



## **Trustworthy Federated Learning for Medical Imaging: A Structured Survey of Heterogeneity, Privacy Leakage, Inference Attacks, and Secure Collaboration**

**Kazi Shaharair Sharif<sup>1</sup>; Mohammed Majbah Uddin<sup>2</sup>;**

- [1]. Department of Computer Science, Kennesaw State University, GA, USA;  
Email: [kazi.sharif@ieee.org](mailto:kazi.sharif@ieee.org)  
[2]. Department of Computer Science, University of Florida, Gainesville, FL, USA;  
Email: [uddinmajbah@ieee.org](mailto:uddinmajbah@ieee.org)

**Doi:** [10.63125/mrbsbt46](https://doi.org/10.63125/mrbsbt46)

**Received:** 11 April 2026; **Revised:** 21 May 2026; **Accepted:** 15 June 2026; **Published:** 08 July 2026;

### **Abstract**

Federated learning (FL) enables hospitals and imaging centers to train shared artificial-intelligence models while retaining raw patient data locally. This design is attractive for medical imaging because clinically useful models often require multi-institutional data, yet conventional pooling is constrained by privacy, governance, ownership, infrastructure, and regulatory requirements. Data locality, however, does not by itself establish trustworthiness. Federated systems can still fail under inter-hospital distribution shift, leak information through model updates or predictions, remain vulnerable to membership and reconstruction attacks, and incur substantial overhead when differential privacy or cryptographic defenses are added. This survey synthesizes the literature through five connected dimensions: predictive utility, heterogeneity robustness, formal privacy, empirical attack resistance, and system efficiency. It reviews foundations and architectures, application domains, non-IID optimization, privacy leakage, differential privacy, membership-inference auditing, secure aggregation, and cross-study evidence gaps. The central finding is that the field is comparatively mature within individual technical streams but weak at their intersection. Studies commonly demonstrate accuracy, non-IID robustness, privacy budgets, attacks, or cryptographic protection in isolation, while relatively few evaluate these properties under one consistent medical-imaging pipeline. The survey therefore proposes a compact five-dimensional taxonomy and a research agenda centered on patient-level integrity, realistic site heterogeneity, matched privacy baselines, strong empirical auditing, secure update aggregation, reproducible systems-cost measurement, and real multi-institutional validation.

### **Keywords**

Federated Learning, Medical Imaging, Trustworthy AI, Non-IID Data, Differential Privacy, Membership Inference, Secure Aggregation, Privacy-Preserving Machine Learning.

## **INTRODUCTION**

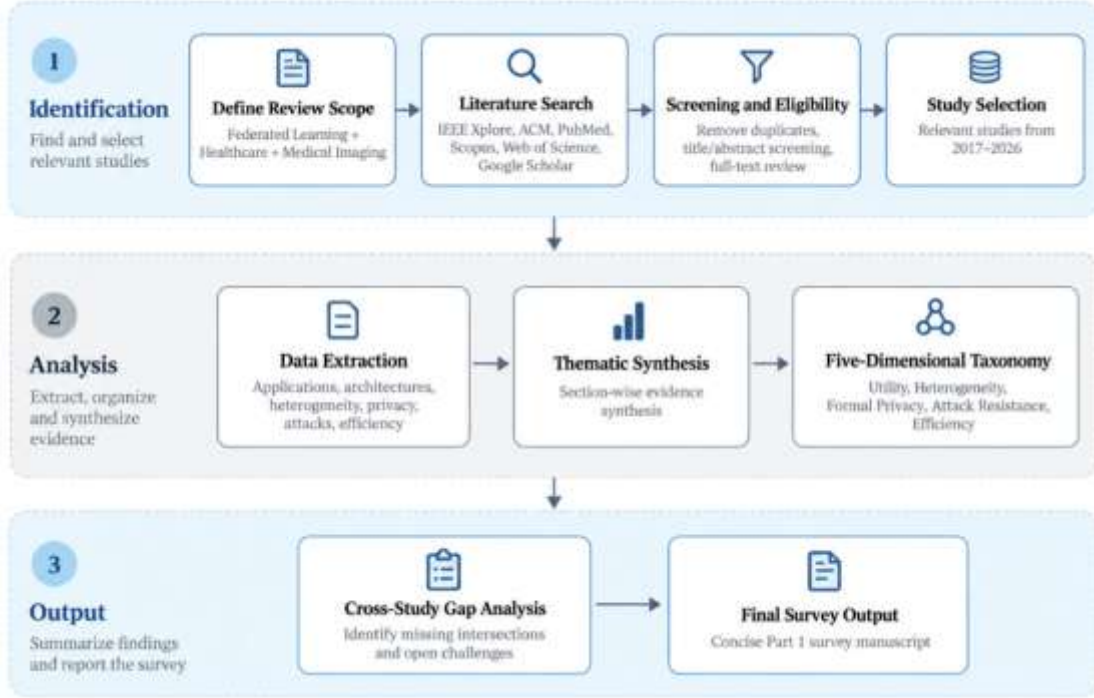
Deep learning has improved medical-image classification, segmentation, detection, prognosis, and decision support, but performance depends on large, diverse, representative datasets. Medical images are naturally distributed across hospitals, diagnostic centers, laboratories, and health systems. Centralizing those data can create legal, governance, ownership, cybersecurity, and institutional barriers, while isolated local training limits sample size and external generalization. Federated learning was developed to bridge these constraints by training a shared model across decentralized clients without requiring the physical pooling of raw records (McMahan et al., 2017; Rieke et al., 2020). Medical imaging is a particularly important FL setting because institutional variation is clinically meaningful. Hospitals differ in disease prevalence, demographics, referral patterns, scanner vendors, acquisition protocols, annotation practices, and sample volume. Landmark collaborations show that federation can support clinically relevant learning at scale. Sheller et al. (2020) demonstrated multi-institutional brain-tumor learning without direct patient-data sharing; Dayan et al. (2021) trained a COVID-19 outcome model across 20 institutions; and Pati et al. (2022) reported a 71-site glioblastoma collaboration spanning six continents. These studies establish the feasibility and scientific value of cross-silo medical FL.

The key limitation is conceptual: keeping images local is not equivalent to providing complete privacy or security. Gradients and model updates are derived from private training records and can leak information. The final model can also expose membership through its prediction behavior. At the same time, severe non-IID data can destabilize optimization, differential privacy can reduce utility, and cryptographic defenses can add communication and computation cost (Kaissis et al., 2020). Trustworthy medical FL must therefore be evaluated as a multi-dimensional systems problem rather than as an accuracy-only learning technique. This survey focuses on five evidence dimensions: predictive utility, robustness to heterogeneity, formal privacy, empirical attack resistance, and system efficiency. Its specific goal is to connect research streams that are often studied separately and to identify what evidence is still required before federated medical imaging can be treated as robust, auditable, and clinically deployable.

## **2. Structured Review Methodology**

The review follows a structured evidence-synthesis process inspired by PRISMA 2020 and PRISMA-S reporting principles (Page et al., 2021; Rethlefsen et al., 2021). The scope covers the intersection of federated learning, healthcare, and medical imaging, with emphasis on heterogeneity, privacy, attacks, security, and efficiency. The primary publication window is 2017 through 2026, reflecting the modern FedAvg era while allowing earlier foundational privacy or cryptographic work when necessary for technical context. Eligible studies must contribute substantively to federated medical imaging or to a privacy/security method directly needed to interpret medical FL. The information sources searched included IEEE Xplore, the ACM Digital Library, PubMed/MEDLINE, Scopus, Web of Science, ScienceDirect, and SpringerLink, with Google Scholar used for supplementary literature discovery and citation chaining. The search strategy combined three concept blocks covering federated learning, medical imaging, and trustworthiness. Federated-learning terms were combined with imaging-related terms, including MRI, CT, radiology, X-ray, mammography, pathology, histopathology, ultrasound, and ophthalmology, and with trustworthiness-related terms, including non-IID data, privacy, differential privacy, membership inference, gradient inversion, secure aggregation, homomorphic encryption, secure multiparty computation, and system overhead. Relevant studies identified through the database searches were supplemented through backward and forward citation tracing where appropriate to capture foundational and closely related work that informed the technical synthesis. Screening was performed through identification, deduplication, title/abstract review, and full-text assessment according to the stated eligibility criteria. The review was designed as a structured evidence synthesis rather than a numerical meta-analysis; accordingly, the selected literature was synthesized narratively and organized through the proposed taxonomy to compare evidence across predictive utility, heterogeneity robustness, formal privacy, empirical attack resistance, and system efficiency.

**Figure 1. Structured Survey Workflow for Trustworthy Federated Learning in Medical Imaging.**



For each study, the extraction framework records application domain, imaging modality, task, clients/sites, model and FL algorithm, type of heterogeneity, privacy mechanism, attack model, secure-computation mechanism, predictive metrics, efficiency metrics, and reproducibility details. A central methodological rule is that data locality alone is not coded as formal privacy, generic transport encryption is not coded as secure aggregation, and image-level privacy is not silently treated as patient-level privacy. These distinctions are essential for fair cross-study comparison.

### 3. Foundations and Architectures of Federated Medical Learning

Federated medical learning is an architecture for collaborative optimization over decentralized data, not a single algorithm. In the dominant cross-silo setting, hospitals retain their local imaging datasets, receive a global model from a coordinating server, train that model locally, and return model updates for aggregation. This differs from centralized learning, where raw data are pooled, and from isolated local learning, where institutions receive no benefit from one another's data. Cross-silo FL is particularly appropriate for hospitals because the clients are stable, identifiable organizations with comparatively reliable network and computing resources (Rieke et al., 2020).

The foundational algorithm is Federated Averaging, or FedAvg, which repeatedly distributes a common model, performs several local optimization steps, and combines participating client updates using data-dependent weights (McMahan et al., 2017). Its efficiency comes from allowing local computation between communication rounds. Its weakness is that local updates may diverge when institutional distributions differ. FedProx constrains local training so that client models do not move excessively far from the global model; SCAFFOLD explicitly corrects client drift; and FedBN retains batch-normalization behavior locally to better handle site-specific feature distributions (Li et al., 2020; Karimireddy et al., 2020; Li et al., 2021).

Medical FL can also be horizontal, vertical, transfer-based, personalized, clustered, or decentralized. Horizontal FL is most common in imaging: hospitals have similar feature spaces but different patients. Vertical FL is relevant when organizations hold different modalities or clinical features for overlapping patients, such as imaging at one institution and genomics at another. Personalized FL retains site-specific components when a single global model is not optimal for all institutions. Decentralized

architectures remove the permanent central aggregator, as illustrated by Swarm Learning (Warnat-Herresthal et al., 2021).

The architecture itself creates the trust boundary. The server may observe individual updates, client-specific behavior, timing, and metadata; peer institutions may infer information from global changes; and external users may query the final model. Consequently, the end-to-end design must cover local data integrity, harmonized preprocessing, client authentication, protected communication, aggregation, validation, audit logging, and deployment rather than treating model averaging as the entire system.

#### 4. Applications Across Medical Imaging Domains

Federated learning has been applied across neuroimaging, thoracic imaging, mammography, digital pathology, ophthalmology, abdominal imaging, and multimodal clinical prediction. The application literature demonstrates a consistent benefit: institutions can improve shared learning without surrendering direct custody of raw images. It also shows that each imaging domain stresses a different aspect of trustworthiness (Guan et al., 2024).

*Table 1. Representative applications of federated learning across major medical-imaging domains.*

| Medical-imaging domain | Representative task                                           | Typical modality                    | Representative study                      | Main contribution                                                                                 | Key trustworthy-FL challenge                                 |
|------------------------|---------------------------------------------------------------|-------------------------------------|-------------------------------------------|---------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| Neuroimaging           | Brain-tumor segmentation and classification                   | MRI                                 | Sheller et al. (2020); Pati et al. (2022) | Demonstrated multi-institutional collaboration without centralizing patient images                | Site heterogeneity, rare-disease data, large 3D models       |
| Thoracic imaging       | COVID-19 outcome prediction and disease classification        | Chest CT                            | X-ray, Dayan et al. (2021)                | Demonstrated large-scale multi-hospital federated prediction using imaging and clinical variables | Population shift, device variation, multimodal heterogeneity |
| Breast imaging         | Breast-density and cancer-related classification              | Mammography                         | Roth et al. (2020)                        | Improved cross-site generalization compared with isolated local models                            | Scanner/vendor shift, class imbalance                        |
| Digital pathology      | Cancer classification, segmentation, and molecular prediction | Histopathology, whole-slide imaging | Lu et al. (2022); Saldanha et al. (2022)  | Enabled privacy-aware collaboration on large pathology data                                       | Stain variation, scanner shift, computational burden         |

|                              |                                                         |                               |                     |                                                                       |                                                                  |
|------------------------------|---------------------------------------------------------|-------------------------------|---------------------|-----------------------------------------------------------------------|------------------------------------------------------------------|
| <b>Ophthalmology</b>         | Diabetic-retinopathy and retinal-disease screening      | Fundus, OCT                   | Teo et al. (2025)   | Supports distributed large-scale screening while retaining local data | Device heterogeneity, calibration, demographic shift             |
| <b>Multimodal healthcare</b> | Outcome prediction using images plus clinical variables | Imaging + EHR/laboratory data | Dayan et al. (2021) | Shows that FL can extend beyond image-only models                     | Feature alignment, missing modalities, privacy across data types |

Brain MRI is one of the strongest demonstrations. Sheller et al. (2020) showed that multi-institutional collaboration can outperform isolated learning while keeping data at participating sites. Pati et al. (2022) expanded the scale dramatically through a 71-site glioblastoma federation using 25,256 MRI scans from 6,314 patients. Neuroimaging exposes rare-disease scale, three-dimensional model cost, class imbalance, scanner shift, and annotation variability. Thoracic imaging provided another real-world test during COVID-19. Dayan et al. (2021) combined chest X-rays with clinical variables across 20 institutions and reported strong outcome prediction and cross-site generalizability, illustrating both natural heterogeneity and multimodal federation.

Mammography emphasizes vendor and population shift. Roth et al. (2020) showed that federated breast-density classification across clinical institutions could improve over local training and generalize across sites. Digital pathology creates even stronger feature shift because tissue preparation, stain, scanner, magnification, and whole-slide processing vary substantially across centers. Ophthalmology is similarly well suited to federation because fundus photography and OCT are produced at high volume across devices and screening programs, but deployment must account for camera variation, population differences, label consistency, and calibration.

A critical methodological distinction is whether a study uses real institutions or simulated clients created by splitting one pooled dataset. Simulated clients are valuable for controlled stress tests, especially for non-IID severity, but they do not reproduce scanner, workflow, network, governance, and operational differences among real hospitals. The strongest evidence therefore combines reproducible simulations with natural multi-site validation rather than treating the two designs as equivalent.

## 5. Statistical Heterogeneity and Non-IID Optimization

Statistical heterogeneity is a defining property of medical FL. Institutions differ in label prevalence, imaging features, sample size, annotation policy, and temporal behavior. Label skew occurs when disease frequencies differ by hospital; feature shift occurs when scanner or acquisition characteristics change image appearance; quantity skew occurs when one institution contributes much more data than another; concept shift can arise from different diagnostic or annotation conventions; and temporal drift appears when equipment, protocols, or patient populations change over time.

FedAvg is sensitive to these differences because local models may move toward incompatible institutional optima before aggregation. The result can be slower convergence, oscillation, reduced global performance, and particularly poor performance at minority sites. Controlled Dirichlet partitioning is widely used to create reproducible label skew, with smaller concentration values producing more severe class imbalance across clients (Hsu et al., 2019). This is useful experimentally but should not be mistaken for a complete representation of real hospital domain shift.

FedProx addresses heterogeneity by limiting excessive local deviation from the current global model, while SCAFFOLD corrects systematic local drift using additional control information (Li et al., 2020; Karimireddy et al., 2020). FedBN is especially relevant to imaging because it keeps normalization statistics local, allowing hospitals to preserve site-specific feature behavior while sharing the rest of the network (Li et al., 2021). Adaptive, personalized, and clustered approaches further acknowledge that one global parameter set may not serve every institution equally well.

Evaluation should therefore go beyond global accuracy. Trustworthy heterogeneity analysis should report per-site performance, worst-client performance, class-specific sensitivity, convergence behavior, variation across clients, and repeated-run stability. A global improvement can conceal degradation at a small specialty hospital, which is clinically important because rare or minority clients may contain exactly the cases that motivate collaboration.

## **6. Privacy Leakage and Adversarial Threats**

FL reduces raw-data transfer but does not eliminate leakage because updates and predictions are functions of the training data. A curious server can inspect client gradients or parameter changes; a malicious server can manipulate training to amplify information; participating clients can infer properties of peers; and external users can attack a released model. Medical imaging intensifies the risk because successful inference may reveal anatomy, disease status, institutional characteristics, or cohort membership (Kaissis et al., 2020).

Gradient inversion provides one of the clearest demonstrations that shared updates can expose private inputs. Zhu et al. (2019) showed that an attacker can optimize a synthetic input until its gradient resembles an observed private gradient, and Geiping et al. (2020) demonstrated stronger reconstructions from deep networks. In medical imaging, even imperfect reconstruction can reveal clinically sensitive structures or lesions. Property inference can expose population-level characteristics that are not part of the intended task, while model inversion seeks representative features associated with outputs.

Membership inference asks whether a particular example was used for training. Shokri et al. (2017) established the modern black-box attack setting; Nasr et al. (2019) showed that white-box and federated access can provide stronger signals; and Melis et al. (2019) demonstrated unintended information leakage from collaborative updates. The medical interpretation is especially sensitive: if a model is known to have been trained on a disease-specific cohort, successful membership inference can disclose that a person may belong to that cohort.

Privacy threats are not limited to confidentiality. Malicious clients can poison data or model updates, insert backdoors, or degrade selected classes. Defenses therefore need distinct layers: differential privacy for record influence, secure aggregation for update visibility, robust aggregation for malicious-client resilience, and empirical attacks for auditing. No single defense covers the full threat surface.

## **7. Differential Privacy and Privacy-Utility Trade-offs**

Differential privacy (DP) provides a formal guarantee that the output of a randomized learning process does not change excessively when one protected unit is added or removed. In medical imaging, the protected unit must be stated explicitly. Protecting one image, all images from one patient, or an entire institutional client are fundamentally different guarantees. Patient-level privacy is usually more clinically meaningful when a patient contributes multiple slices, sequences, visits, or pathology tiles.

DP-SGD, introduced for deep learning by Abadi et al. (2016), limits the influence of each training example through gradient clipping and adds calibrated random noise during optimization. Privacy accounting then tracks the total privacy loss across repeated optimization steps. This is different from simply adding arbitrary noise to model weights. A valid privacy claim requires a defined protected unit, a stated privacy budget, the clipping strategy, noise mechanism, training duration, and accounting procedure.

The principal practical challenge is the privacy-utility trade-off. Stronger privacy generally increases optimization difficulty, but the size of the performance penalty is task dependent. Small datasets, rare

classes, severe non-IID partitions, and training from random initialization can make private optimization particularly difficult. Pretraining can reduce the amount of private fine-tuning required, but the provenance of pretrained data and the scope of the privacy guarantee must then be clear. Differential privacy may also affect underrepresented classes disproportionately, which makes global accuracy alone an insufficient utility measure.

Formal privacy should be paired with empirical auditing. A small privacy budget does not directly report real attack success, while a failed empirical attack is not a mathematical guarantee. Jayaraman and Evans (2019) showed that formal bounds and observed leakage can differ substantially. The strongest medical FL reporting therefore combines privacy accounting, matched non-private baselines, task utility, and attack evaluation under the same architecture and data conditions.

## **8. Membership Inference and Empirical Privacy Auditing**

Membership inference is one of the most practical empirical tools for testing whether a trained model exposes evidence that a record was part of its training set. The attacker can use prediction confidence, per-example loss, entropy, or learned shadow-model behavior. These attacks exploit the fact that models can behave differently on training members and unseen examples, particularly when they memorize rare, difficult, or distinctive cases (Shokri et al., 2017; Yeom et al., 2018).

The attacker model matters. Black-box attacks observe only outputs, while white-box attacks may access parameters, gradients, hidden representations, or training dynamics. Salem et al. (2019) showed that attacks can remain effective with weaker assumptions about the target. Nasr et al. (2019) demonstrated strong passive and active white-box attacks in federated settings, emphasizing that privacy must be considered during collaborative training as well as after model release.

Average attack accuracy or ROC-AUC is useful but incomplete. Carlini et al. (2022) showed that privacy leakage can be concentrated among a small subset of records and argued for low-false-positive-rate analysis. This is particularly relevant to medicine, where unusual or rare cases may be the most sensitive. An attack AUC near chance therefore means only that the tested attack failed to distinguish members well on average; it should not be described as proof of privacy.

A robust medical privacy audit should define the protected unit, use patient-disjoint member and non-member sets where possible, control class balance, document attacker knowledge, apply multiple attack signals, separate attack training from evaluation, report both average and low-FPR metrics, and repeat sampling when feasible. The same attack pipeline should be used across privacy settings so that changes in attack success can be attributed to the defense rather than to a changed audit design.

## **9. Secure Aggregation and Cryptographic Defenses**

In ordinary FL, the server receives individual client updates before combining them. Secure aggregation changes this visibility: the server should recover only the aggregate required for learning and not each hospital's update in plaintext. Bonawitz et al. (2017) introduced a practical secure-aggregation protocol based on coordinated masks, key agreement, secret sharing, and dropout recovery. This protects a different layer from differential privacy. DP limits the influence of individual records, whereas secure aggregation hides client updates from the coordinating server.

Other cryptographic approaches include secure multiparty computation (MPC), homomorphic encryption (HE), and trusted execution environments (TEEs). MPC allows multiple parties to compute an aggregate without revealing private inputs. HE permits selected operations directly over ciphertexts but can add substantial computation, memory, and bandwidth. TEEs protect server-side computation inside a hardware-isolated environment but rely on hardware trust and remain subject to implementation and side-channel concerns. Transport encryption such as TLS is necessary for network security but is not equivalent to secure aggregation because the server still decrypts and sees each update.

The main deployment trade-off is confidentiality versus system cost. Pairwise masking has relatively modest arithmetic cost, while complete secure aggregation adds protocol rounds, key management, secret sharing, authentication, and dropout handling. Homomorphic encryption can expand payloads

and substantially increase computation. A masking-only experiment should therefore be described as a prototype or lower-bound overhead estimate rather than as the cost of a production secure-aggregation protocol.

Secure aggregation also conflicts with some robustness mechanisms. Poisoning detection, anomaly screening, and Byzantine-robust aggregation often require visibility into individual updates, while secure aggregation intentionally hides them. Trustworthy hospital federation must reconcile confidentiality with integrity through secure clipping, verifiable computation, privacy-preserving anomaly detection, or related protocols rather than assuming that update secrecy and malicious-client resilience are automatically compatible.

#### 10. Five-Dimensional Taxonomy of Trustworthy Federated Medical Learning

The literature uses terms such as secure, privacy-preserving, robust, and trustworthy inconsistently. To make evidence comparable, this survey organizes trustworthiness into five dimensions: predictive utility, heterogeneity robustness, formal privacy, empirical attack resistance, and system efficiency. The taxonomy is an evidence framework, not a single numerical score. Strong evidence in one dimension cannot automatically compensate for failure in another.

**Table 2. Proposed five-dimensional taxonomy for evaluating trustworthy federated medical learning.**

| Dimension                              | Central question                                                                 | What should be evaluated?                                                                   | Representative evidence                                                                  | Typical weakness in existing studies                                          |
|----------------------------------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| <b>Predictive Utility (U)</b>          | Does federation produce clinically useful predictions?                           | Performance relative to local and centralized learning                                      | Accuracy, macro F1, ROC-AUC, Dice, sensitivity, specificity, calibration                 | <b>Utility is often reported without privacy or security analysis</b>         |
| <b>Heterogeneity Robustness (H)</b>    | Does the model remain reliable when hospitals have different data distributions? | Label skew, feature shift, quantity imbalance, scanner/domain shift, convergence behavior   | Non-IID stress tests, client-level results, worst-client performance, convergence rounds | <b>Many studies evaluate only IID or simplified label skew</b>                |
| <b>Formal Privacy (P)</b>              | What explicit protection is provided for individual training records?            | Differential privacy mechanism, protected unit, privacy budget, clipping, noise, accountant | Achieved privacy budget, privacy settings, matched private/non-private utility           | <b>Data locality is sometimes incorrectly treated as formal privacy</b>       |
| <b>Empirical Attack Resistance (A)</b> | What can a realistic adversary infer from the trained system?                    | Membership inference, gradient reconstruction, model/property inference                     | Attack ROC-AUC, attack advantage, low-FPR TPR, reconstruction quality                    | <b>Formal privacy and empirical attacks are commonly evaluated separately</b> |
| <b>System Efficiency (E)</b>           | Can the protected federation operate practically?                                | Runtime, communication, model payload, memory, secure aggregation overhead                  | Time per round, total communication, bandwidth, memory, cryptographic overhead           | <b>Efficiency is frequently omitted from trustworthy-FL evaluation</b>        |

Predictive utility asks whether the system performs the medical task effectively and should be evaluated against local and, when possible, centralized or external-site baselines. Heterogeneity robustness asks whether performance remains stable across different institutions, classes, scanners, sample sizes, and non-IID severity. Formal privacy asks what explicit guarantee is provided, including the protected unit and privacy-accounting or cryptographic assumptions. Empirical attack resistance asks what a stated adversary can actually infer through membership, reconstruction, property, or related attacks. System efficiency asks whether the protected federation can operate at acceptable runtime, communication, memory, synchronization, and cryptographic cost.

The dimensions interact. Severe heterogeneity can worsen utility and can alter privacy leakage. Differential privacy can reduce predictive performance and may disproportionately affect minority clients. Secure aggregation improves confidentiality but can limit server-side anomaly inspection. Large 3D imaging models increase communication and cryptographic cost. These interactions explain why isolated evaluation can produce a misleading picture of deployment readiness.

A practical taxonomy-guided pathway is therefore: first establish federated utility; then stress-test realistic heterogeneity; add a formal privacy mechanism; audit residual leakage; protect institutional updates; measure system overhead; and finally interpret the dimensions jointly. Specialized studies do not need to implement every layer, but research claiming broad trustworthiness should make clear which dimensions are demonstrated and which remain untested.

### 11. Cross-Study Gap Analysis

The main weakness in the literature is fragmentation rather than lack of technical ideas. Application papers often provide detailed predictive results but limited attack auditing. Optimization papers deeply examine non-IID convergence while omitting privacy. DP studies report privacy budgets and utility without testing strong attacks. Attack papers demonstrate leakage but may not evaluate clinical performance or protected-system cost. Cryptographic studies measure protocol behavior without using the same medical-imaging pipeline as downstream utility studies.

**Table 3. Cross-study gap analysis and research priorities for trustworthy federated medical learning.**

| Interaction                               | Current pattern                                  | evidence   | Major missing evidence                                                        | Research priority |
|-------------------------------------------|--------------------------------------------------|------------|-------------------------------------------------------------------------------|-------------------|
| <b>Utility × Heterogeneity</b>            | Relatively mature for synthetic non-IID settings |            | Real scanner, demographic, and protocol, institutional domain shift           | High              |
| <b>Utility × Formal Privacy</b>           | Increasing DP studies                            | DP utility | Matched architecture and training conditions across privacy budgets           | High              |
| <b>Utility × Attack Resistance</b>        | Limited                                          |            | Predictive utility and strong empirical attacks evaluated on the same model   | High              |
| <b>Heterogeneity × Formal Privacy</b>     | Sparse                                           |            | How privacy noise and clipping behave under progressively severe non-IID data | <b>Very high</b>  |
| <b>Heterogeneity × Attack Resistance</b>  | Sparse                                           |            | Whether client drift, rare classes, and domain shift increase privacy leakage | <b>Very high</b>  |
| <b>Formal Privacy × Attack Resistance</b> | Growing incomplete                               | but        | Formal DP combined with strong membership and reconstruction audits           | <b>Very high</b>  |
| <b>Formal Privacy × Efficiency</b>        | Partial                                          |            | Privacy guarantees evaluated together with runtime and communication burden   | High              |

|                                                                 |                                                        |                                                                        |                |
|-----------------------------------------------------------------|--------------------------------------------------------|------------------------------------------------------------------------|----------------|
| <b>Secure Aggregation × Utility × Efficiency</b>                | Fragmented across security and medical-imaging studies | End-to-end medical evaluation of protected aggregation and system cost | High           |
| <b>Secure Aggregation × Robust Aggregation</b>                  | Limited                                                | Confidentiality combined with poisoning/backdoor detection             | High           |
| <b>Utility × Heterogeneity × Privacy × Attacks × Efficiency</b> | Rare                                                   | Integrated trustworthy-FL evaluation under one controlled pipeline     | <b>Highest</b> |

Several intersections are therefore underdeveloped. First, heterogeneity and formal privacy are rarely studied together across controlled severity levels even though clipping and noise operate on an optimization process already affected by client drift. Second, formal privacy and empirical attack resistance are frequently reported separately even though they answer complementary questions. Third, secure aggregation and clinical utility are seldom evaluated end to end with complete protocol cost. Fourth, privacy auditing is still frequently image based even when patient-level membership is the clinically meaningful unit.

Synthetic non-IID evaluation also remains narrower than real institutional heterogeneity. Dirichlet label skew is reproducible and valuable, but it cannot capture scanner vendor, acquisition protocol, annotation style, demographic composition, workflow, and infrastructure differences simultaneously. Similarly, reported numerical precision often exceeds statistical readiness because expensive FL experiments are frequently run only once. Point estimates are useful, but claims of superiority should be separated from replicated evidence.

The most important missing evidence is integrated evaluation under one controlled architecture and dataset. A strong pipeline should determine whether federation improves over isolated learning; whether performance remains stable under realistic heterogeneity; what utility is lost under formal privacy; what leakage remains under strong attacks; whether updates can be hidden securely; and what computational and communication cost those protections introduce. This intersection is the central research gap identified by the survey.

## 12. Unresolved Challenges and Research Agenda

Real clinical deployment introduces challenges that are only partially represented by current benchmarks. Data heterogeneity remains multidimensional and evolves over time. Patient-level integrity is difficult when datasets contain multiple correlated slices, sequences, views, or longitudinal studies without reliable identifiers. Small specialty institutions and rare-disease clients can be overwhelmed by sample-weighted aggregation, yet their cases may be among the most valuable and privacy sensitive. Multimodal and vertical FL add secure patient alignment, missing modalities, heterogeneous encoders, and new intermediate-representation leakage risks.

Privacy, security, and robustness remain tightly coupled. Medical benchmarks lack broadly accepted task-specific privacy budgets. Formal privacy and empirical auditing are not routinely combined. Secure aggregation can reduce the visibility required for poisoning detection. Differential privacy may reduce already limited utility for minority groups. Repeated continual learning can accumulate privacy loss over time. Foundation and vision-language models bring large communication payloads, expensive fine-tuning, heterogeneous architectures, and possible memorization of pretraining data.

Operational deployment is equally important. Hospitals must integrate FL with PACS, DICOM workflows, EHR systems, identity and access management, local firewalls, model registries, validation procedures, audit logs, and governance policies. Questions of consent, legal basis, model ownership, responsibility for adverse outcomes, cross-border collaboration, client enrollment, key management, and version approval are not solved by the learning algorithm. Standardized reporting of bytes per round, convergence time, hardware, protocol overhead, and failure recovery is still uncommon.

The research priorities are therefore clear: use patient-level and institutionally meaningful partitions; combine synthetic stress tests with natural site shift; report local, federated, and centralized references under matched conditions; pair formal privacy with strong empirical attacks; quantify full secure-computation cost; report client-level and subgroup utility; replicate stochastic experiments where feasible; preserve seeds, partitions, privacy-accountant settings, attack scripts, and hardware details; and expand external validation at institutions that did not participate in training. The next step for the field is not merely to show that federation works, but to show when it remains trustworthy over its full lifecycle.

### **13. Conclusion**

Federated learning provides a practical route to multi-institutional medical AI without requiring hospitals to place raw images in a single repository. Evidence from brain MRI, COVID-19 outcome prediction, mammography, pathology, and other applications demonstrates that collaborative learning can improve scale and generalizability while preserving local data custody (Sheller et al., 2020; Dayan et al., 2021; Pati et al., 2022). Yet federation alone is only the starting point.

The survey shows that trustworthy federated medical learning requires five forms of evidence. The model must provide clinically meaningful utility, remain stable under real inter-hospital heterogeneity, offer an explicit formal privacy guarantee where privacy is claimed, resist realistic empirical attacks, and operate at feasible computational and communication cost. Differential privacy and secure aggregation protect different layers and are complementary. Membership inference and reconstruction attacks are not alternatives to formal privacy; they are empirical audits of residual leakage. Non-IID robustness cannot be inferred from an IID benchmark, and secure communication cannot be inferred from ordinary transport encryption.

The cross-study analysis identifies integrated evaluation as the central unmet need. The field has strong components but relatively few studies that evaluate utility, heterogeneity, formal privacy, attacks, and efficiency together under a consistent medical-imaging design. Future progress should therefore be judged not only by higher predictive scores, but also by whether the entire collaborative pipeline can be defended, audited, reproduced, governed, and deployed across real institutions.

The central conclusion is concise: data locality reduces one barrier to collaboration, but trustworthiness requires evidence across the complete learning and security lifecycle. Medical FL becomes clinically credible when collaboration, robustness, privacy, attack resistance, and operational feasibility are demonstrated together rather than assumed from the federated architecture alone.

### **REFERENCES**

- [1]. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. <https://doi.org/10.1145/2976749.2978318>
- [2]. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191. <https://doi.org/10.1145/3133956.3133982>
- [3]. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramèr, F. (2022). Membership inference attacks from first principles. *2022 IEEE Symposium on Security and Privacy*, 1897–1914.
- [4]. Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27, 1735–1743. <https://doi.org/10.1038/s41591-021-01506-3>
- [5]. Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients: How easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33, 16937–16947.
- [6]. Guan, H., Yap, P.-T., Bozoki, A., & Liu, M. (2024). Federated learning for medical image analysis: A survey. *Pattern Recognition*, 151, 110424. <https://doi.org/10.1016/j.patcog.2024.110424>
- [7]. Hsu, T.-M. H., Qi, H., & Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *NeurIPS Workshop on Federated Learning*.
- [8]. Jayaraman, B., & Evans, D. (2019). Evaluating differentially private machine learning in practice. *28th USENIX Security Symposium*, 1895–1912.
- [9]. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2, 305–311. <https://doi.org/10.1038/s42256-020-0186-1>
- [10]. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.

- [11]. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. *Proceedings of the 37th International Conference on Machine Learning*, 119, 5132–5143.
- [12]. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, 429–450.
- [13]. Li, X., Jiang, M., Zhang, X., Kamp, M., & Dou, Q. (2021). FedBN: Federated learning on non-IID features via local batch normalization. *International Conference on Learning Representations*.
- [14]. Lu, M. Y., Chen, R. J., Kong, D., Lipkova, J., Singh, R., Williamson, D. F. K., Chen, T. Y., & Mahmood, F. (2022). Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76, 102298. <https://doi.org/10.1016/j.media.2021.102298>
- [15]. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54, 1273–1282.
- [16]. Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. *2019 IEEE Symposium on Security and Privacy*, 691–706. <https://doi.org/10.1109/SP.2019.00029>
- [17]. Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. *2019 IEEE Symposium on Security and Privacy*, 739–753. <https://doi.org/10.1109/SP.2019.00065>
- [18]. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- [19]. Pati, S., Baid, U., Edwards, B., Sheller, M. J., Wang, S.-H., Reina, G. A., et al. (2022). Federated learning enables big data for rare cancer boundary detection. *Nature Communications*, 13. <https://doi.org/10.1038/s41467-022-33407-5>
- [20]. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews*, 10, 39. <https://doi.org/10.1186/s13643-020-01542-z>
- [21]. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119. <https://doi.org/10.1038/s41746-020-00323-1>
- [22]. Roth, H. R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., et al. (2020). Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* (pp. 181–191). Springer.
- [23]. Saldanha, O. L., Quirke, P., West, N. P., James, J. A., Loughrey, M. B., Grabsch, H. I., Salto-Tellez, M., Alwers, E., Cifci, D., Ghaffari Laleh, N., et al. (2022). Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nature Medicine*, 28, 1232–1239. <https://doi.org/10.1038/s41591-022-01768-5>
- [24]. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., & Backes, M. (2019). ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. *Network and Distributed Systems Security Symposium*.
- [25]. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 12598. <https://doi.org/10.1038/s41598-020-69250-1>
- [26]. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy*, 3–18. <https://doi.org/10.1109/SP.2017.41>
- [27]. Teo, Z. L., Zhang, X., Yang, Y., Jin, L., Zhang, C., Poh, S. S. J., Yu, W., Chen, Y., Jonas, J. B., Wang, Y. X., Wu, W. C., Lai, C. C., Liu, Y., Goh, R. S. M., & Ting, D. S. W. (2025). Privacy-preserving technology using federated learning and blockchain in protecting against adversarial attacks for retinal imaging. *Ophthalmology*, 132(4), 484–494. <https://doi.org/10.1016/j.ophtha.2024.10.017>
- [28]. Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019). A hybrid approach to privacy-preserving federated learning. *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, 1–11.
- [29]. Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594, 265–270. <https://doi.org/10.1038/s41586-021-03583-3>
- [30]. Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium*, 268–282.
- [31]. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32.