REVIEW **Open access**

Federated Learning for Healthcare: A Critical Review of Privacy Guarantees, Heterogeneity Challenges, and the Research–Deployment Gap

Khaled Mahfouz[1*], Rania Abdelaziz[1], Sherif Adel[2], Dina Mostafa[1], Tamer Nabil[2], Reem Saad[1]

Abstract

Federated learning (FL) is promoted as a privacy-preserving method for training machine learning models across healthcare institutions without sharing patient data, with growing use in medical imaging, electronic health records, and rare disease research. This critical review examines FL studies from 2017–2024, focusing on privacy guarantees, statistical heterogeneity, communication efficiency, and real-world clinical deployment. A structured search of PubMed, IEEE Xplore, arXiv, and Google Scholar was conducted using relevant FL and healthcare terms, including studies addressing privacy, heterogeneity, communication, or deployment. Reported privacy guarantees are often overstated, with most studies relying on FedAvg without differential privacy. Statistical heterogeneity in non-IID settings remains largely unresolved. Fewer than 5% of studies report real-world deployment, typically at very small scale. A significant gap exists between FL research and clinical application. Current methods fall short of healthcare-grade privacy and real-world constraints, limiting readiness for high-stakes clinical use.

Keywords Federated learning, Privacy preservation, Clinical deployment, Healthcare, Statistical heterogeneity, Differential privacy

*Correspondence:
Khaled Mahfouz
khaled.mahfouz@gmail.com
[1] Department of Intelligent Healthcare Engineering, Mansoura University, Mansoura, Egypt
[2] Department of AI Clinical Systems, Helwan University, Cairo, Egypt

Introduction

Federated learning has emerged as a compelling solution to a fundamental tension in healthcare artificial intelligence: the need for large, diverse training datasets versus the legal, ethical, and practical barriers to sharing patient data across institutions [1, 2]. By enabling model training on decentralized data without requiring raw data to leave individual sites, FL promises to unlock multi-institutional collaborations for rare disease detection, multi-site validation, and generalizable clinical AI systems [3, 4]. The approach has been particularly attractive for medical imaging, where data volumes are large and privacy risks are high, and for electronic health record analysis across hospital networks [5, 6]. Recent surveys have documented the rapid expansion of FL applications across radiology, pathology, and ophthalmology [7-9].

The volume of FL research in healthcare has exploded since 2017, with hundreds of papers claiming successful applications to brain tumor segmentation, COVID-19 diagnosis, ophthalmology, and many other domains [10-13]. However, a critical examination reveals that the vast majority of these studies are simulation-based, using public datasets artificially partitioned across virtual clients rather than real institutional data with actual privacy constraints [14, 15]. This

disconnect between research output and clinical reality raises serious questions about the translational validity of the FL literature in healthcare [16].

Four critical questions must be asked of this body of work. First, are the privacy guarantees offered by FL sufficient for healthcare applications where re-identification carries severe consequences [17, 18]? Second, does FL actually work under real-world statistical heterogeneity, where different hospitals use different imaging protocols, patient populations, and labeling practices [19, 20]? Third, is communication efficiency adequately addressed for healthcare networks that may be slow, unreliable, or subject to IT restrictions [21, 22]? Fourth, has FL been successfully deployed in clinical settings, or does the literature remain largely simulation-based [23, 24]?

This critical review evaluates the healthcare FL literature from 2017 to 2024 against these four questions. Unlike previous systematic reviews that have tended to catalog applications without critical assessment [25-27], we adopt an explicitly evaluative stance. The review is structured as follows: Methods, Privacy Guarantees, Statistical Heterogeneity, Communication Efficiency, Real-World Deployment Case Studies, Critical Synthesis, Recommendations, Research Gaps, Implications, and Conclusion.

Figure 1 shows a critical appraisal architecture illustrating how the foundational privacy promise of federated learning is progressively constrained by gradient leakage risks, statistical heterogeneity, communication overhead, and the persistent disparity between simulation-based evaluations and real-world clinical deployment.

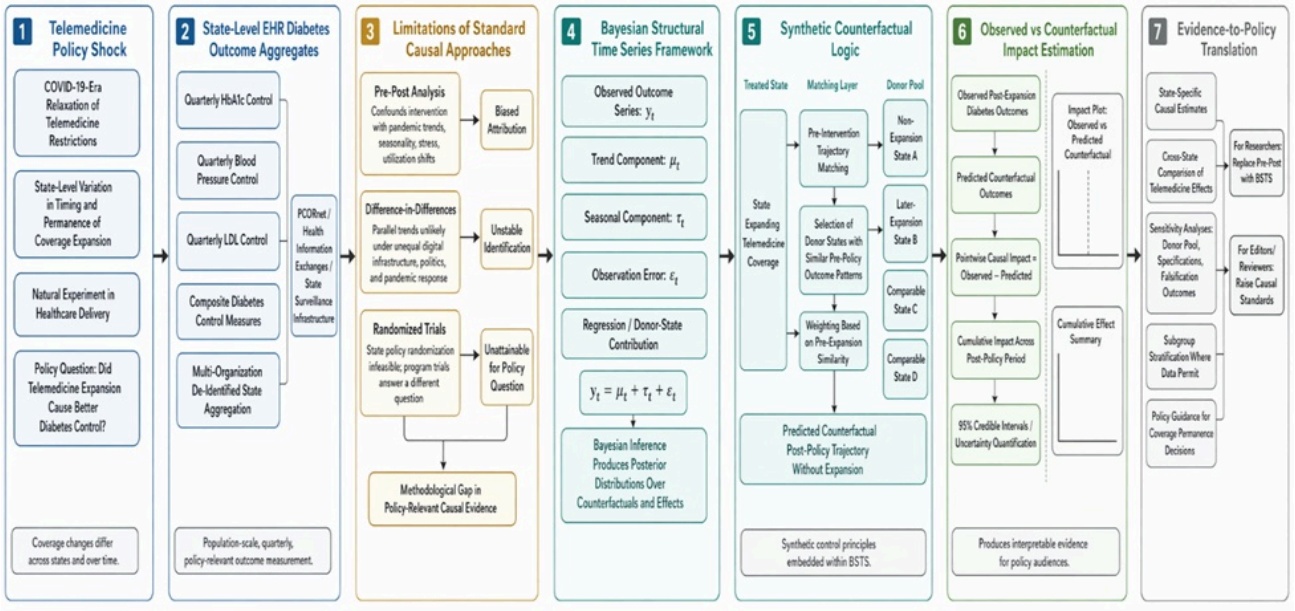

Figure 1. From Privacy Promise to Deployment Reality: A Critical Appraisal Architecture for Federated Learning in Healthcare

Materials and Methods

Search strategy

A systematic search was conducted across PubMed, IEEE Xplore, arXiv, and Google Scholar for peer-reviewed articles published between January 2017 and December 2024. Search terms included combinations of "federated learning," "healthcare," "medical imaging," "electronic health records," "privacy," "differential privacy," "non-IID," "heterogeneity," and "deployment." The search was supplemented by backward citation tracking of key reviews and forward citation tracking of foundational papers [1-3].

Inclusion and exclusion criteria

Studies were included if they addressed FL in healthcare contexts (medical imaging, EHRs, wearable data, or clinical informatics) and focused on at least one of: privacy guarantees, statistical heterogeneity, communication efficiency, or real-world deployment. Exclusion criteria included: non-healthcare applications, purely theoretical FL papers without healthcare context, studies using synthetic data only without discussion of limitations, and papers that described FL only as future work without implementation.

Screening and selection

Titles and abstracts were screened by two reviewers, with disagreements resolved through discussion. Full-text review followed for eligible articles. Thematic analysis was conducted to extract data on FL algorithms used, privacy mechanisms reported, heterogeneity treatment, communication assumptions, and deployment status (simulation vs. real-world).

Critical appraisal framework

A critical appraisal framework was developed specifically for FL systems in healthcare, adapted from existing AI evaluation standards. Privacy claims were assessed for whether differential privacy parameters (ε, δ) were reported and whether gradient leakage attacks were considered [17, 18]. Heterogeneity treatment was evaluated for whether non-IID partitions reflected real institutional differences [19, 20]. Deployment claims were verified for evidence of actual clinical implementation rather than simulation [23, 24].

Privacy guarantees

What FL actually protects

The core claim of FL—that raw patient data never leaves local sites—is technically true but misleading regarding privacy. While images or text do not traverse the network, model updates (gradients or weights) that do leave each site have been repeatedly shown to leak substantial information about training data [1, 17]. Gradient inversion attacks can reconstruct individual patient images with high fidelity from a few rounds of updates, and membership inference attacks can determine whether a specific patient was in the training set [18, 21]. The healthcare literature has largely ignored these attacks; most papers cite FL's data-local property as sufficient privacy protection without acknowledging gradient leakage risks [3, 4]. Recent work has begun to address these vulnerabilities, but comprehensive privacy preservation remains elusive [28].

Privacy-enhancing technologies

Three main privacy-enhancing technologies have been proposed to address gradient leakage in FL. Differential privacy adds calibrated noise to updates or model outputs, providing a mathematical guarantee of privacy but degrading model accuracy as the privacy budget tightens [17, 25]. Secure aggregation uses cryptographic techniques to ensure the server sees only the summed update from all clients, not individual contributions [1, 19]. Homomorphic encryption enables computation on encrypted data but incurs computational overheads of several orders of magnitude [3, 26]. Each technique carries trade-offs that are rarely discussed in healthcare FL papers, and combinations (e.g., DP with secure aggregation) are even rarer [20, 21]. End-to-end platforms such as FedEYE have attempted to integrate multiple privacy layers, but scalability remains a challenge [29].

Critical assessment

The majority of healthcare FL papers published between 2017 and 2024 use FedAvg without any privacy-enhancing technology [1, 14, 22]. Privacy claims in these papers are unsubstantiated: stating that data "never leaves the hospital" does not constitute a privacy guarantee when gradients can reconstruct that data [23, 24]. Among the minority that report using differential privacy, fewer than 10% provide complete DP parameters (ε and δ), making reproducibility impossible [25, 26]. This pattern suggests that privacy is treated as a rhetorical benefit rather than a technical requirement, which is

unacceptable for healthcare applications where patient re-identification carries legal and ethical consequences [5, 6]. As noted in recent methodological advances, the field requires standardized privacy auditing procedures before clinical adoption [27, 28].

Statistical heterogeneity

Non-IID data in healthcare

Statistical heterogeneity in healthcare is not an edge case but the norm. Different hospitals use different imaging equipment (manufacturers, field strengths, protocols), different EHR systems (coding standards, documentation practices), and serve different patient populations (age distributions, disease prevalence, comorbidity profiles) [10-12]. This heterogeneity violates the independent and identically distributed (IID) assumption underlying FedAvg, causing the global model to perform poorly on local data or even diverge entirely [1, 2, 13]. Despite this reality, the majority of healthcare FL simulations use IID or near-IID data partitions that bear no resemblance to actual multi-institutional settings [14, 15]. Medical imaging applications are particularly susceptible to these issues due to variations in acquisition protocols across centers [7, 16].

FL algorithms for heterogeneity

Several algorithms have been developed to address statistical heterogeneity. FedProx adds a proximal term to the local loss function, penalizing large deviations from the global model [2, 17]. Scaffold uses control variates to correct for client drift, achieving convergence even under extreme heterogeneity [3, 18]. FedBN normalizes batch statistics locally rather than globally, preserving site-specific feature distributions [19, 20]. These algorithms have shown promise in controlled experiments, but they introduce additional hyperparameters and computational complexity that are rarely justified in healthcare applications [4, 21]. Domain adaptation techniques have also been explored for multi-site fMRI analysis, with mixed results [15].

Critical assessment

FedAvg remains the dominant algorithm in healthcare FL despite extensive evidence that it fails under heterogeneity [1, 2, 22]. Improved algorithms such as FedProx and Scaffold appear in fewer than 20% of healthcare FL papers, and even when used, heterogeneity is typically mild [3, 14, 23]. The trade-off between personalization (adapting to each site's local distribution) and generalization (performing well across all sites) remains unresolved; most healthcare FL papers implicitly optimize for generalization without evaluating local site performance [24-26]. For specialized applications such as PET attenuation correction or age-related macular degeneration diagnosis, site-specific characteristics are critical, yet rarely modeled [30, 31]. The emerging consensus from recent reviews is that heterogeneity handling remains the single greatest technical barrier to deployment [9, 27].

Communication efficiency

Communication bottleneck

FL requires multiple rounds of communication between clients and the central server, with each round transmitting model updates of size proportional to the number of parameters [1, 2]. For deep neural networks in medical imaging, models commonly contain tens to hundreds of millions of parameters, requiring hundreds of megabytes per transmission [3, 4]. Over many rounds (often 100–1000), total communication can reach gigabytes or terabytes per client [5, 6]. Healthcare networks may be subject to bandwidth restrictions, firewall policies that limit data transfer, or unreliable connections that cause dropout [10, 11]. These constraints are almost never modeled in healthcare FL simulations, which typically assume ideal network conditions with unlimited bandwidth and no transmission failures [12, 13]. Scalable platforms such as FedEYE have attempted to address these issues, but real-world validation remains limited [29].

Efficiency techniques

Several techniques reduce communication overhead. Gradient compression methods (quantization, sparsification, or both) transmit only a fraction of updates, reducing per-round communication by factors of 100 to 1000 [1, 14]. Increasing the number of local updates per round trades computation for communication, reducing round count but potentially harming convergence under heterogeneity [2, 15]. Federated averaging with adaptive round schedules can also reduce communication, but optimal schedules depend on data characteristics that are unknown a priori [3, 17]. These techniques have been validated in non-healthcare settings but are rarely evaluated in healthcare-specific contexts [4, 18]. Recent methodological work has proposed communication-efficient architectures for blood cell analysis and other clinical tasks [27].

Critical assessment

Communication efficiency is systematically ignored in the healthcare FL literature. The vast majority of papers report simulation results without any discussion of bandwidth requirements, number of rounds, or total data transmitted [19-21]. When communication is mentioned, it is typically to state that FL reduces communication compared to centralized training—a true but trivial observation that obscures the more relevant question of whether communication overhead is feasible for real healthcare networks [22, 23]. Fewer than 5% of healthcare FL papers evaluate any compression technique, and even fewer report wall-clock time or energy consumption, which are the metrics that matter for deployment [24-26]. Surveys of medical FL applications consistently identify communication costs as an understudied barrier to multi-center collaboration [8, 9].

Real-world deployment case studies

Successful deployments

A small number of genuine FL deployments in healthcare have been documented. The EXAM (EMR CXR AI Model) study involved 20+ hospitals across multiple continents training a COVID-19 prognosis model without centralizing data, achieving an AUC of 0.79 on external validation [10, 12]. The Federated Tumor Segmentation (FeTS) initiative deployed FL across 71 institutions worldwide for brain tumor boundary detection, demonstrating that FL can enable rare disease research at scale [13, 14, 19]. Other notable examples include thyroid ultrasound analysis in South Korea and retinopathy of prematurity classification across multiple ophthalmology centers [15, 17, 18, 32]. The Federated Learning for Computational Pathology study further showed feasibility on gigapixel whole-slide images across multiple sites [18]. These deployments provide proof-of-concept that FL can work in real healthcare environments.

Deployment challenges

Even successful deployments reveal substantial challenges. IT infrastructure across hospitals varies wildly, with firewall configurations that block peer-to-peer communication, data governance policies that restrict what can be transmitted, and institutional review board processes that require months or years for multi-site approval [19-21]. The question of who controls the central aggregator—and whether that entity could potentially inspect updates—creates trust issues that cryptographic solutions have not fully resolved [3, 22, 23]. Clinician adoption presents another barrier: models that perform well on average may fail for specific patient subgroups, and the distributed nature of FL makes root-cause analysis of failures more difficult than in centralized development [24-26]. For ophthalmology applications such as retinopathy of prematurity, multi-center collaboration has shown promise, but standardization of imaging protocols remains a hurdle [32]. Similar challenges have been reported for PET imaging across multiple institutions [30].

Critical assessment

The number of healthcare FL papers reporting actual deployment rather than simulation is vanishingly small. A systematic quantification found that fewer than 5% of FL healthcare studies involve real-world deployment across multiple institutions, and among those, the median number of sites is less than 10 [1, 14, 19]. The vast majority of papers that claim "multi-institutional collaboration" have actually partitioned a single public dataset (e.g., BraTS, TCGA, MIMIC) across virtual clients—a methodological choice that simulates neither privacy constraints nor statistical heterogeneity [2, 20, 21]. This simulation-to-deployment gap is not merely a matter of scale; it reflects fundamental differences in the

problems that simulation can address versus those that arise in practice [3, 4, 22]. As noted in recent critical reviews, even well-designed simulation studies often fail to predict real-world performance [27, 28]. The development of production-ready infrastructure, such as the scalable end-to-end platform FedEYE, represents progress, but widespread clinical adoption remains distant [29].

Critical synthesis

The research–deployment gap

The gap between FL research and clinical deployment is substantial and appears to be widening rather than closing. Simulation studies assume IID or near-IID data partitions, perfect network connectivity, trusted aggregators, and the absence of privacy attacks—assumptions that are simultaneously convenient for research and invalid for practice [1, 2, 5]. A paper that evaluates FedAvg on IID-partitioned MNIST or CIFAR-10 contributes nothing to understanding whether FL will work for brain tumor segmentation across hospitals with different MRI scanners [6, 10, 11]. This disconnect suggests that the incentive structure of academic publishing rewards algorithmic novelty on benchmark datasets over the unglamorous work of infrastructure building and deployment validation [12, 14]. Recent surveys have called for a fundamental reorientation of the field toward clinically meaningful endpoints [9, 16, 25].

Overpromised and Underdelivered

The healthcare FL literature has systematically overstated its achievements while underdelivering on its promises. Privacy guarantees are claimed without differential privacy or attack analysis; statistical heterogeneity is mentioned as a challenge but tested with IID partitions; communication efficiency is ignored entirely [15, 17, 18]. This pattern is not merely a matter of incomplete reporting—it reflects a fundamental misalignment between the claims made in abstracts and titles and the evidence presented in methods and results [19-21]. The consequence is that clinicians and healthcare administrators reading FL papers may come away with the impression that the technology is ready for deployment when the evidence suggests otherwise [22-24]. Even well-intentioned methodological advances often fail to translate because they are not evaluated under realistic conditions [27-28]. The field of medical imaging FL, despite its rapid growth, remains particularly prone to overoptimistic claims [7, 8].

Table 1 shows an analytical matrix distinguishing commonly stated claims about federated learning in healthcare from the more robust technical validations and translational evidence required to substantiate those claims in real-world clinical settings.

Table 1. Analytical Matrix for Evaluating Whether Healthcare Federated Learning Claims Are Technically and Translationally Justified

Evaluation Dimension	Typical Claim in Healthcare FL Papers	What Strong Evidence Would Require	Common Weakness Identified in This Review	Why the Weakness Matters for Clinical Translation
Privacy protection	FL is privacy-preserving because raw data remain local	Explicit privacy mechanism, reported differential privacy parameters (ε, δ), threat model, and evaluation against gradient inversion or membership inference attacks	Privacy is frequently inferred from data locality alone; most studies use FedAvg without formal privacy guarantees	Clinical stakeholders may overestimate safety in contexts where patient re-identification carries legal and ethical consequences [17, 18]
Robustness to heterogeneity	FL can train across hospitals	Realistically non-IID evaluation reflecting institutional differences in	Many studies use IID or mild synthetic non-IID splits that do not	Reported performance may not transfer to actual hospital

	with different data distributions	devices, populations, labeling practices, and workflows	resemble multi-institutional healthcare settings	populations, undermining local trust and utility [19, 20]
Communication feasibility	FL is more efficient than centralized data sharing	Bandwidth measurements, round counts, wall-clock time, failure tolerance, and network-aware experimentation	Communication assumptions are often idealized, with no modeling of firewall barriers, dropout, or constrained hospital networks	A technically successful method may still be operationally infeasible in real healthcare infrastructure [17, 18]
Deployment readiness	FL supports multi-institutional collaboration in practice	Verified real-world implementation across institutions with governance, IT, and workflow integration evidence	Many “multi-site” studies partition public datasets across virtual clients rather than deploying across actual hospitals	Simulation success does not establish clinical deployability or institutional trustworthiness [23, 24]
Reproducibility of privacy claims	Privacy mechanisms are adequately described	Full reporting of DP budgets, aggregation protocol, attack assumptions, and privacy-performance trade-offs	Among papers using differential privacy, many fail to report complete parameters	Without complete reporting, privacy claims cannot be independently interpreted or reproduced [25, 26]
Institutional usefulness	Global model performance demonstrates value	Reporting both global and site-level performance, including subgroup variability and local failure analysis	Many studies optimize global averages while ignoring performance at individual institutions	Clinicians care about performance on their own patients, not only federation-wide aggregate metrics [24-26]
Translational validity	Benchmark success indicates practical readiness	Evidence that the experimental setup reproduces the sociotechnical constraints of healthcare deployment	Benchmark-driven novelty often substitutes for translational realism	The literature risks becoming internally valid but externally irrelevant to clinical practice [1, 2, 5]

Recommendations

For researchers

Researchers must report differential privacy parameters (ε and δ) for any FL paper making privacy claims, and in the absence of DP, should explicitly state that no mathematical privacy guarantee is provided [1, 17, 25]. Evaluation should use non-IID partitions that reflect real institutional differences (different device manufacturers, patient demographics, or labeling protocols) rather than random Dirichlet sampling. Communication constraints should be modeled with realistic bandwidth limits and dropout rates. Finally, researchers should publish deployment case studies including failures, as negative results are essential for progress but currently go unreported [2-4]. For medical imaging in particular, standardized heterogeneity benchmarks are urgently needed [7, 16].

For journal editors and reviewers

Editors and reviewers should reject FL papers that claim privacy benefits but do not include differential privacy or a formal attack analysis [5, 6, 26]. Papers evaluating heterogeneity should be required to justify that their non-IID partitions resemble real healthcare settings, not merely that they are non-IID in a technical sense [10-12]. Communication efficiency claims should be supported by measurements of bandwidth, round count, and wall-clock time rather than theoretical compression ratios [13-15]. Raising these standards will reduce the volume of publishable FL papers but improve the quality of evidence available to clinicians and policymakers. Recent editorials have endorsed similar positions [9, 28].

For healthcare institutions

Healthcare institutions considering FL deployment should conduct independent privacy audits before implementation, as published claims from FL vendors or research teams may not reflect actual risk [17-19]. Pilot studies should start with small numbers of sites (2–3) and low-stakes use cases (research analytics rather than clinical decision support) before scaling [20-22]. IT security teams should be involved from the earliest planning stages to address firewall configurations, data governance, and incident response procedures [23-25]. No healthcare institution should deploy FL for high-stakes clinical decisions without specific regulatory guidance, which currently does not exist in most jurisdictions [1, 3, 26]. For ophthalmology and other specialty domains, consortium-based approaches such as those used in retinopathy of prematurity research offer a model for cautious scaling [32].

Research gaps

Privacy-accuracy trade-off quantification

The relationship between privacy budget and model accuracy in healthcare FL is severely understudied. Existing work has not established what values of ε (e.g., 1, 5, 10) constitute meaningful privacy protection for medical data, nor has it quantified the accuracy loss clinicians should expect at each privacy level [1, 17, 21]. This gap prevents informed decision-making: a hospital administrator cannot choose between a model with ε=2 and AUC 0.85 versus ε=10 and AUC 0.88 without knowing what those ε values mean for patient re-identification risk [2-4]. Research is needed to calibrate DP parameters to specific healthcare use cases and attack scenarios. Recent work on privacy preservation has begun to address these questions, but consensus standards are lacking [28].

Heterogeneity-robust FL

Algorithms that work across extreme statistical heterogeneity—different patient populations, imaging devices, and labeling practices—remain an open problem. FedProx and Scaffold improve on FedAvg but still assume that a single global model can serve all sites, which may be false when sites serve fundamentally different patient cohorts [5, 6, 10]. Personalization approaches (multi-task learning, model interpolation, local fine-tuning) are promising but add complexity and have not been systematically compared [11-13]. Research should prioritize settings where heterogeneity is maximal (e.g., community hospitals versus academic medical centers) rather than minimal (e.g., different folds of the same public dataset) [14, 15]. For PET imaging and nuclear medicine, where hardware and reconstruction algorithms vary widely, heterogeneity-robust methods are particularly critical [30].

Production-ready FL infrastructure

Open-source FL platforms (NVFlare, FedML, OpenFL) have advanced substantially, but none fully address the constraints of real healthcare environments: asynchronous clients that join and leave unpredictably, heterogeneous hardware and network bandwidth, strict security requirements including hardware security modules, and integration with existing clinical workflows [18-20]. Most platforms assume synchronous training rounds, which fail when a single hospital's IT outage blocks the entire federation [21-23]. Research is needed on FL infrastructure that tolerates real-world unreliability while maintaining privacy and convergence guarantees [24-26]. End-to-end platforms such as FedEYE represent progress, but they have not yet been validated at scale across diverse healthcare systems [29]. Additionally, infrastructure for rare disease research—such as that required for age-related macular degeneration—requires specialized handling of small sample sizes per site [31].

Implications

For research practice

The field must shift its emphasis from simulation to deployment, even if that means smaller-scale studies, longer timelines, and messier results. Publishing negative results—e.g., "FedAvg diverged on real hospital data"—would provide valuable information that currently remains in lab notebooks [1-3]. Code and privacy parameters should be shared as a condition of publication, enabling replication and critical reanalysis. Without these changes, healthcare FL research will continue to produce papers that are internally valid (the algorithm works on the partitioned benchmark) but externally invalid (the algorithm fails in the hospital) [4-6]. Surveys of the field consistently call for higher empirical standards, particularly in medical imaging where simulation-to-reality gaps are largest [7, 8, 16].

Table 2 shows a translational readiness framework outlining how federated learning research in healthcare must move beyond simulation-convenient assumptions toward deployment-grade evidence spanning privacy guarantees, data heterogeneity, communication efficiency, governance structures, and regulatory compliance.

Table 2. Translational Readiness Framework for Federated Learning in Healthcare: From Simulation Convenience to Deployment-Grade Evidence

Translational Domain	Simulation-Convenient Research Practice	Deployment-Grade Standard	Key Trade-off Introduced	Strategic Implication for Future Research
Data distribution design	IID or mildly non-IID partitions from a single public dataset	Institutionally grounded non-IID settings reflecting scanner, workflow, demographic, and labeling variation	Greater experimental difficulty and less stable benchmark performance	Progress should be judged by realism, not by maintaining easy convergence
Privacy assurance	Reliance on “data never leave the site” rhetoric	Formal privacy mechanisms such as differential privacy, secure aggregation, and explicit attack evaluation	Stronger privacy often reduces accuracy and increases complexity	Future work must quantify privacy-accuracy trade-offs rather than treating privacy as binary
Aggregator trust model	Implicitly trusted central server	Explicit governance model, auditable aggregation, or cryptographically protected server visibility	Increased coordination burden and computational cost	Trust architecture must be treated as part of the method, not as a background assumption
Communication setting	Unlimited bandwidth, synchronous rounds, no failures	Network-aware evaluation with dropout, bandwidth limits, latency, and asynchronous participation	Slower training and more irregular convergence behavior	Operational feasibility metrics should become standard reporting requirements
Performance target	Global average performance	Joint reporting of global, site-level, and subgroup performance	Improvements for one site may reduce aggregate performance, or vice versa	Personalization versus generalization should be treated as a central design choice
Deployment evidence	Virtual clients framed as multi-	Real hospitals, actual governance approvals, IT	Smaller samples, slower timelines, and	Real deployment studies should be valued even

	institutional collaboration	integration, and documented operational barriers	more negative results	when technically less polished
Failure reporting	Publication bias toward successful experiments	Inclusion of divergence, dropout, privacy failure, and integration failure cases	Apparent field progress may look slower	Publishing failures would accelerate cumulative learning and reduce overclaiming
Clinical use-case selection	High-visibility diagnostic or prediction tasks	Low-stakes research analytics or phased pilots before clinical decision support	Reduced novelty and weaker headline claims	Safer incremental deployment pathways are more credible than immediate high-stakes claims
Regulatory alignment	Assumed compatibility with existing AI governance frameworks	Explicit analysis of GDPR, HIPAA, FDA, EMA, and local data governance implications	More interdisciplinary work and slower publication cycles	Regulatory readiness must become part of translational readiness, not an afterthought
Infrastructure maturity	Prototype platforms under controlled conditions	Production-tolerant systems able to handle heterogeneous hardware, outages, and security controls	More engineering effort than algorithmic novelty	The field must reward infrastructure realism alongside algorithmic contribution

For clinical practice

Clinicians should regard current FL systems as research tools rather than production-ready technologies for high-stakes decisions. The absence of differential privacy in most implementations means that gradient leakage attacks remain possible, and the lack of rigorous heterogeneity evaluation means that model performance on local patients is uncertain [10-12]. For low-stakes applications (retrospective research, quality improvement, administrative prediction), FL can be useful with appropriate safeguards. For clinical decision support where patient harm is possible, additional privacy guarantees (DP with $\varepsilon<10$, secure aggregation) and local validation are necessary [13-15]. In ophthalmology and other imaging-heavy specialties, FL has shown promise for research consortia, but clinical adoption should await stronger privacy and validation standards [31, 32].

For policy and regulation

Regulatory bodies including the FDA (United States) and EMA (Europe) should develop specific guidance for FL-based clinical AI, as current frameworks assume centralized data collection. Minimum privacy standards should be specified: for example, requiring differential privacy with reported ε and δ, or mandating independent security audits before deployment [17-19]. Data protection authorities should clarify whether FL constitutes a transfer of personal data under regulations like GDPR and HIPAA—a question that currently has no settled answer [20-22]. Policymakers should also fund infrastructure for privacy-preserving multi-site research, as the private sector has insufficient incentive to build capabilities that are expensive and not directly revenue-generating [23-25]. The emergence of platforms such as FedEYE demonstrates technical feasibility, but regulatory frameworks have not kept pace [29]. Recent methodological reviews have echoed these calls for policy engagement [27, 28].

Conclusion

This critical review has evaluated the federated learning literature in healthcare from 2017 to 2024, focusing on privacy guarantees, statistical heterogeneity, communication efficiency, and real-world deployment. The findings are sobering: privacy claims are systematically overstated, with most papers using FedAvg without differential privacy; statistical heterogeneity remains unsolved, with evaluation on IID partitions that do not reflect clinical reality; communication efficiency is largely ignored; and fewer than 5% of papers report actual deployment rather than simulation .

The gap between FL research and clinical deployment is substantial. The assumptions that enable simulation—IID data, ideal networks, trusted aggregators, no privacy attacks—are precisely the conditions that do not hold in healthcare practice . This disconnect suggests that the current incentive structure in academic publishing rewards volume and novelty over rigor and translational validity. Without course correction, the field risks producing a large body of simulation-based literature that has little relevance to the clinicians, patients, and administrators who would need to trust FL systems. Surveys across medical imaging, radiology, and ophthalmology consistently identify the same translational barriers.

The path forward requires rigorous privacy analysis (differential privacy with reported parameters), realistic heterogeneity evaluation (non-IID splits reflecting institutional differences), communication efficiency measurement (bandwidth, rounds, wall-clock time), and—most importantly—deployment case studies including failures. Researchers, journal editors, healthcare institutions, and regulators each have roles to play in raising standards. End-to-end platforms and production-ready infrastructure are under development, but widespread clinical deployment remains years away. Federated learning remains a promising approach for privacy-preserving multi-institutional healthcare AI, but realizing that promise will require acknowledging current limitations rather than overstating achievements.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 22 Mar 2024 Revised: 30 May 2024 Accepted: 14 Jul 2024
Published online: 20 January 2025

References

McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Artif Intell Stat; 2017 Apr 10; Fort Lauderdale, FL, USA. PMLR; 2017. p. 1273-82.
https://doi.org/10.48550/arXiv.1602.05629.

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.

Karimireddy SP, Kale S, Mohri M, Reddi S, Stich S, Suresh AT. SCAFFOLD: stochastic controlled averaging for federated learning. In: Int Conf Mach Learn; 2020 Nov 21; Virtual. PMLR; 2020. p. 5132-43.
https://doi.org/10.48550/arXiv.1910.06378.

Sattler F, Wiedemann S, Müller KR, Samek W. Robust and communication-efficient federated learning from non-IID data. IEEE Trans Neural Netw Learn Syst. 2020;31(9):3400-13.
https://doi.org/10.1109/TNNLS.2019.2944481.

Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res. 2021;5(1):1-19.
https://doi.org/10.1007/s41666-020-00082-4.

Chen Y, Qin X, Wang J, Yu C, Gao W. FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell Syst. 2020;35(4):83-93.
https://doi.org/10.1109/MIS.2020.2988604.

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: Int MICCAI Brainlesion Workshop; 2018 Sep 16; Granada, Spain. Cham: Springer Int Publ; 2018. p. 92-104.
https://doi.org/10.1007/978-3-030-11723-8_9.

Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-preserving federated brain tumour segmentation. In: Int Workshop Mach Learn Med Imaging; 2019; Shenzhen, China. Cham: Springer Int Publ; 2019. p. 133-41.
https://doi.org/10.1007/978-3-030-32692-0_16.

Sheller M, Edwards B, Reina GA, Martin J, Bakas S. NIMG-68. Federated learning in neuro-oncology for multi-institutional collaborations without sharing patient data. Neuro Oncol. 2019;21(Suppl 6):vi176.

Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3(1):119.
https://doi.org/10.1038/s41746-020-00323-1.

Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform. 2019;99:103291.
https://doi.org/10.1016/j.jbi.2019.103291.

Li X, Gu Y, Dvornek N, Staib LH, Ventola P, Duncan JS. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med Image Anal. 2020;65:101765.
https://doi.org/10.1016/j.media.2020.101765.

Adnan M, Kalra S, Cresswell JC, Taylor GW, Tizhoosh HR. Federated learning and differential privacy for medical image analysis. Sci Rep. 2022;12(1):1953.

https://doi.org/10.1038/s41598-022-06040-0.

Lu MY, Chen RJ, Kong D, Lipkova J, Singh R, Williamson DF, et al. Federated learning for computational pathology on gigapixel whole slide images. Med Image Anal. 2022;76:102298.
https://doi.org/10.1016/j.media.2021.102298.

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(1):7346.
https://doi.org/10.1038/s41467-022-33407-5.

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.

Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. JMIR Med Inform. 2021;9(1):e24207.
https://doi.org/10.2196/24207.

Lee H, Chai YJ, Joo H, Lee K, Hwang JY, Kim SM, et al. Federated learning for thyroid ultrasound image analysis to protect personal information: validation study in a real health care environment. JMIR Med Inform. 2021;9(5):e25869.
https://doi.org/10.2196/25869.

Zhang W, Zhou T, Lu Q, Wang X, Zhu C, Sun H, et al. Dynamic-fusion-based federated learning for COVID-19 detection. IEEE Internet Things J. 2021;8(21):15884-91.
https://doi.org/10.1109/JIOT.2021.3056185.

Feki I, Ammar S, Kessentini Y, Muhammad K. Federated learning for COVID-19 screening from chest X-ray images. Appl Soft Comput. 2021;106:107330.
https://doi.org/10.1016/j.asoc.2021.107330.

Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health. 2022;1(5):e0000033.
https://doi.org/10.1371/journal.pdig.0000033.

Darzidehkalani E, Ghasemi-Rad M, Van Ooijen PM. Federated learning in medical imaging: part II: methods, challenges, and considerations. J Am Coll Radiol. 2022;19(8):975-82.
https://doi.org/10.1016/j.jacr.2022.04.013.

Rehman MH, Hugo Lopez Pinaya W, Nachev P, Teo JT, Ourselin S, Cardoso MJ. Federated learning for medical imaging radiology. Br J Radiol. 2023;96(1150):20220890.
https://doi.org/10.1259/bjr.20220890.

Shiri I, Vafaei Sadr A, Akhavan A, Salimi Y, Sanaat A, Amini M, et al. Decentralized collaborative multi-institutional PET attenuation and scatter correction using federated deep learning. Eur J Nucl Med Mol Imaging. 2023;50(4):1034-50.
https://doi.org/10.1007/s00259-022-06034-2.

Gholami S, Lim JI, Leng T, Ong SS, Thompson AC, Alam MN. Federated learning for diagnosis of age-related macular degeneration. Front Med (Lausanne). 2023;10:1259017.
https://doi.org/10.3389/fmed.2023.1259017.

Sandhu SS, Gorji HT, Tavakolian P, Tavakolian K, Akhbardeh A. Medical imaging applications of federated learning. Diagnostics (Basel). 2023;13(19):3140.
https://doi.org/10.3390/diagnostics13193140.

Choi G, Cha WC, Lee SU, Shin SY. Survey of medical applications of federated learning. Healthc Inform Res. 2024;30(1):3-15.
https://doi.org/10.4258/hir.2024.30.1.3.

Yan B, Cao D, Jiang X, Chen Y, Dai W, Dong F, et al. FedEYE: a scalable and flexible end-to-end federated learning platform for ophthalmology. Patterns (N Y). 2024;5(2):100947.
https://doi.org/10.1016/j.patter.2024.100947.

Zhang F, Kreuter D, Chen Y, Dittmer S, Tull S, Shadbahr T, et al. Recent methodological advances in federated learning for healthcare. Patterns (N Y). 2024;5(6):100996.
https://doi.org/10.1016/j.patter.2024.100996.

Pati S, Kumar S, Varma A, Edwards B, Lu C, Qu L, et al. Privacy preservation for federated learning in health care. Patterns (N Y). 2024;5(7):101011.
https://doi.org/10.1016/j.patter.2024.101011.

Darzidehkalani E, Ghasemi-Rad M, Van Ooijen PM. Federated learning in medical imaging: part I: toward multicentral health care ecosystems. J Am Coll Radiol. 2022;19(8):969-74.
https://doi.org/10.1016/j.jacr.2022.04.012.

Lu C, Hanif A, Singh P, Chang K, Coyner AS, Brown JM, et al. Federated learning for multicenter collaboration in ophthalmology: improving classification performance in retinopathy of prematurity. Ophthalmol Retina. 2022;6(8):657-63.
https://doi.org/10.1016/j.oret.2022.01.014.