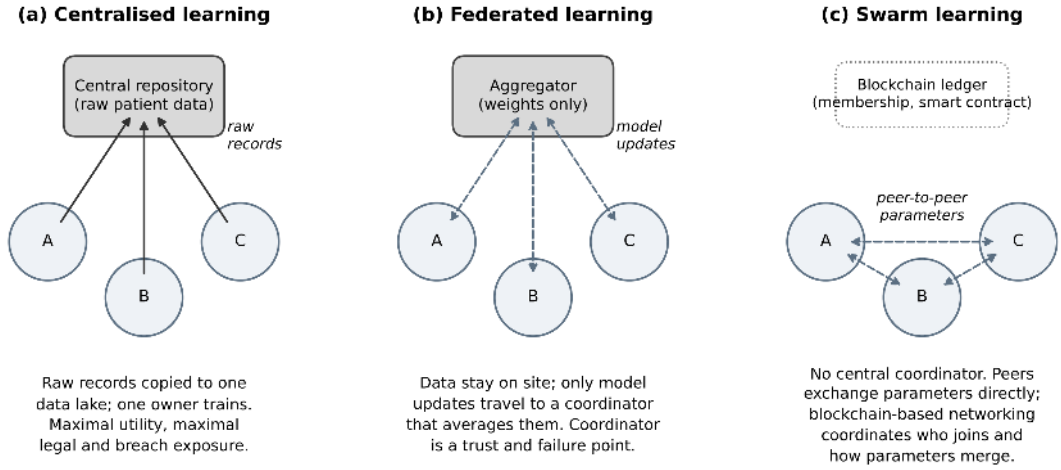

Figure 8.1: Centralised, federated and swarm learning compared. In (a) raw records move to a central repository; in (b) only model updates move, to an aggregator; in (c) peers exchange parameters directly, with blockchain-based coordination and no central server. Schematic based on descriptions in [5] and [17].

Before FL became standard, medical imaging researchers explored simpler schemes. In institutional incremental learning a model is trained to convergence at one site and then passed to the next; in cyclic institutional incremental learning it circulates repeatedly, a few epochs per stop. Chang and colleagues showed that cyclic weight transfer could approach the performance of centrally trained models in simulated multi-institution imaging tasks [18]. Sheller's group found, however, that these serial methods suffer from catastrophic forgetting: incremental learning favoured the last institution in the sequence, and cyclic learning produced quasi-periodic validation curves, whereas FL tracked centralised performance more closely [14]. Serial schemes also leave each site idle while it waits for the model to come round, which scales poorly.

3.2 Horizontal and vertical partitions; cross-silo and cross-device settings

Federated problems differ in how the data are split. In **horizontal FL**, the common case in hospital networks, sites hold different patients described by the same kinds of feature: every hospital has chest radiographs and admission vital signs. In **vertical FL**, sites hold different features for the same individuals, for example a hospital holding clinical records and an insurer holding claims for an overlapping population. Vertical FL needs privacy-preserving entity matching to align records and a split model whose parts are trained jointly, and it is considerably harder to engineer [16].

A second distinction concerns the participants. **Cross-device** FL involves very large numbers of unreliable clients, such as phones, each holding little data and participating occasionally; communication cost dominates and techniques such as update compression matter [16]. **Cross-silo** FL involves a handful to a few hundred organisations, each with substantial data, reliable connectivity and the ability to participate in every round [16]. Multi-hospital networks are cross-silo by nature. The distinction matters because several results from the cross-device literature, including the statistical properties of client sampling and some differential privacy guarantees that rely on large populations, transfer poorly to a federation of twenty hospitals.

4. Federated Averaging: The Core Algorithm

Consider K hospitals. Hospital k holds n_k labelled examples, and the total is n. With a model parameterised by weights w and a per-example loss ℓ, the goal is to minimise the same objective that centralised training would minimise:

$$f(w) \;=\; \sum_{k=1}^{K} \frac{n_k}{n}\, F_k(w), \qquad F_k(w) \;=\; \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} \ell\left(x_i, y_i; w\right)$$

where $\mathcal{P}$_k is the set of examples held at site k. If every hospital computed one gradient on its data and sent it to the server, which averaged and applied them, the result would be exactly centralised gradient descent, distributed across sites. The trouble is communication: deep networks need thousands of steps, and every step would require a round trip.

McMahan and colleagues proposed instead that each client perform several local steps before communicating [5]. In round t the server sends the current global weights w^t to a fraction C of clients. Each selected client initialises its local model at w^t and runs E epochs of stochastic gradient descent with mini-batch size B and learning rate η:

$$w \;\leftarrow\; w - \eta\, \nabla \ell(b; w) \quad \text{for each mini-batch } b \text{ of local data}$$

The client returns its final local weights w_k^{t+1}, and the server forms a data-weighted average:

$$w^{t+1} \;=\; \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j}\; w_k^{t+1}$$

where S_t is the set of clients participating in round t. Setting E = 1 with a single full batch recovers distributed gradient descent, which the authors call FedSGD. Increasing local computation trades communication for drift. On the benchmarks in the original paper, FedAvg needed 10 to 100 times fewer communication rounds than synchronised stochastic gradient descent, and it remained usable on unbalanced and non-identically distributed partitions [5]. Figure 8.2 shows one round as it would run in a hospital network.

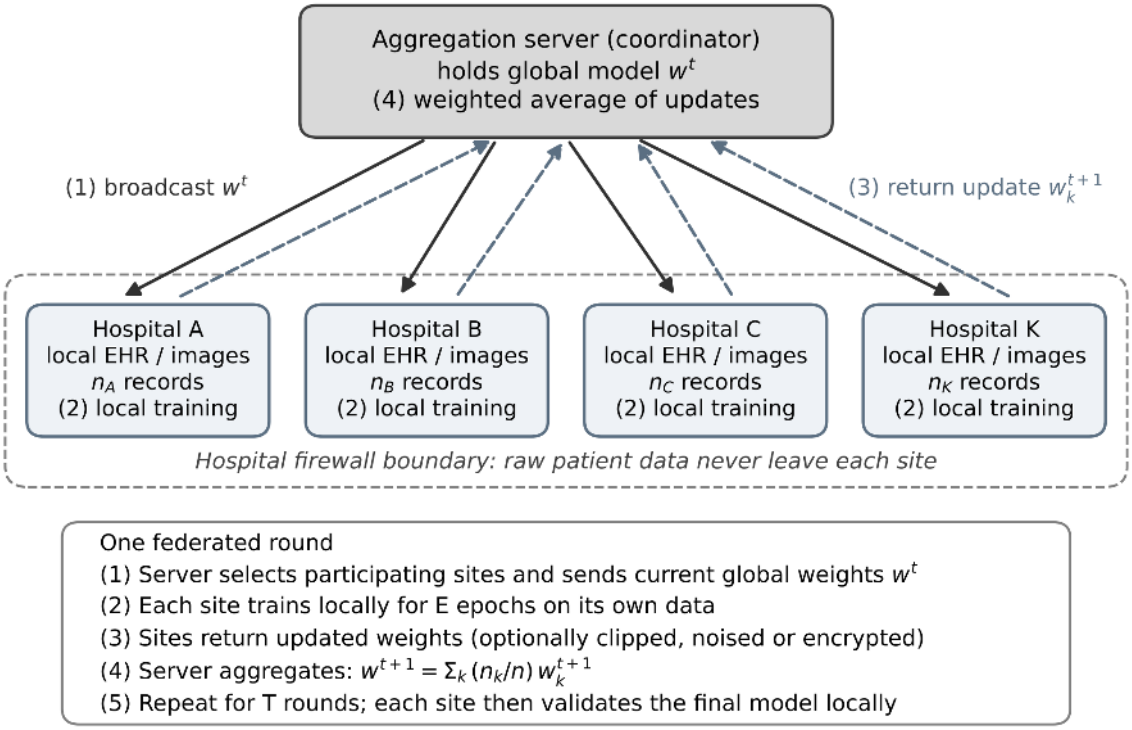

Figure 8.2: Hub-and-spoke federated learning across a hospital network, showing one round of Federated Averaging. Only model weights cross the institutional boundary; raw records remain behind each hospital's firewall. Original schematic based on the algorithm in [5].

Three practical points follow. First, the weighting by n_k means a large hospital pulls the global model toward its own distribution, which is efficient if sites are similar and unfair if they are not. Second, averaging is done on weights, not on predictions, so the procedure assumes every site trains an identical architecture from a common starting point. Third, the server learns a great deal from the sequence of updates, which motivates the privacy measures in Section 6. In cross-silo medical settings every site typically participates in every round (C = 1), and the number of rounds is modest: the EXAM study ran 200 federated rounds with one local epoch per round [19].

5. Statistical Heterogeneity and Its Remedies

5.1 What "non-IID" means in a hospital network

Classical convergence arguments assume that training examples are independent and identically distributed. Across hospitals they are not. Heterogeneity takes several forms. *Label skew* arises when case mix differs: a tertiary oncology centre sees far more tumours than a district hospital. *Feature or covariate shift* arises from differing scanners, acquisition protocols,

laboratory assays and documentation habits. *Concept shift* arises when the same label means different things, as when two sites apply different thresholds for oxygen therapy or annotate tumour boundaries differently. *Quantity skew* arises because sites differ by orders of magnitude in size. Figure 8.3 illustrates label skew with synthetic data. Partitioning a dataset with a Dirichlet distribution is a common simulation device: a large concentration parameter yields near-identical sites, and a small one yields sites dominated by one or two classes.

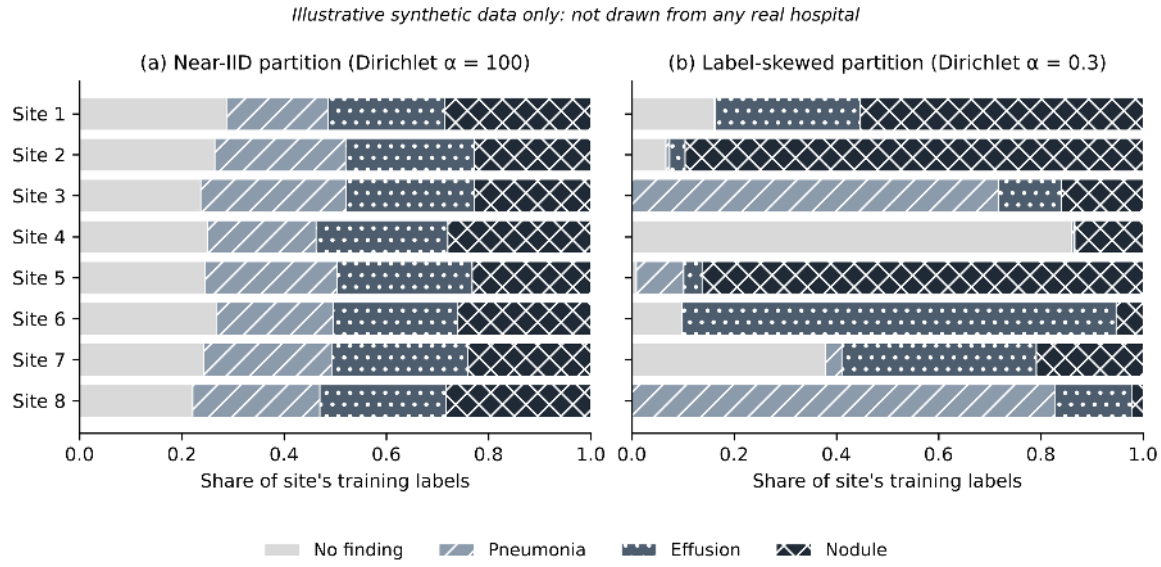

Figure 8.3: Illustrative non-IID label skew across eight synthetic sites. Panel (a) shows a near-IID partition and panel (b) a strongly skewed one, generated by Dirichlet sampling over four hypothetical radiograph labels. Synthetic data for illustration only; no real hospital data are shown.

Heterogeneity hurts FedAvg because each site's local optimum differs. During E local epochs every client moves toward its own optimum; the average of those endpoints is not, in general, the optimum of the global objective. Karimireddy and colleagues call this *client drift* [20]. Zhao and colleagues showed that on artificially label-partitioned benchmarks the accuracy of FedAvg could drop by as much as 55 per cent, and that sharing a small, globally available subset of data recovered much of the loss [21]. Sheller's group argued that such pathological partitions overstate the problem relative to real institutional splits, where FL came within about one Dice point of centralised training [14]. Both observations are useful. Severity depends on the task and the federation, and it has to be measured rather than assumed.

Heterogeneity also interacts with network architecture in ways that are easy to miss. Batch normalisation layers keep running estimates of feature statistics, and averaging those estimates across sites with different scanners can destabilise training. The EXAM investigators reported batch-normalisation instability in their non-IID setting and froze the pretrained image backbone in response [19].

5.2 FedProx

FedProx, from Li and colleagues, modifies the local objective with a proximal term that penalises distance from the current global model [22]:

$$\min_{w} h_k(w; w^t) \;=\; F_k(w) + \frac{\mu}{2} \parallel w - w^t \parallel^2$$

The coefficient μ controls how far a client may wander. With μ = 0 the method reduces to FedAvg. FedProx also tolerates *systems heterogeneity*: clients may perform variable amounts of local work and still contribute, which matters when one hospital has a modern GPU server and another a single workstation. The authors provide convergence guarantees under both forms of heterogeneity and report markedly more stable convergence than FedAvg in highly heterogeneous settings [22]. The cost is one additional hyperparameter, whose best value differs between tasks.

5.3 SCAFFOLD

SCAFFOLD attacks client drift directly using control variates, a variance-reduction technique [20]. The server maintains an estimate c of the global update direction and each client maintains its own estimate c_k. During local training the client corrects each gradient step:

$$y \;\leftarrow\; y - \eta \,(g_k(y) - c_k + c)$$

The correction subtracts the client's habitual bias and adds the global direction, so local steps track the global objective rather than the local one. Karimireddy and colleagues prove that SCAFFOLD needs significantly fewer communication rounds than FedAvg and is unaffected by data heterogeneity or client sampling [20]. In practice it doubles the information exchanged per round, because control variates travel alongside weights, and it is sensitive to clients that drop out mid-training. For a stable cross-silo federation these are acceptable costs.

5.4 Adaptive server optimisers and personalisation

A different line of work treats the averaged client change as a pseudo-gradient and applies an adaptive optimiser, such as Adam or Yogi, on the server. Reddi and colleagues formalised this family, often called FedOpt, analysed its convergence when client data are heterogeneous, and reported that adaptive server optimisers can markedly improve federated performance [23].

Personalised FL goes further and abandons the assumption that one model should serve all sites. If two hospitals genuinely differ, a single global model may be a poor compromise for both. Approaches include fine-tuning the global model locally, keeping some layers private to each site, and multi-task formulations. Ditto, from Li and colleagues, trains a personalised model at each client regularised toward the global model and frames personalisation as a route to both fairness and resilience against corrupted clients [24]. Keeping normalisation statistics

local is another simple form of personalisation, and it also helps defend against certain reconstruction attacks [25]. MELLODDY adopted a structural form: a shared trunk network learned jointly and private task-specific heads retained by each partner [26].

Personalisation has a clinical cost that is often unstated. A model tuned to each hospital is, by construction, a different model at each hospital, and each one needs local validation and potentially its own regulatory assessment. The governance problem shifts rather than disappears.

6. Privacy: Threats and Defences

6.1 What model updates reveal

Federated learning was initially promoted on the intuition that gradients and weights are aggregate quantities from which individual records cannot be recovered. That intuition is wrong in general.

Membership inference asks whether a particular individual's record was in the training set. Shokri and colleagues showed that an adversary with only query access to a model could train "shadow" models to recognise the subtle differences in a target model's outputs on members and non-members [27]. In medicine, membership alone can be sensitive: knowing that a person's record trained a model for HIV-associated complications discloses diagnosis. Model inversion attacks go further and reconstruct representative inputs from a model's outputs or confidence scores [28], and neural networks can memorise rare sequences in training data verbatim [29]. In collaborative training specifically, Hitaj and colleagues showed that a malicious participant could use a generative adversarial network to synthesise samples resembling another participant's private class [30]. Sheller's group noted that, because a participant can tamper with its own training, any collaborative scheme in which participants receive the shared model is open to attacks of this kind [14].

Gradient inversion is the most direct threat. In Deep Leakage from Gradients (DLG), Zhu, Liu and Han initialised random dummy inputs and labels, computed the gradients they would produce, and optimised the dummies until their gradients matched the observed ones [31]. For small batches the recovered images were close to pixel-accurate. The limits were real but not reassuring: the method worked up to a batch size of eight and images of 64 by 64 pixels, and needed a twice-differentiable network [31]. Geiping and colleagues then showed that a magnitude-invariant loss allows faithful reconstruction of higher-resolution images even from trained networks, and that aggregating gradients over several iterations or several images does not guarantee privacy [32].

The healthcare-specific evidence is more nuanced. Hatamizadeh and colleagues tested gradient inversion against realistic FL settings on chest radiographs. Earlier attacks failed when clients

updated batch-normalisation statistics during training, and succeeded only when the attacker could estimate those statistics and held the current global model [25]. Larger batches, more local images and more local iterations made inversion substantially harder. A client training on a single image at batch size one remained highly exposed. The authors concluded that earlier work had probably exaggerated the risk for typical FL configurations, while still recommending that the global model state, the architecture and normalisation statistics be protected [25]. This is the right reading for hospital consortia: the risk is real for small or degenerate local datasets and modest for large sites with standard training settings. A federation that admits a clinic with forty patients needs stronger protection than one of ten tertiary centres.

6.2 Differential privacy

Differential privacy (DP) offers a mathematical guarantee. A randomised mechanism M is (ε, δ)-differentially private if, for any two datasets D and D′ that differ in one individual's record and any set of outcomes S, Pr[M(D) ∈ S] ≤ e^ε Pr[M(D′) ∈ S] + δ [33]. Smaller ε means stronger protection. In deep learning the standard mechanism is DP-SGD: clip each example's gradient to a fixed norm, add calibrated Gaussian noise, and track the cumulative privacy loss over training. Abadi and colleagues showed that deep networks with non-convex objectives could be trained this way under a modest privacy budget, with manageable costs in training efficiency and model quality [34]. In FL, noise can be added locally by each site, protecting against a curious server, or centrally at aggregation, protecting the released model against outsiders.

The trade-off is unavoidable. In the DLG experiments, Gaussian noise with variance 10^{-2} blocked reconstruction, but accuracy on the CIFAR-100 task fell from 76.3 per cent to 45.3 per cent [31]. Hatamizadeh and colleagues found that DP-SGD reduced leakage for their most exposed client but hurt global accuracy more than a simpler noise protocol [25]. Kaissis and co-authors caution that perturbation can degrade performance in data-scarce medical fields and that more work is needed before DP is used routinely in medical imaging [3]. Lighter defences also help. Pruning the smallest 20 per cent or more of gradient entries already produced visible artefacts in DLG reconstructions and higher sparsity defeated the attack, at little cost to accuracy [31]. EXAM shared only a fraction of the largest weight updates and reported comparable performance when sharing 25 per cent [19]. Such heuristics reduce leakage without a formal guarantee.

6.3 Secure aggregation, homomorphic encryption and secure multi-party computation

Cryptographic defences protect the updates in transit and at the server, rather than the model that results. Secure aggregation, as designed by Bonawitz and colleagues, lets the server compute the sum of client updates without seeing any individual update. Clients add pairwise

random masks that cancel in the sum, and secret sharing lets the protocol tolerate clients that drop out mid-round [35]. The overhead is modest: for 16-bit inputs, 2^{10} users and 2^{20}-dimensional vectors, communication expanded by a factor of 1.73 relative to sending data in the clear [35]. Secure aggregation removes the curious-server threat but does nothing about what the final model reveals, so it is usually paired with DP.

Homomorphic encryption (HE) allows computation directly on ciphertexts. In FL, clients can encrypt updates under a scheme that supports addition, the server sums the ciphertexts, and only key-holders decrypt the aggregate. The obstacle has been computational cost. Froelicher and colleagues demonstrated multiparty homomorphic encryption for federated analytics in precision medicine, reproducing published Kaplan-Meier survival analyses and genome-wide association studies across distributed datasets without revealing intermediate data [36]. Secure multi-party computation splits data or model parameters into secret shares held by different parties; Kaissis and colleagues note that it scales poorly beyond a few parties and requires all of them to remain online [3].

PriMIA, an open-source framework from Kaissis and colleagues, combined these elements in an end-to-end demonstration. It used differentially private, securely aggregated FL to train a paediatric chest radiograph classifier and secure multi-party computation for encrypted inference. The resulting model performed comparably to locally trained models without protection, and the protections prevented a gradient-based model inversion attack from reconstructing usable images [37].

6.4 Trusted execution environments

Trusted execution environments (TEEs) are hardware-isolated enclaves in which code and data are protected even from the host operating system. In FL, aggregation, or even local training, can run inside an enclave whose integrity remote parties verify by attestation. Kaissis and co-authors observe that such hardware can preserve data and algorithm privacy even after a breach of the operating system kernel and expect it to become more common [3]. TEEs avoid much of the computational overhead of cryptographic methods but shift trust to the hardware vendor and its attestation service. They are best viewed as one layer among several.

6.5 Integrity as well as confidentiality

Privacy is not the only security property at stake. A participant that submits poisoned updates, accidentally through corrupted labels or deliberately, can degrade the global model or implant behaviour that activates on particular inputs. Secure aggregation makes such behaviour harder to detect, because the server cannot inspect individual updates. Byzantine-resilient aggregation rules and personalised formulations such as Ditto offer partial protection [16,24]. In a clinical federation the more realistic threat is not a malicious hospital but a careless one: a site whose labels were extracted with a faulty script. Table 8.2 summarises the threat model.

Table 8.2: Privacy and integrity threats in federated learning, with defences and their trade-offs

Threat	Mechanism and evidence	Main defences	Trade-offs
Gradient inversion	Optimise dummy inputs until gradients match; near pixel-accurate at small batch sizes [31,32]	Larger batches and local epochs; keep BN statistics local; secure aggregation; DP [25]	Risk concentrated in small sites; DP lowers accuracy [25,31]
Membership inference	Shadow models detect output differences between members and non-members [27]	Differential privacy; regularisation [33,34]	Strong ε reduces utility, especially for rare classes [3]
Model inversion and memorisation	Reconstruct representative inputs from confidence scores; verbatim memorisation of rare sequences [28,29]	DP; limit output detail; audit for memorisation [3]	Hard to quantify residual risk
Malicious participant (GAN attack)	Adversarial participant synthesises samples of others' classes [30]	Identity management, tamper-resistant hardware, vetting of participants [14]	Requires legal and organisational controls, not only technical ones
Curious server	Server inspects individual updates	Secure aggregation [35]; homomorphic encryption [36]; TEEs [3]	Communication or compute overhead; HE is costly; TEEs shift trust to vendor
Poisoning or faulty site	Corrupted or adversarial updates degrade or backdoor the model	Byzantine-resilient aggregation; personalisation; per-site validation [16,24]	Conflicts with secure aggregation, which hides individual updates

7. Landmark Deployments in Medicine

7.1 Brain tumours: from feasibility to seventy-one sites

The brain-tumour programme led from the University of Pennsylvania with Intel illustrates how medical FL matured. A 2018 feasibility study first showed, on BraTS data split by institution, that multi-institutional deep learning was possible without sharing patient data [38]. The extended 2020 study in *Scientific Reports* reported that FL among ten institutions produced models reaching about 99 per cent of the quality of centrally trained models [14]. On the BraTS test set, the centrally trained model scored a Dice coefficient of 0.863 and the federated model 0.858; on leave-one-institution-out validation the figures were 0.840 and 0.835 [14]. The federated model needed about 2.26 times as many epochs to converge. In the same period, Li and colleagues at NVIDIA and King's College London studied differentially private federated brain-tumour segmentation and examined the trade-off between privacy protection and segmentation quality [39].

The Federated Tumor Segmentation (FeTS) initiative then took the method into the field. Pati and colleagues coordinated what they described as the largest FL study to date: 71 sites across six continents, contributing data from 6,314 cases to train a glioblastoma sub-compartment segmentation model [15]. Compared with a model trained on publicly available data, the final federated model improved delineation by 33 per cent for the surgically targetable tumour and 23 per cent for the complete tumour extent [15]. Validation used a local set of 1,043 cases and

an out-of-sample set of 518 cases [15]. The study's contribution is as much organisational as numerical: it showed that a federation of this size, spanning very different health systems, can be coordinated to a common protocol at all.

7.2 EXAM: COVID-19 oxygen requirements across twenty institutes

The EXAM study (Electronic Medical Record Chest X-ray AI Model) remains the clearest example of FL producing a clinically framed model under time pressure. Dayan and colleagues brought together twenty institutes in the United States, Europe, Asia and Brazil to predict the oxygen therapy a symptomatic COVID-19 patient would need 24 and 72 hours after presenting to the emergency department [19]. Inputs were a chest radiograph and routinely collected vital signs and laboratory values; the image branch used a ResNet-34 with spatial attention, and training used client-server Federated Averaging over 200 rounds on 16,148 cases [19]. The model achieved an average area under the curve (AUC) above 0.92 for both horizons, improved average AUC by 16 per cent relative to locally trained models, and increased generalisability by 38 per cent [19]. At the largest independent test site, sensitivity was 0.950 and specificity 0.882 for predicting mechanical ventilation or death at 24 hours, and at Cooley Dickinson Hospital, which did not take part in training, AUCs were 0.944 at 24 hours and 0.924 at 72 hours [19].

The limitations the authors report are as instructive as the results. Oxygen delivery served as a proxy for oxygen need, and prophylactic oxygen may have distorted labels. Data were not harmonised across sites, and failures traced to missing data, motion artefacts and out-of-distribution patients. Post hoc analysis of failures was hampered because investigators could not see the data, and population drift was not assessed [19]. Recruitment ran from March to May 2020 and data curation from June to September, overlapping the 140 federated runs conducted from August to October [19]. Curation took at least as long as the experiments, a pattern that recurs throughout this literature. The model was not approved by a regulator and was intended for research use.

7.3 Other clinical studies

Several smaller studies fill in the picture. Brisimi and colleagues built predictive models from federated electronic health records, an early demonstration outside imaging [40]. Roth and colleagues reported a real-world federated implementation of mammographic breast-density classification across seven clinical institutions, a task in which differences in equipment and reading practice between institutions are expected [41]. Sarma and colleagues found that federated training of a prostate segmentation model across multiple academic centres improved performance at the participating sites relative to models trained locally, without data sharing [42].

7.4 MELLODDY: federated learning among pharmaceutical competitors

FL is not confined to hospitals. The MELLODDY project brought together ten pharmaceutical companies: Amgen, Astellas, AstraZeneca, Bayer, Boehringer Ingelheim, GSK, Janssen, Merck KGaA, Novartis and Servier. Over three years they trained quantitative structure-activity relationship (QSAR) models on more than 2.6 billion confidential activity data points covering over 21 million compounds and more than 40,000 assays [26]. The platform was built on the open-source Substra software. Each partner kept its data and a task-specific prediction head; only updates to a shared trunk were exchanged, combined by secure aggregation, and compound fingerprints were shuffled with a key held by the platform operator so that chemical structures could not be read from the shared features [26].

The results were positive but modest, which makes them credible. Every partner saw some improvement. For classification tasks, eight partners recorded median relative gains in area under the precision-recall curve above 2.5 per cent and one above 7.5 per cent; for regression, all but one partner improved median R^2; and the applicability domain, measured by conformal efficiency, widened by more than 10 per cent for six partners [26]. Gains were largest for absorption, distribution, metabolism and excretion (ADME) and safety-panel assays, and the best results came from adding auxiliary high-throughput screening and imaging data [26]. For a pharmacy and pharmaceutical-sciences readership, MELLODDY is the reference case for collaboration among competitors: a group of rivals accepted that a shared trunk improved everyone’s models without exposing any company’s proprietary chemistry.

Table 8.3: Landmark federated and swarm learning studies in healthcare and drug discovery

Study	Sites and data	Task	Federated approach	Key result
Sheller et al. 2020 [14]	10 institutions (BraTS glioma MRI)	Brain-tumour segmentation	FedAvg compared with serial institutional learning	FL reached about 99% of centralised quality; Dice 0.858 vs 0.863
Dayan et al. 2021, EXAM [19]	20 institutes, 4 continents; 16,148 cases	Oxygen requirement at 24 and 72 h in COVID-19	Client-server FedAvg, 200 rounds; partial weight sharing	Mean AUC above 0.92; +16% AUC and +38% generalisability vs local models
Warnat-Herresthal et al. 2021 [17]	Over 16,400 transcriptomes from 127 studies; over 95,000 chest X-rays	COVID-19, tuberculosis, leukaemia, lung pathologies	Swarm learning, blockchain-coordinated, no central server	Outperformed classifiers built at individual sites
Pati et al. 2022, FeTS [15]	71 sites, 6 continents; 6,314 cases	Glioblastoma sub-compartment segmentation	FL on open-source stack	33% and 23% delineation gains vs public-data model
Kaissis et al. 2021, PriMIA [37]	Multi-institutional paediatric chest X-rays	Pneumonia classification	DP and secure aggregation; encrypted inference	Performance comparable to unprotected local models; inversion attack thwarted

Study	Sites and data	Task	Federated approach	Key result
Sarma et al. 2021 [42]	Multiple academic centres	Prostate MRI segmentation	Client-server FL	Improved site performance vs local training
Heyndrickx et al. 2024, MELLODDY [26]	10 pharmaceutical companies; over 2.6 billion data points, 21 million compounds	Multi-task QSAR	Shared trunk, private heads; secure aggregation	Gains for every partner, largest for ADME and safety assays

8. Software Frameworks

The maturity of FL in medicine owes much to open-source frameworks, which absorb the plumbing of networking, security and orchestration so that clinical teams can concentrate on data and models.

NVIDIA FLARE (NVFlare) is an open-source Python SDK whose documented use cases include COVID-19 outcome prediction of the EXAM type. Its design separates a server-side controller, which assigns tasks, from client-side executors, which carry them out. Communication runs over gRPC with mutual TLS; provisioning issues each site a signed start-up kit; and filters on the client side can apply differential privacy, homomorphic encryption or exclusion of variables before an update leaves the site [43]. It supports FedAvg, FedProx, FedOpt, SCAFFOLD and cyclic weight transfer, and also peer-to-peer patterns including swarm learning [43].

Flower, described by Beutel and colleagues, is framework-agnostic. A pluggable *strategy* object encapsulates client selection, configuration and aggregation, and clients can wrap PyTorch, TensorFlow or other libraries. The authors report simulations of up to 15 million clients on a pair of high-end GPUs [44]. TensorFlow Federated, from Google, is tightly integrated with TensorFlow and has been oriented toward simulation and research; the Flower authors characterise its multi-node deployment as planned rather than available at the time of their comparison [44,45]. PySyft, a generic framework for privacy-preserving deep learning described by Ryffel and colleagues, emphasises primitives such as secure multi-party computation [46]; PriMIA, discussed in Section 6.3, is a dedicated framework for privacy-preserving medical imaging [37].

OpenFL, developed by Intel with the University of Pennsylvania, was designed as the back end of the FeTS platform and is described in *Physics in Medicine and Biology* [47]. Its project page now states that OpenFL is no longer under active development, will be archived, and recommends migration to Flower [48]. Institutions choosing a stack should weigh community activity and long-term support as heavily as features. A framework that loses its maintainers leaves a hospital network carrying security patches it may not be equipped to write.

Table 8.4: Open-source frameworks for federated learning in healthcare (status as of November 2026 where noted)

Framework	Origin	Architecture and notable features	Medical use and status
NVIDIA FLARE [43]	NVIDIA	Controller-executor model; mutual TLS provisioning; DP and HE filters; FedAvg, FedProx, FedOpt, SCAFFOLD, cyclic and swarm workflows	Use cases include mammography, prostate segmentation and COVID-19 outcome prediction [19,43]
Flower [44]	Beutel et al.; open-source community	Framework-agnostic; strategy abstraction; gRPC; simulation engine scaling to millions of virtual clients	General-purpose; recommended migration path for OpenFL users [48]
TensorFlow Federated [45]	Google	TensorFlow-native; simulation-focused research platform [44]	Mainly research and simulation
OpenFL [47]	Intel and University of Pennsylvania	Python library; designed as back end of the FeTS platform	Used in FeTS-related work; project page states it is no longer actively developed [48]
PySyft / PriMIA [37,46]	Ryffel et al.; Kaissis et al.	Secure multi-party computation, DP and secure aggregation; encrypted inference	Paediatric chest X-ray classification demonstration [37]

9. Governance, Incentives and Infrastructure for Indian Hospital Networks

9.1 Governance under the DPDP regime

For an Indian consortium, the first governance question is who is the data fiduciary for what. Each hospital that trains locally remains the fiduciary for its patients' data and carries the obligations that the Act and Rules attach to that role: notice, consent or another lawful ground, security safeguards and breach reporting [8,10]. The coordinator, whether a university, a government programme or a vendor, receives model updates. Section 6 showed that updates can, in some configurations, leak information about individuals. A conservative reading therefore treats the coordinator as a party that processes derived personal data and binds it by contract to the same security standards, with secure aggregation or encryption designed to make that exposure negligible. Consortium agreements should allocate in advance who reports a breach within the seventy-two-hour window, who conducts impact assessments, and who performs the algorithmic due diligence that the Rules require of Significant Data Fiduciaries [10].

Consent deserves care. Indian hospitals have historically relied on broad treatment consent, and the Rules require notices that are clear and itemised and consent that is specific and freely given [10]. Federated training for a defined clinical purpose is easier to explain to patients than open-ended data transfer, and the consent-manager architecture, which the Rules bring into force from November 2026, could in principle let patients see and withdraw consent for research uses of their records [8,9]. Whether withdrawal requires retraining a model that has

already learned from a patient's data is unresolved in law and in technique. DP offers one answer, by bounding any individual's influence, but no Indian guidance on the point existed in the sources reviewed for this chapter.

9.2 Incentives

Federations form when each member gains more than it gives. The evidence shows that small sites gain most: in Sheller's data the institutions with the least data produced the poorest single-site models, and in EXAM average gains over local models were large [14,19]. Large sites contribute most and gain least, and in a competitive private hospital sector that asymmetry can stall collaboration. MELLODDY's design suggests mitigations: partners kept private heads, so each retained a model tuned to its own assays, and each could verify its own gains [26]. Transparent per-site evaluation, shared authorship and publication, and agreed rules on commercial exploitation of the resulting model are probably more important than any algorithmic contribution-scoring scheme. Public funders can tip the balance by making participation a condition of grants for AI development.

9.3 Infrastructure and data readiness

The binding constraint for most hospitals is not the federation software; it is the data. EXAM's curation phase took four months across twenty well-resourced institutions, and data that were never fully harmonised still caused failures [19]. FeTS had to bring seventy-one sites on six continents to a common protocol [15]. A hospital that cannot export structured laboratory values, or whose radiographs lack consistent metadata, cannot participate meaningfully however good the network. Interoperability programmes, such as the national digital-health infrastructure discussed in Chapter 10, matter here because they standardise exactly the identifiers, terminologies and record formats that federated models consume.

Compute requirements are modest compared with centralised training at scale, but they are not trivial. Each site needs a GPU-capable server inside its security perimeter, stable outbound connectivity, certificate management and staff who can run and monitor jobs. In NVFlare, client sites open no inbound ports and authenticate with mutual TLS, which simplifies firewall approvals [43]. For smaller district or private hospitals, a realistic model may be a regional node, hosted by a medical college or state health agency, that trains on behalf of several facilities under appropriate agreements. That is a hybrid of federation and pooling, and it should be described honestly as such.

10. Open Problems: Fairness and Evaluation

10.1 Whose performance improves?

A federated model optimises a data-weighted average objective. Unless the design says otherwise, the largest sites dominate and the smallest, often those serving rural or poorer populations, may see little benefit or even a decline relative to a local model. Kairouz and colleagues discuss fairness in FL as a distinct research problem, including objectives that weight worse-performing clients more heavily [16], and Ditto explicitly targets fairness across clients through personalisation [24]. Fairness across demographic groups within sites is harder, because FL makes it more difficult to audit the training data at all. A coordinator cannot easily check whether women, older patients or particular ethnic groups are under-represented without collecting, at least in aggregate, the very attributes that federation was designed to keep local.

10.2 How should a federated model be evaluated?

Evaluation in FL is complicated by the inability to inspect data. Several principles emerge from the studies reviewed here. Performance should be reported per site, not only pooled, because a high average can hide failure at a minority of sites. Validation on institutions that did not contribute to training is essential, as EXAM and FeTS both did [15,19]. Comparisons should include both local models and, where feasible, a centralised model on the same data, as Sheller's group did, so that the cost of federation is visible [14]. Post hoc failure analysis, which EXAM's authors found hampered by restricted access [19], may need privacy-preserving tools of its own, such as federated error analysis or secure enclaves for clinicians to review flagged cases.

Federated studies also inherit every weakness of conventional clinical AI evaluation. Retrospective AUCs do not demonstrate clinical benefit, labels can be proxies, as EXAM's oxygen-delivery outcome showed, and models drift after deployment [19]. Prospective, multi-site evaluation of a federated model's effect on decisions and outcomes is still rare. Shared benchmarks with realistic institutional splits, rather than artificial Dirichlet partitions, would help the field compare algorithms on the heterogeneity that hospitals actually face [14,21].

10.3 Quantifying residual privacy risk

Finally, the privacy of a deployed FL system is rarely quantified end to end. Most clinical studies report which defences were used but not the resulting ε or the outcome of an attack audit. Hatamizadeh and colleagues proposed measures of data leakage intended to make such audits reproducible [25], and PriMIA reported the result of an inversion attack against its own model [37]. Regulators and ethics committees would be better served if federated studies routinely published a threat model, the privacy parameters chosen and the results of a standard attack battery.

11. Conclusion and Future Directions

Federated learning has moved, in less than a decade, from a technique for mobile devices to the method behind some of the largest multi-institutional studies in medical imaging and drug discovery. The evidence supports a measured conclusion. Across brain-tumour segmentation, COVID-19 outcome prediction and pharmaceutical QSAR, federated models approached centrally trained models and clearly outperformed models trained at single sites [14,15,19,26]. The method works.

What it does not do is equally clear. FL does not make models private by default; gradient inversion and membership inference are genuine threats whose severity depends on configuration, and formal guarantees require differential privacy, with a cost in accuracy [25,31,34]. It does not remove heterogeneity, which must be addressed with algorithms such as FedProx, SCAFFOLD or personalisation, and above all by measurement [20,22,24]. It does not remove the need for data curation, which in EXAM took at least as long as the federated experiments [19]. And it does not remove legal obligations: under India's DPDP Act and Rules, as under GDPR and HIPAA, each hospital remains accountable for the personal data it holds and processes [6-8].

Several directions deserve priority. Federated fine-tuning of large pretrained foundation models could reduce the data each site must contribute, though it raises new questions about memorisation. Standard reporting of per-site performance, external validation and privacy parameters would make federated studies easier to judge. For India, the most useful near-term step is probably less algorithmic than institutional: a small number of well-governed federations around specific clinical questions, built on interoperable records and contracts aligned with the DPDP Rules, that show hospitals and patients what the method can deliver and what it costs.

REFERENCES

1. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683. https://doi.org/10.1371/journal.pmed.1002683
2. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993-2024. https://doi.org/10.1109/TMI.2014.2377694
3. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-11. https://doi.org/10.1038/s42256-020-0186-1
4. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119. https://doi.org/10.1038/s41746-020-00323-1

5. McMahan B, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR. 2017;54:1273-82. Available from: https://proceedings.mlr.press/v54/mcmahan17a.html
6. Annas GJ. HIPAA regulations: a new era of medical-record privacy? N Engl J Med. 2003;348:1486-90.
7. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR): a practical guide. Cham: Springer; 2017.
8. Press Information Bureau, Government of India. Digital Personal Data Protection Rules, 2025 [backgrounder]. New Delhi: Ministry of Electronics and Information Technology; 17 November 2025 [cited 2026 Nov 3]. Available from: https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc20251117695301.pdf
9. AZB & Partners. India's Digital Personal Data Protection Act: phased rollout and key compliance milestones [Internet]. Mumbai: AZB & Partners; 14 November 2025 [cited 2026 Nov 3]. Available from: https://www.azbpartners.com/bank/indias-digital-personal-data-protection-act-phased-rollout-and-key-compliance-milestones/
10. EY India. Transforming data privacy: Digital Personal Data Protection Rules, 2025 [Internet]. 2025 [cited 2026 Nov 3]. Available from: https://www.ey.com/en_in/insights/cybersecurity/transforming-data-privacy-digital-personal-data-protection-rules-2025
11. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One. 2011;6(12):e28071. https://doi.org/10.1371/journal.pone.0028071
12. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy. IEEE; 2008. p. 111-25.
13. Schwarz CG, Kremers WK, Therneau TM, Sharp RR, Gunter JL, Vemuri P, et al. Identification of anonymous MRI research participants with face-recognition software. N Engl J Med. 2019;381(17):1684-6. https://doi.org/10.1056/NEJMc1908881
14. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598. https://doi.org/10.1038/s41598-020-69250-1
15. Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun. 2022;13:7346. https://doi.org/10.1038/s41467-022-33407-5
16. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. Found Trends Mach Learn. 2021;14(1-2):1-210. https://doi.org/10.1561/2200000083
17. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature. 2021;594(7862):265-70. https://doi.org/10.1038/s41586-021-03583-3

18. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc. 2018;25(8):945-54. https://doi.org/10.1093/jamia/ocy017
19. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021;27(10):1735-43. https://doi.org/10.1038/s41591-021-01506-3
20. Karimireddy SP, Kale S, Mohri M, Reddi S, Stich S, Suresh AT. SCAFFOLD: stochastic controlled averaging for federated learning. In: Proceedings of the 37th International Conference on Machine Learning. PMLR. 2020;119:5132-43. Available from: https://proceedings.mlr.press/v119/karimireddy20a.html
21. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-IID data. arXiv [Preprint]. 2018. arXiv:1806.00582.
22. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. Proc Mach Learn Syst. 2020;2:429-50.
23. Reddi S, Charles Z, Zaheer M, Garrett Z, Rush K, Konečný J, et al. Adaptive federated optimization. arXiv [Preprint]. 2020. arXiv:2003.00295.
24. Li T, Hu S, Beirami A, Smith V. Ditto: fair and robust federated learning through personalization. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021). 2021. p. 6357-68.
25. Hatamizadeh A, Yin H, Molchanov P, Myronenko A, Li W, Dogra P, et al. Do gradient inversion attacks make federated learning unsafe? IEEE Trans Med Imaging. 2023. https://doi.org/10.1109/TMI.2023.3239391
26. Heyndrickx W, Mervin L, Morawietz T, Sturm N, Friedrich L, Zalewski A, et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model. 2024;64(7):2331-44. Epub 2023 Aug 29. https://doi.org/10.1021/acs.jcim.3c00799
27. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017. https://doi.org/10.1109/SP.2017.41
28. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM; 2015. p. 1322-33. https://doi.org/10.1145/2810103.2813677
29. Carlini N, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: evaluating and testing unintended memorization in neural networks. In: Proceedings of the 28th USENIX Security Symposium. USENIX Association; 2019. p. 267-84.
30. Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM; 2017. p. 603-18.
31. Zhu L, Liu Z, Han S. Deep leakage from gradients. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019). 2019. p. 14774-84.

32. Geiping J, Bauermeister H, Dröge H, Moeller M. Inverting gradients: how easy is it to break privacy in federated learning? In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020). 2020. arXiv:2003.14053.
33. Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci. 2014;9(3-4):211-407. https://doi.org/10.1561/0400000042
34. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM; 2016. https://doi.org/10.1145/2976749.2978318
35. Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, et al. Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM; 2017. https://doi.org/10.1145/3133956.3133982
36. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun. 2021;12:5910. https://doi.org/10.1038/s41467-021-25972-y
37. Kaissis G, Ziller A, Passerat-Palmbach J, Ryffel T, Usynin D, Trask A, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat Mach Intell. 2021;3(6):473-84. https://doi.org/10.1038/s42256-021-00337-8
38. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2018). Lecture Notes in Computer Science, vol 11383. Cham: Springer; 2019. p. 92-104. https://doi.org/10.1007/978-3-030-11723-8_9
39. Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-preserving federated brain tumour segmentation. In: Machine Learning in Medical Imaging (MLMI 2019). Lecture Notes in Computer Science, vol 11861. Cham: Springer; 2019. p. 133-41. https://doi.org/10.1007/978-3-030-32692-0_16
40. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform. 2018;112:59-67.
41. Roth HR, Chang K, Singh P, Neumark N, Li W, Gupta V, et al. Federated learning for breast density classification: a real-world implementation. In: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (DART 2020, DCL 2020). Lecture Notes in Computer Science, vol 12444. Cham: Springer; 2020. p. 181-91. https://doi.org/10.1007/978-3-030-60548-3_18
42. Sarma KV, Harmon S, Sanford T, Roth HR, Xu Z, Tetreault J, et al. Federated learning improves site performance in multicenter deep learning without data sharing. J Am Med Inform Assoc. 2021;28(6):1259-64. https://doi.org/10.1093/jamia/ocaa341
43. Roth HR, Cheng Y, Wen Y, Yang I, Xu Z, Hsieh YT, et al. NVIDIA FLARE: federated learning from simulation to real-world. arXiv [Preprint]. 2022. arXiv:2210.13291.
44. Beutel DJ, Topal T, Mathur A, Qiu X, Fernandez-Marques J, Gao Y, et al. Flower: a friendly federated learning research framework. arXiv [Preprint]. 2020. arXiv:2007.14390.

45. Google. TensorFlow Federated: machine learning on decentralized data [Internet]. [cited 2026 Nov 3]. Available from: https://www.tensorflow.org/federated
46. Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D, et al. A generic framework for privacy preserving deep learning. arXiv [Preprint]. 2018. arXiv:1811.04017.
47. Foley P, Sheller MJ, Edwards B, et al. OpenFL: the open federated learning library. Phys Med Biol. 2022. https://doi.org/10.1088/1361-6560/ac97d9
48. OpenFL contributors. OpenFL: an open-source framework for federated learning [Internet]. GitHub; [cited 2026 Nov 3]. Available from: https://github.com/intel/openfl