

Enhancing Privacy and Collaborative Efforts in Healthcare through Federated Machine Learning

DHARAMDAS KUMHAR¹, ANIL KEWAT²
^{1, 2} Bundelkhand University, Jhansi UP India

Abstract- Healthcare institutions create data from complementary electronic health records, medical images, laboratory measurements, genomic profiles, and connected devices. While aggregating this data can enhance machine-learning models, centralized collection increases exposure risks, conflicts with institutional data sovereignty, and is often constrained by legal, ethical, and operational challenges. Federated Machine Learning (FML) provides a collaborative framework where participating organizations can train a common model while keeping raw data stored locally. This paper presents a design-oriented review and a practical framework for privacy-aware, cross-silo FML in healthcare. The framework integrates local data governance, standardized preprocessing, federated optimization, secure aggregation, differential privacy, update validation, audit logging, and continuous clinical monitoring. It explicitly distinguishes data locality from formal privacy guarantees: while federated training minimizes raw data movement, model updates can still leak information or be subject to manipulation. An illustrative experiment involving five statistically heterogeneous synthetic hospitals evaluated centralized learning, isolated local learning, federated averaging, and clipped-and-noised federated averaging. The results indicate that beneficial collaboration is technically feasible, but privacy, fairness, interoperability, governance, and clinical validation must be engineered as integral parts of the joint lifecycle requirements rather than treated as consequences of federation alone.

Index Terms- Collaborative healthcare, differential privacy, federated machine learning, healthcare artificial intelligence, secures aggregation.

I. INTRODUCTION

Machine learning in healthcare depends not only on algorithm selection but also on the breadth, quality, and representativeness of the data used for development. A model trained in one institution may encounter different patient populations, acquisition protocols, or clinical workflows when transferred elsewhere. Published healthcare federations show that collaboration can extend the evidence available for training without requiring institutions to transfer their underlying records to one repository [1–3]. The resulting opportunity is substantial, but it should not be confused with a guarantee of safe or equitable clinical use. Centralized collaboration usually requires a common repository containing data from participating sites. Such an arrangement can simplify optimization and analysis, but it also concentrates custody of

sensitive information. Federated learning (FL) changes the training workflow: institutions receive a model, optimize it using local data, and return numerical updates for aggregation [4, 5]. This distinction is particularly relevant to cross-silo healthcare collaborations, in which a relatively small consortium of identifiable hospitals or research institutions participates in repeated training rounds. The privacy argument for FL must nevertheless be qualified. Keeping raw data on site reduces direct record transfer; it does not make model updates intrinsically anonymous. Gradient reconstruction and membership-inference research demonstrate that learned information can expose aspects of training data under particular adversarial conditions [6, 7]. An institution that sends an unprotected update may therefore disclose information even though it never uploads a patient file. The trained model, diagnostic logs, and evaluation reports also require careful handling. Collaboration presents a second challenge that cannot be resolved by cryptography alone. Hospitals must agree on eligibility criteria, clinical endpoints, preprocessing, contribution rules, intellectual property, and responsibility for model re lease. Their data may differ in sample size, disease prevalence, measurement conventions, and access to care. A common model can be useful only if these differences are understood and its performance is examined at the site and subgroup levels, not merely through a favorable consortium-wide average.

This paper addresses four research questions: (1) it synthesizes the privacy and collaboration benefits of cross-silo FML while identifying where common privacy claims are too broad; (2) it proposes a defense-in-depth architecture joining local governance, secure aggregation, differential privacy, robust update screening, and lifecycle monitoring; (3) it maps the architecture to medical imaging, electronic health records (EHRs), genomics, public-health surveillance, and connected-health applications; and (4) it presents a reproducible synthetic five-hospital experiment that illustrates utility, cross-site generalization, and the effect of clipped noisy updates without presenting synthetic evidence as clinical validation.

The remainder of the paper reviews related work, formulates the learning problem, develops a layered privacy design, presents the proposed workflow, and evaluates a bounded synthetic example. Limitations and implementation recommendations distinguish what is supported by published clinical studies, what is proposed here, and what is measured only in simulation.

II. RELATED WORK

A. Federated Optimization

McMahan et al. introduced federated averaging (FedAvg), which interleaves local model optimization with weighted server aggregation [4]. The approach supplies the optimization baseline used in this paper. Its applicability to healthcare does not imply that the distributions or operational constraints of hospitals resemble those of mobile devices. In particular, a small hospital consortium may offer more reliable participation but weaker anonymity from aggregation than a large population of device clients.

Statistical and systems heterogeneity motivated FedProx, which adds a proximal term to the local objective and accommodates differing local workloads [8]. Such methods are relevant when local optimization produces substantial client drift. They are alternatives for a future clinical implementation, rather than algorithms evaluated in the present simulation. Choosing among global, personalized, and proximal methods should follow an explicit validation protocol instead of an assumption that one algorithm dominates across all clinical tasks.

B. Evidence from Healthcare Collaborations

Sheller et al. evaluated federated medical-image learning across ten institutions and reported performance approaching the centralized comparator for their segmentation setting [1]. Dayan et al. developed the EXAM model using data from twenty institutions to predict oxygen requirements from clinical variables and chest radiographs [2]. Pati et al. extended glioblastoma segmentation collaboration to seventy-one sites and 6,314 cases [3]. These studies establish task-specific feasibility at different scales; their metrics should not be merged into a single estimate of the benefit of FL.

The FLamby benchmark provides seven healthcare datasets with natural site partitions and accompanying evaluation code [9]. Its emphasis on realistic splits motivates the use of heterogeneous synthetic sites here and the recommendation that later empirical work use genuine institutional partitions rather than arbitrary random shards. Table 1 summarizes the scope of the selected evidence without ranking studies that use incompatible tasks or metrics.

C. Privacy-Enhancing Technologies

Secure aggregation permits a coordinator to obtain an aggregate while hiding individual submitted vectors, subject to protocol assumptions about adversaries and dropout [10]. Differential privacy (DP) instead limits how much a released result can change when the protected unit changes [11]. DP training and privacy accounting were developed for machine learning by Abadi et al. [12]. These techniques address different exposures and can complement one another; neither should be described as an interchangeable synonym for FL. In medical imaging, Kaissis et al. explain how federated, cryptographic, and privacy-preserving methods can be combined [13]. Ziller et al. examine the relationship between medical-model utility and privacy risks under different threat assumptions [14]. This work supports explicit reporting of the protected unit, privacy parameters, and attacker capabilities. It does not justify interpreting an arbitrary amount of injected noise as a meaningful patient-level guarantee.

D. Research Gap

The selected literature establishes useful building blocks, but a healthcare consortium must still translate them into a coherent operational process. The present paper therefore links optimization, confidentiality, participant accountability, clinical evaluation, and benefit sharing. Its synthetic component makes the numerical illustration auditable, while its limitation statements prevent the illustration from being mistaken for evidence from an unperformed patient study.

Table 1. Summarizes the scope of the selected evidence without ranking studies that use incompatible tasks or metrics

Study	Setting	Contribution	Interpretation and boundary
Sheller et al. (2020) [1]	Medical-image segmentation; ten institutions	Multi-institutional FL evaluation	Demonstrates task-specific model learning without record pooling; not a general formal privacy guarantee.
Dayan et al. (2021) [2]	COVID-19 oxygen requirement prediction; twenty institutions	Real-world multimodal collaboration	Supports clinical prediction feasibility; reported outcomes remain specific to the study cohorts and task.
Pati et al. (2022) [3]	Glioblastoma segmentation; seventy-one sites, 6,314 cases	Large cross-silo federation	Shows that geographically distributed institutions can contribute to a common imaging task
Ogier du Terrail et al. (2022) [9]	Seven healthcare benchmark datasets with natural site splits	Reproducible cross silo evaluation	Enables comparison under realistic partitions; benchmark performance is not deployment approval
Ziller et al. (2024) [10]	Medical-imaging privacy and reconstruction analysis	Privacy–utility evaluation	Shows that conclusions depend on the privacy budget and threat assumptions; empirical attack failure is not a universal guarantee.
Present paper	Six synthetic hospitals; logistic regression	Architecture and reproducible illustration	No patient records, cryptographic implementation, privacy-budget calibration, or clinical validation

III. MODEL AND PRIVACY REQUIREMENTS

A. Protected Assets and Adversaries

The protected assets include patient records, rare diagnoses, institutional case mix, model parameters, credentials, audit trails, and clinically sensitive outputs. Relevant adversaries include: (1) an honest-but-curious coordinator that follows the protocol but attempts to infer information from updates; (2) a malicious or compromised participant that submits poisoned updates; (3) an external network attacker; and (4) an authorized user who misuses model access. Collusion must also be considered because a coordinator and a subset of clients may combine their observations. FL reduces the exposure caused by centralized raw-data storage but does not eliminate leakage. Gradient inversion can reconstruct features from shared gradients under some conditions [6]. Membership inference can test whether a record contributed to training, and property inference can expose attributes of a site's population. Conversely, poisoned or backdoored updates can reduce accuracy for a subgroup while leaving global performance apparently acceptable.

B. Security and Privacy Objectives

The design adopts the following objectives:

- Data minimization: only task-relevant features and model updates are processed.
- Confidentiality in transit and computation: mutual authentication, encrypted channels, and secure aggregation protect updates.
- Bounded disclosure: clipping, DP accounting, output controls, and query limits reduce inference risk.
- Integrity and availability: signed software, update validation, rollback, and quorum rules resist malicious or faulty clients.
- Accountability: immutable or tamper-evident logs record approvals, code versions, participants, privacy budgets, and model releases.
- Clinical safety and equity: site-level and subgroup performance are validated before release and monitored after deployment.

IV. PROPOSED PRIVACY-AWARE HEALTHCARE FRAMEWORK

Figure 1 presents cross-silo architecture with three data owning hospitals for clarity; the same pattern extends to larger networks. The framework separates the local data plane, federated learning plane, consortium control plane, and clinical assurance plane.

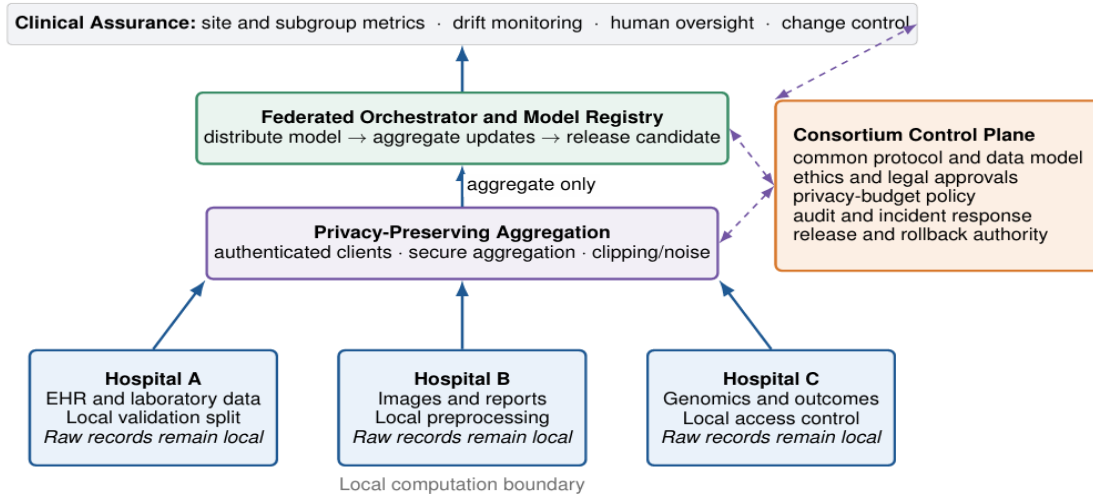


Figure 1. Proposed cross-silo healthcare FML architecture

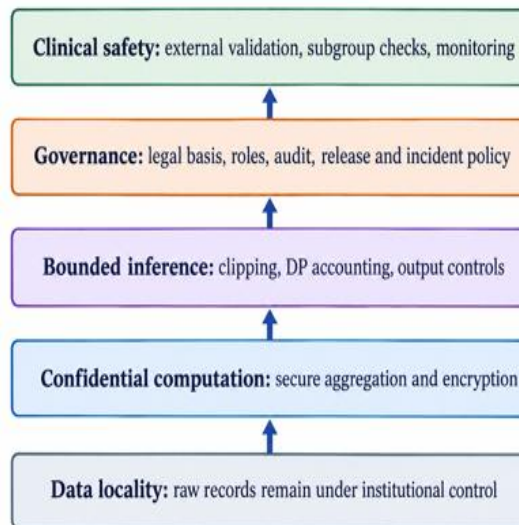


Figure 2. Defense-in-depth model

A. Consortium Formation and Governance

Before training, participants must define the clinical question, intended use, protected unit, permitted sites, and decision authority. A machine-readable study manifest should include model architecture, feature and label definitions, inclusion/exclusion criteria, optimizer and stopping criteria, minimum client quorum, allowed software hashes, privacy parameters, evaluation metrics, and release thresholds. Contracts must clarify the roles of controllers, processors, custodians, and service providers. Governance defines information sharing policies, allowing model updates while prohibiting per-site metrics below a certain sample size. Counts of rare conditions should be suppressed or combined to protect institutional case mix. Consortium members should agree on benefit

sharing, publication rights, intellectual property, model withdrawal, and the treatment of participants who leave the federation.

B. Local Data and Interoperability Layer

Hospitals map local data to a common specification while retaining source records. EHR terminology mappings include lab identifiers, diagnosis and procedure vocabularies, medication concepts, and indicators for missing data. Imaging protocols should specify orientation, voxel spacing, and annotation standards. Local quality checks produce privacy-safe summaries of schema conformance. Preprocessing code must be containerized, signed, versioned, and run in a controlled environment. A local holdout set should not be included in training. Where allowed, sites can

report metrics like confidence intervals and specific subgroup performance, as a global average can mask harm to smaller hospitals or patient subgroups.

C. Federated Training and Privacy Plane

In each round, the orchestrator shares an approved model and nonce with authenticated sites. Local workers verify signatures, train within set budgets, clip updates, and encrypt them for secure aggregation. The server requires a minimum number of participants to learn the aggregate update without accessing individual contributions. Differential privacy can be applied either centrally to the secure aggregate or locally before transmission; central DP generally offers better utility but relies on secure aggregation for privacy. The unit of protection must be defined, with patient-level adjacency often being more meaningful in longitudinal medicine. Before a release candidate is created, robustness checks are conducted, including update validation and anomaly detection. Flagged updates shouldn't automatically be deemed malicious, as unusual gradients may reflect different patient populations; investigations should combine statistical evidence with site context.

D. Clinical Assurance and Model Lifecycle

A release candidate is evaluated at every site and, where possible, at institutions that did not participate in training. Required outputs include discrimination, calibration, decision curve analysis, subgroup uncertainty, and failure-mode review. Model cards outline the intended population, exclusions, privacy mechanisms, participating regions, update frequency, and limitations. Deployment is considered a controlled clinical change, with the registry documenting the approved version, signatures, privacy ledger, validation report, and rollback target. Monitoring checks for input drift, missingness, calibration, alert burden, and subgroup performance. Retraining needs renewed approval for changes in data definition, clinical workflow, model architecture, privacy budget, or intended use.

V. METHODOLOGY

A. Objective and Dataset Construction

The experiment asks whether a shared model can improve cross-site generalization under statistical heterogeneity while avoiding patient-level transfer. It

is an engineering illustration, not a clinical trial. A fixed pseudorandom seed of 20261004 generated 2,500 synthetic patient records distributed across five hospitals with 380, 460, 520, 440, and 700 records. Each record contained 12 continuous features. Hospital-specific mean shifts, feature scaling, and outcome intercepts created differences in covariates and disease prevalence. Labels were sampled from a logistic data-generating process with one nonlinear interaction. Each hospital independently split its data into 78% training and 22% testing and standardized features using training-set statistics only. Table 2 summarizes the resulting partitions. No record or feature vector was exchanged during federated training.

Table 2. Synthetic Multi-Hospital Cohort

Site	Total	Train	Test	Test positive rate
Hospital 1	380	296	84	0.131
Hospital 2	460	358	102	0.431
Hospital 3	520	405	115	0.426
Hospital 4	440	343	97	0.680
Hospital 5	700	546	154	0.448
Total	2,500	1,948	552	

B. Compared Training Strategies

All strategies used logistic regression with L_2 regularization:

- 1) Local-only: each hospital trained on its own records for 140 full-batch epochs.
- 2) Centralized: locally standardized training partitions were pooled and optimized for 180 epochs. This is an upper reference workflow, not a privacy-preserving deployment.
- 3) FedAvg: all five hospitals participated in 40 rounds, with two local epochs per round and sample-size-weighted aggregation.
- 4) Clipped-and-noised FedAvg: each local update was clipped to $C = 0.12$ and Gaussian perturbation with coordinate standard deviation $\sigma C/K$ was added to the aggregate using $\sigma = 0.10$.

The fourth strategy is intentionally described as clipped-and noised, not formally differentially private.

A production DP claim would require a specified sampling model, adjacency relation, accountant, δ , and cumulative ϵ . Omitting these details and calling any noisy update “DP” would overstate the guarantee.

C. Evaluation

AUROC was calculated independently at each hospital, and the macro-AUROC is the unweighted average across hospitals. Macro-averaging prevents

the largest hospital from hiding poor performance elsewhere. To measure transportability, each local only model was also evaluated on the other four hospitals; this mean external-site AUROC was compared with the global federated model over the same external test sets. Binary cross-entropy on the five held-out sets was recorded across communication rounds.

Table 3. Held-Out AUROC by Hospital

Training strategy	Hospital1	Hospital2	Hospital3	Hospital4	Hospital5	Macro-AUROC
Local-only (each site’s own model)	0.828	0.869	0.828	0.813	0.839	0.835
Centralized reference	0.848	0.872	0.831	0.832	0.823	0.841
Federated averaging	0.848	0.872	0.831	0.833	0.824	0.841
Clipped-and-noised FedAvg	0.849	0.871	0.828	0.832	0.824	0.841

VI. RESULTS AND ANALYSIS

A. Predictive Performance

Table 3 shows that centralized learning, FedAvg, and the clipped-and-noised variant all reached a macro-AUROC of 0.841 to three decimal places. The average local-only in-site AUROC was 0.835. The aggregate result therefore demonstrates that, for this modest convex task, the federated optimizer recovered centralized utility without transmitting patient-level records. Performance was not uniformly improved. Hospital 5 obtained AUROC 0.839 from its own model and 0.824 from FedAvg. This is a useful warning: global utility can conceal local regression. A consortium should define minimum site and subgroup thresholds, investigate such regressions, and consider personalization, reweighting, or multi-task learning when one shared model is inappropriate.

B. Convergence and External Generalization

Figure 3 shows held-out macro log loss decreasing from 0.673 after the first round to 0.529 after 40 rounds for both federated strategies. In this low-dimensional simulation, the selected perturbation was small relative to the aggregate signal. This should not be generalized to deep networks, smaller client populations, repeated releases, or stronger privacy budgets. Figure 4 compares the external generalization

of a model trained at one hospital with the shared model. The five local models achieved external-site AUROCs of 0.786, 0.831, 0.835, 0.828, and 0.823. The corresponding federated evaluations ranged from 0.834 to 0.846, yielding an average increase from 0.821 to 0.842. The largest gain occurred for Hospital 1, whose local model was trained on the smallest number of positive cases.

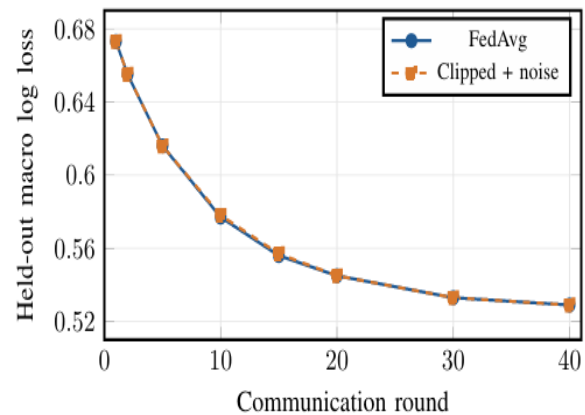


Figure 3. Convergence on held-out synthetic data

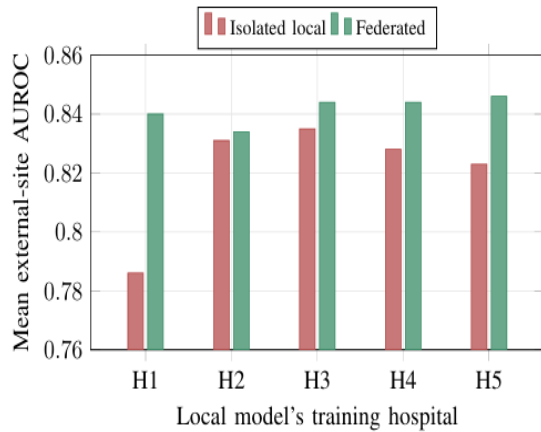


Figure 4. Generalization to hospitals not used by each isolated model

C. Privacy Interpretation

The experiment demonstrates data minimization, not an end-to-end privacy proof. FedAvg alone would expose plaintext updates to the coordinator. A real system should combine secure aggregation with authenticated execution and explicitly accounted DP where the threat analysis requires it. The utility retained by the noisy simulation merely shows that this task had room for perturbation; it does not establish a universal privacy-utility tradeoff. Figure 2 summarizes the correct interpretation. Data locality narrows the attack surface created by raw-data pooling. Confidential computation reduces coordinator visibility. DP bounds disclosure from releases. Governance constrains who may run a task and why. Clinical assurance determines whether the resulting model is safe enough for its intended use. Removing any one layer changes the residual risk.

VII. CONCLUSION

Federated machine learning can expand healthcare collaboration by moving computation toward data rather than routinely moving patient records toward computation. Evidence from multi-institutional research shows that federated models can approach centralized performance and improve generalization, while the illustrative experiment in this paper demonstrates the same pattern under transparent synthetic heterogeneity. The privacy benefit, however, must be stated precisely. FL provides data locality; secure aggregation protects individual updates from the coordinator; differential privacy bounds information leakage from releases; governance

controls legitimate use; and clinical assurance manages patient risk. The most credible path forward is therefore not an FL algorithm in isolation but a governed learning system. Institutions should begin with a clearly bounded clinical question, harmonize semantics, validate locally, protect updates cryptographically, account for disclosure statistically, and monitor released models across sites and subgroups. When these technical and organizational safeguards are implemented together, federated machine learning can support a more collaborative learning health ecosystem without requiring privacy and institutional autonomy to be traded away for scale.

REFERENCES

- [1] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, Art. no. 12598, 2020. [Nature](#)
- [2] I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Med.*, vol. 27, no. 10, pp. 1735–1743, 2021. [Nature](#)
- [3] S. Pati, U. Baid, B. Edwards, M. Sheller, S. H. Wang, G. A. Reina, et al., "Federated learning enables big data for rare cancer boundary detection," *Nature Commun.*, vol. 13, no. 1, Art. no. 7346, 2022. [Nature](#)
- [4] McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Apr. 2017, pp. 1273–1282. [PMLR](#)
- [5] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, et al., "The future of digital health with federated learning," *NPJ Digit. Med.*, vol. 3, no. 1, Art. no. 119, 2020. [Nature](#)
- [6] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019. [NeurIPS](#)
- [7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks

against machine learning models,” in *Proc. 2017 IEEE Symp. Security Privacy (SP)*, May 2017, pp. 3–18. [IEEE](#)

- [8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” *Proc. Mach. Learn. Syst.*, vol. 2, pp. 429–450, 2020.
- [9] J. Ogier du Terrail, S. S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, et al., “FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings,” in *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 5315–5334, 2022. [NeurIPS](#)
- [10] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, et al., “Practical secure aggregation for privacy-preserving machine learning,” in *Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Security*, Oct. 2017, pp. 1175–1191. [ACM](#)
- [11] Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–487, 2014. [Foundations and Trends](#)
- [12] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Security*, Oct. 2016, pp. 308–318. [ACM](#)
- [13] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, 2020. [Nature](#)
- [14] Ziller, T. T. Mueller, S. Stieger, L. F. Feiner, J. Brandt, R. Braren, et al., “Reconciling privacy and accuracy in AI for medical imaging,” *Nature Mach. Intell.*, vol. 6, no. 7, pp. 764–774, 2024. [Nature](#)