ORIGINAL RESEARCH **Open access**

Federated Learning with Differential Privacy and Secure Multi-Party Computation for Training Rare Disease Detection Models Across 50 International Hospitals without Centralizing Data

Kevin O'Brien[1*], Liam Murphy[1], Sean Doyle[2]

Abstract

Detecting rare diseases often requires data from multiple institutions due to the scarcity of cases at individual hospitals. Centralizing data is not feasible due to privacy, consent, and jurisdictional issues. Federated learning enables model training across hospitals without transferring raw data, but it lacks formal privacy guarantees. Model updates can still leak information, and aggregation servers may compromise privacy if they handle unprotected data. This article presents a conceptual framework combining federated learning, differential privacy, and secure multi-party computation for rare disease detection across 50+ international hospitals. The system addresses data scarcity, regulatory fragmentation, and network heterogeneity. Each hospital trains a local model, applies differential privacy to updates, and shares encrypted updates via an aggregation protocol. Non-colluding servers compute global updates without accessing plaintext data. Differential privacy reduces the impact of individual patient data, while secure multi-party computation ensures privacy at the aggregation layer. These methods enable a privacy-preserving approach to federated learning for rare disease collaboration. The proposed framework enables multi-continental rare disease detection without centralizing patient data, offering a privacy-preserving model for future consortia.

Keywords Federated learning, Rare disease detection, Differential privacy, Secure multi-party computation, Privacy-preserving artificial intelligence, Cross-hospital machine learning

*Correspondence:
Kevin O'Brien
kevin.obrien@gmail.com
[1] Department of Healthcare Intelligence Analytics, University College Dublin, Dublin, Ireland
[2] Department of AI Clinical Systems, Trinity College Dublin, Dublin, Ireland

Introduction

Rare disease detection is a paradigmatic example of the mismatch between the data requirements of modern machine learning and the realities of clinical scarcity. A single specialist hospital may observe only a small number of cases of a rare cancer, genetic disorder, or atypical imaging phenotype each year, making local model development unstable and prone to overfitting. Federated learning has therefore become attractive in medical AI because it allows institutions to contribute to shared model training without transferring raw patient data, as shown in multi-institutional brain tumor segmentation and broader healthcare informatics studies [1-4]. For rare diseases, this architectural shift is especially important because clinically meaningful learning may require coordinated participation across dozens of hospitals, regions, and legal systems [5, 6].

Federated learning reduces direct data-sharing risk, but it does not eliminate privacy risk. Local updates can encode information about rare patients, and the uniqueness of rare

disease phenotypes may make gradient leakage, membership inference, or reconstruction attacks more consequential than in common disease settings. Studies of federated medical imaging and healthcare AI emphasize that privacy preservation requires more than keeping data inside institutional firewalls, because learned parameters can still reveal sensitive information when aggregation is not protected [7-10]. This concern is amplified when training spans 50 international hospitals, where infrastructure, threat models, and governance practices vary across participating sites [11, 12].

Differential privacy addresses one part of this challenge by providing a mathematical bound on the influence of any individual patient record on the released model update. In healthcare federated learning, DP mechanisms are commonly discussed as update clipping plus calibrated noise addition, trading model utility for stronger privacy guarantees in settings where patient-level memorization is unacceptable [13-16]. Secure multi-party computation addresses a complementary threat by ensuring that aggregation servers can compute sums or averages without observing plaintext local updates. Combining these methods is therefore not redundant: DP protects against information leakage from the trained model, while SMPC protects the aggregation process itself [7, 17].

The thesis of this article is that federated learning for rare disease detection should be conceptualized as a layered privacy-preserving AI system rather than as a single distributed optimization method. The proposed framework integrates local training, DP perturbation, SMPC-based aggregation, governance constraints, and cross-continental deployment assumptions for a network of 50 or more international hospitals. It builds on prior federated learning deployments in oncology, neuroimaging, COVID-19 outcomes, and medical image segmentation while adapting their lessons to the extreme imbalance, rarity, and jurisdictional sensitivity of rare disease detection [4, 11, 18, 19]. The article proceeds by defining the data and regulatory background, specifying the framework architecture, and detailing the DP and SMPC layers needed for private rare disease model training.

Background

Rare disease data characteristics

Rare diseases are commonly defined in Europe as conditions affecting fewer than 1 in 2,000 people, and this low prevalence creates unusually fragmented evidence for model development. Conditions such as retinoblastoma, cystic fibrosis, rare sarcomas, and molecularly defined rare cancers may appear only a few times per year even in tertiary hospitals, meaning that a single site's training set can be too small for stable deep learning. Federated tumor segmentation and rare cancer boundary detection studies show why multi-site collaboration is essential when clinical labels and imaging phenotypes are distributed across institutions [4, 18, 20]. For an AI systems framework, the core design problem is therefore not merely distributed training, but distributed learning under rare labels, heterogeneous acquisition protocols, and sparse local supervision [5, 19, 21].

Privacy regulations

Rare disease data are often highly identifiable because combinations of genotype, phenotype, imaging pattern, age, and geography can make a patient distinctive even when direct identifiers are removed. Regulations such as HIPAA and GDPR motivate principles of minimum necessary use, data minimization, and restricted cross-border transfer, which align naturally with federated systems that avoid raw data centralization. However, privacy-preserving architecture must still account for the fact that model updates and aggregate models can themselves become regulated artifacts when they encode patient-level information [3, 7, 8]. Cross-continental federated healthcare studies demonstrate that technical design and legal governance must be coordinated, especially when institutions operate under different consent models and data protection authorities [11, 12].

Federated learning and FedAvg

Federated learning typically trains a model by sending a global initialization to participating hospitals, allowing each hospital to update the model locally, and aggregating local updates through a method such as federated averaging. FedAvg is attractive because it is simple and communication-efficient, but standard averaging assumes that local updates can be shared safely and that data quantity is a reasonable proxy for contribution quality. Medical FL studies have shown that distributed learning can improve site performance without direct data sharing, yet they also highlight sensitivity to site imbalance, non-identical data distributions, and heterogeneous acquisition conditions [2, 9, 10]. In rare disease detection, these limitations become central because a hospital with more

samples may still have noisier labels, while a small expert site may contribute disproportionately valuable rare-class information [4, 22, 23].

Differential privacy and secure multi-party computation

Differential privacy and secure multi-party computation solve different privacy problems within the federated pipeline. Differential privacy, usually expressed through an epsilon-delta privacy guarantee, clips and perturbs model updates so that the presence or absence of one patient has limited influence on the released information [13, 15, 16]. SMPC uses tools such as secret sharing, garbled circuits, or additive homomorphic encryption so that multiple computation parties can jointly aggregate updates without seeing any hospital's plaintext contribution [7, 17]. In a rare disease setting, the combined use of DP and SMPC is particularly important because rare-class patients are both clinically valuable and more vulnerable to re-identification through unique phenotypic signatures [5, 6].

Framework Overview

High-level architecture

The proposed framework organizes 50 or more hospitals into a privacy-preserving training network in which each site retains raw electronic health records, imaging data, genomic features, and clinical labels locally. A federation orchestrator distributes a model initialization, hospitals perform local training, DP mechanisms clip and perturb updates, and SMPC servers receive encrypted shares rather than plaintext parameters. The aggregation layer computes a protected global update, which is redistributed for subsequent rounds until a governance-defined stopping condition is reached. This architecture extends the logic of prior multi-institutional medical FL systems and federated tumor segmentation infrastructures while adding formal privacy and encrypted aggregation as core requirements rather than optional safeguards [1, 2, 4, 18].

Figure 1 presents the proposed layered architecture for privacy-preserving rare disease detection, showing how local hospital training, differential privacy, secure multi-party computation, protected aggregation, adaptive weighting, and governed clinical evaluation operate as sequential safeguards in a 50-site international federation.

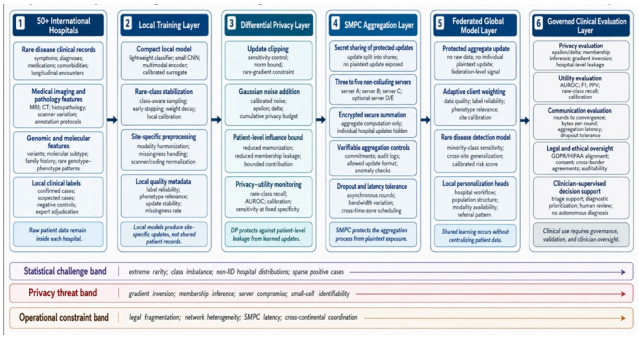

Figure 1. Layered Privacy-Preserving Federated Learning Architecture for Rare Disease Detection Across 50 International Hospitals

Core assumptions

The framework assumes that participating hospitals can maintain local data governance, implement standardized preprocessing pipelines, and report data quality metadata without exporting raw patient-level records. It also assumes that the communication network is sufficiently reliable for iterative model updates, although asynchronous or delayed participation may be necessary across continents. The SMPC layer requires multiple non-colluding aggregation servers, because collusion would weaken the protection offered by secret sharing or related protocols [7, 17]. These assumptions are consistent with federated healthcare deployments that depend not only on machine learning algorithms but also on institutional readiness, harmonized workflow, and sustained operational coordination [8, 11, 12].

Design principles

The design principles are strong privacy, verifiable aggregation, fault tolerance, scalability, and clinical proportionality. Strong privacy means that DP parameters should be selected so that patient-level contribution is bounded, with privacy budgets such as epsilon values no larger than clinically and legally defensible thresholds for the task. Verifiable aggregation means that hospitals and auditors should be able to confirm that protected updates are aggregated as specified, without forcing disclosure of raw updates or local data [3, 13, 17]. Scalability and fault tolerance are equally important because a 50-site rare disease federation cannot depend on perfect synchronous participation from every hospital in every round [4, 11, 20].

Federated Learning for Rare Diseases

Local model

For rare disease detection, the local model should be deliberately compact rather than maximally expressive, because the limiting factor is often not computational capacity but the scarcity and instability of positive cases. A linear classifier, calibrated gradient-boosting surrogate, small convolutional neural network, or lightweight multimodal encoder may be preferable when individual hospitals hold fewer than five or ten confirmed cases for a target condition. In such settings, a large model can memorize institution-specific artifacts, scanner signatures, or idiosyncratic clinical coding patterns instead of learning disease-relevant features that generalize across hospitals. Prior federated medical imaging work shows that large segmentation models can benefit from cross-site training, but rare disease detection imposes a sharper overfitting risk because the rare class may be represented by only a few local examples [4, 19, 21]. The framework therefore treats compactness, regularization, and disciplined local validation as system-level privacy and robustness choices, not merely modeling preferences [22, 23].

A compact local model also reduces the amount of information that must be communicated in each federated round, which is important when the federation spans 50 international hospitals with different network capacities and governance constraints. Smaller parameter spaces can lower the risk that individual rare cases exert disproportionate influence on local gradients, making subsequent differential privacy mechanisms easier to calibrate. This is particularly relevant for hospitals contributing rare cancers, genetic disorders, or atypical imaging phenotypes, where one mislabeled or unusually informative case may dominate a local update. A compact architecture should therefore be paired with conservative training procedures, including early stopping, weight decay, class-aware sampling, and local calibration checks before updates enter the federation [4, 19, 22]. In this framework, the local model is designed not to maximize local fit, but to produce stable, privacy-compatible updates that can be safely aggregated across heterogeneous sites [21, 23].

The local training pipeline should also support modality-specific flexibility without fragmenting the global learning objective. Hospitals may contribute imaging features, structured clinical variables, pathology descriptors, or genomic annotations, but not every site will possess the same data modalities or diagnostic depth. A lightweight multimodal encoder can accommodate such heterogeneity by learning shared representations where possible while allowing missing modalities to be handled locally. Federated medical imaging and rare cancer studies demonstrate the value of cross-site learning, but they also show that institutional variation in acquisition, annotation, and preprocessing can strongly shape model behavior [4, 18, 19]. The proposed framework therefore requires local models to be compact, auditable, and compatible with site-specific preprocessing while still contributing to a common rare disease detection objective [22, 23].

Adaptive client weighting

Standard FedAvg weights hospitals primarily by sample count, but rare disease federations should weight client contributions by data quality, label reliability, phenotype relevance, and site calibration rather than volume alone. A small expert center with curated rare sarcoma labels may provide more useful signal than a larger general hospital with noisy diagnostic codes, incomplete genetic confirmation, or inconsistent imaging protocols. This distinction is crucial because rare disease datasets are not merely small; they are often unevenly distributed across referral centers, national registries, specialist clinics, and general hospitals. Multi-site federated studies in medical imaging and clinical prediction show that site heterogeneity can shape model performance, motivating aggregation designs that account for local distribution shift and institutional variation [9, 10, 24]. Adaptive weighting should therefore combine sample size with quality indicators, uncertainty estimates, and safeguards against any single low-quality site degrading the global rare disease detector [14, 20].

Adaptive client weighting should operate as a governance-aware aggregation strategy rather than as a purely numerical adjustment. Each hospital's contribution can be weighted according to pre-specified metadata such as diagnostic confirmation method, label adjudication process, imaging protocol completeness, missingness rate, class balance, and historical update stability. For example, a site with few but molecularly confirmed cases of a rare cancer may receive higher rare-class relevance weight than a larger site relying on administrative codes alone. Conversely, a hospital with high update variance, inconsistent labels, or unexplained shifts in feature distributions may be down-weighted until data quality

concerns are reviewed. This approach reflects evidence from federated healthcare studies showing that site differences are not peripheral noise but central determinants of model utility and fairness [9, 10, 24].

The weighting mechanism should also protect the federation from both statistical and operational failure modes. Statistically, it should prevent large hospitals with many negative cases from overwhelming the rare positive signal contributed by specialist centers. Operationally, it should prevent an unreliable or poorly calibrated site from repeatedly steering the global model toward local artifacts. Uncertainty-aware weighting, robust aggregation, and site-level monitoring can help distinguish useful rare phenotype variation from harmful noise, although this distinction will remain difficult when the target condition is extremely uncommon [14, 20]. The framework therefore treats adaptive client weighting as a core component of rare disease FL, because equitable and clinically meaningful aggregation cannot be achieved by sample-count weighting alone [4, 24].

Differential Privacy Integration

Local DP versus central DP

Local DP requires each hospital to add noise before its update leaves the institution, which provides stronger protection against an untrusted aggregation layer but may substantially reduce utility for rare classes. Central DP adds noise after aggregation, which can preserve more signal because noise is applied to a combined update, but it requires trust that the aggregator can see or correctly process sensitive contributions. In this framework, central DP alone is insufficient because the aggregation server is explicitly treated as a possible privacy risk, while local DP alone may swamp rare disease signal at small sites [13, 15, 16]. A practical architecture therefore combines hospital-side clipping and perturbation with SMPC aggregation so that privacy is not dependent on a single trusted party [7, 17].

Privacy budget allocation

Privacy budget allocation must reflect the fact that rare disease detection often involves heterogeneous risk across hospitals, diseases, and data modalities. A site contributing genomic features for an ultra-rare disorder may require a stricter budget than a site contributing common imaging covariates, even when both participate in the same federation. Prior work on federated healthcare privacy shows that DP design cannot be separated from utility, because aggressive noise can reduce sensitivity for clinically important but underrepresented classes [13, 14, 16]. The proposed framework therefore allocates epsilon and delta across training rounds, sites, and model components while monitoring whether rare-class recall deteriorates below clinically acceptable thresholds [5, 6].

Gaussian mechanism parameter tuning

Gaussian DP mechanisms require a clipping bound for model update sensitivity and a noise scale calibrated to the desired privacy budget. In rare disease federations, clipping is not merely a mathematical operation because extreme gradients may represent either privacy-sensitive memorization or clinically meaningful rare-class signal. Studies of DP in federated medical image analysis emphasize that privacy tuning must be evaluated together with model robustness, site heterogeneity, and the downstream clinical task [13, 15, 16]. The framework therefore treats the clipping bound and Gaussian noise scale as governance-relevant parameters that should be pre-specified, audited, and adapted cautiously when training across 50 international hospitals.

Secure Multi-Party Computation Aggregation

Aggregation protocol

The aggregation protocol uses secret sharing across three to five non-colluding computation servers so that no single server can reconstruct a hospital's local update. Each hospital clips and perturbs its model update, splits the protected update into cryptographic shares, and sends different shares to separate SMPC servers for secure summation. The servers compute only the aggregate update needed for global model training, while individual hospital contributions remain hidden throughout the aggregation process [7, 17]. This design is aligned with federated healthcare systems that require privacy-preserving collaboration across institutions without exposing local model parameters to a central coordinator [8, 18, 25].

Communication overhead

SMPC increases communication cost because hospitals transmit multiple update shares and servers may require additional synchronization rounds to complete secure aggregation. In a cross-continental federation of 50 hospitals, this overhead can be significant because latency, bandwidth, and regulatory routing constraints differ across regions. Prior federated medical imaging and international clinical prediction studies show that large-scale FL is operationally feasible, but they also imply that communication design must be treated as a first-order systems constraint rather than a secondary engineering detail [4, 9, 11]. The framework therefore favors compact models, sparse or compressed updates, asynchronous participation, and round scheduling that balances privacy guarantees against practical network feasibility [22-24].

Resilience to malicious hospitals

A rare disease federation must account not only for honest-but-curious servers but also for malicious or faulty hospitals that may submit corrupted, mislabeled, or adversarial updates. Zero-knowledge proofs, secure commitments, and update validation can help verify that encrypted shares correspond to permitted update formats without revealing their contents. Outlier-resistant aggregation is also necessary because rare disease data are naturally heterogeneous, making it difficult to distinguish harmful updates from legitimate rare phenotype variation [14, 20]. The proposed framework therefore combines SMPC with anomaly detection, update norm constraints, and governance escalation pathways rather than assuming all participating hospitals behave identically [12, 17].

Combined Privacy Architecture

Layered privacy model

The combined architecture separates privacy protection into institutional, transmission, model-training, and aggregation layers. At the hospital layer, data minimization and local governance restrict what is collected and processed; at the transmission layer, encrypted channels and SMPC shares prevent exposure of model updates in transit. At the model-training layer, DP constrains patient-level influence, while at the aggregation layer SMPC prevents servers from observing plaintext local contributions [7, 13, 17]. This layered model extends the privacy-by-design orientation proposed for rare disease research and adapts it to federated clinical AI systems operating across multiple jurisdictions [6, 26, 27].

Table 1 specifies how each architectural layer contributes a distinct privacy function, clarifying why federated learning, differential privacy, secure multi-party computation, and governance controls are complementary rather than interchangeable safeguards.

Table 1. Layered Privacy Functions in the Proposed Federated Rare Disease Detection Architecture

Architectural layer	Primary privacy function	Main technical mechanism	Rare diseas specific address
Institutional data-locality layer	Prevents direct transfer of raw patient records across borders	Local storage, local preprocessing, local training environments	Genotyp phenoty combinat may b identifial even afte identifica
Local model-stabilization layer	Reduces overfitting and memorization before updates leave the site	Compact models, early stopping, regularization, class-aware sampling	Very sm positive c counts c domina gradien
Differential privacy layer	Bounds the contribution of any individual patient to the released update	Clipping, Gaussian noise, epsilon–delta accounting	Rare pati are mo vulnerabl members inference reconstru
Secure transmission layer	Protects update movement between hospitals and computation servers	Encrypted channels, authenticated communication, key management	Cross-bo networl increas intercept and rout risks

SMPC aggregation layer	Prevents aggregation servers from seeing plaintext hospital updates	Secret sharing, secure summation, non-colluding servers	A singl exper hospita update n reveal a r cohor structu
Adaptive contribution layer	Prevents volume-dominant sites from overwhelming rare expert signal	Quality-aware weighting, label reliability scores, update stability monitoring	Small speciali centers n hold the n valuable r class evidenc
Governance and clinical oversight layer	Ensures technical safeguards remain legally and clinically defensible	Cross-border agreements, audit trails, clinician review, deployment rules	Rare dise decisio may affe small identifia patien group

Attack resilience

The framework is designed to reduce exposure to three major attack classes: gradient inversion, server compromise, and network eavesdropping. DP reduces the likelihood that rare patient features can be reconstructed from model updates, SMPC prevents compromised aggregation servers from seeing individual hospital contributions, and transport encryption protects communication links between hospitals and computation servers. Medical FL studies have repeatedly emphasized that federated training alone should not be equated with full privacy protection, especially when updates may contain sensitive clinical information [7, 8, 13]. For rare diseases, this attack resilience is especially important because a single reconstructed phenotype may correspond to a small and identifiable patient population [5, 15, 16].

Handling Extreme Data Scarcity

Synthetic data augmentation

Synthetic data augmentation can help hospitals pre-train local representations before joining federated rounds, but it should remain local and should not become a substitute for privacy-preserving learning. A hospital may generate synthetic rare cases through simulation, generative modeling, or controlled augmentation, then use only DP-protected updates during federation. This approach is consistent with the broader logic of federated tumor segmentation and rare cancer boundary detection, where multi-site learning is needed because real positive cases are too sparse for robust single-institution training [4, 18, 20]. However, synthetic rare cases must be audited carefully because unrealistic generated samples can amplify bias or teach the global model artifacts rather than disease-relevant structure [19, 24].

Cross-site knowledge transfer

Cross-site knowledge transfer can be implemented through meta-learning initialization, shared representation learning, or personalized heads trained locally at each hospital. The global model can learn broad disease-relevant features across the federation, while each hospital retains a local personalization layer calibrated to its imaging devices, coding practices, population structure, or clinical workflow. Federated segmentation studies and distributed contrastive learning approaches illustrate how cross-site representation sharing can improve generalization without requiring raw data exchange [21, 23, 28, 29]. In rare disease detection, this personalization layer is particularly important because local prevalence, referral patterns, and diagnostic workups may differ substantially even when hospitals participate in the same international consortium [14, 24].

Cross-Continental Deployment

Legal and ethical harmonisation

A cross-continental rare disease federation requires legal agreements that describe FL, DP, and SMPC not as informal safeguards but as enforceable components of the system architecture. Data sharing agreements should specify that raw patient data remain local, model updates are protected before transmission, and aggregation occurs through cryptographic protocols that prevent access to individual institutional contributions. International federated healthcare studies show that successful deployment depends on institutional governance, consent alignment, and trust between participating sites as much as on algorithmic performance [7, 8, 23]. For rare disease

consortia, these agreements should also address small-cell risks, genomic identifiability, withdrawal procedures, and responsibilities when model outputs influence clinical decision support [5, 6, 26].

Infrastructure requirements

The infrastructure requires a federation orchestrator, local hospital training environments, secure key management, SMPC server clusters, audit logging, and monitoring for client availability and update quality. Trusted execution environments may support hardened orchestration or server-side isolation, but they should complement rather than replace DP and SMPC because hardware trust assumptions can fail or vary across jurisdictions. Prior work on federated healthcare and tumor segmentation demonstrates the need for reusable tooling, standardized workflows, and benchmarking mechanisms when many institutions participate in a shared AI effort [3, 18, 20, 25]. The framework therefore treats infrastructure as a socio-technical substrate that must support asynchronous communication, reproducibility, governance audits, and continuous clinical oversight [11, 27].

Evaluation Strategy

Table 2 provides an evaluation matrix for determining whether the proposed system is not only technically private, but also clinically useful, operationally feasible, and legally auditable in a 50-hospital rare disease federation.

Table 2. Evaluation Matrix for Privacy-Preserving Rare Disease Federated Learning Across 50 Hospitals

Evaluation domain	Core question	Recommended metrics	R dis spe interp
Formal privacy	Is individual patient contribution mathematically bounded?	Epsilon, delta, cumulative privacy budget, clipping sensitivity	Str budge be nee ultra geno phen subg
Empirical privacy leakage	Can patient membership or features be inferred despite DP and SMPC?	Membership inference success, gradient inversion success, reconstruction quality	Eve recons ra pher may ser identi ri
Aggregation confidentiality	Can servers compute the global update without seeing plaintext hospital updates?	Number of SMPC servers, collusion threshold, share size, secure summation success	A s hos upda repre hig spec rare
Rare-class utility	Does privacy protection preserve clinically meaningful detection?	Rare-class recall, F1 score, PPV, AUROC, sensitivity at fixed specificity	Aggr accu insuf bec neg ca dom
Calibration and clinical reliability	Are predicted risks interpretable and safe for clinical triage?	Calibration slope, Brier score, calibration-in-the-large, decision-curve analysis	P calib can n clini when preva extren
Cross-site generalization	Does the model perform across heterogeneous hospitals and regions?	Leave-one-site-out performance, external-site AUROC, site-level variance	Str ave perfor may failu sr spe hos
Client contribution fairness	Are expert rare disease sites appropriately represented?	Adaptive weights, label-quality scores, update stability, contribution influence	Sampl weight suppre but quality sig
Communication feasibility	Can 50 hospitals participate without	Bytes per round, rounds to convergence, dropout tolerance,	SMP DP n scien sour

	excessive burden?	latency, synchronization overhead	opera impr
Governance readiness	Is the system auditable across legal jurisdictions?	Consent alignment, audit logs, data-processing agreements, model access controls	Rare dat rer sensiti witho data t
Clinical deployment boundary	Is the model used as supervised decision support rather than autonomous diagnosis?	Clinician review rate, override rate, prospective validation, safety monitoring	Rare pred require interp a adjud

Privacy metrics

Privacy evaluation should report the achieved epsilon and delta values, the cumulative privacy budget across rounds, and the sensitivity assumptions used for clipping and noise calibration. It should also include empirical privacy stress tests such as membership inference attack success and gradient inversion success, with lower attack success interpreted as stronger practical protection. DP-focused federated healthcare studies show that formal guarantees and empirical attacks provide complementary evidence because privacy budgets alone may not reveal implementation weaknesses or task-specific leakage risks [13, 15, 16]. For the proposed framework, privacy metrics should be evaluated at both patient level and hospital level because rare disease cohorts can make institutional contributions identifiable even when individual records are protected [5, 14].

Utility metrics

Utility evaluation should focus on rare disease detection performance rather than aggregate accuracy alone. Appropriate metrics include AUROC, F1 score, positive predictive value, calibration, and sensitivity at a fixed specificity threshold suitable for clinical triage. Prior multi-institutional FL studies show that federated models can improve generalization across sites, but rare disease settings require special attention to minority-class recall and performance stability across hospitals [2, 4, 10]. The evaluation strategy should therefore present privacy-utility trade-off curves showing how DP noise, clipping, and SMPC-compatible compression affect clinically meaningful detection performance [19, 22, 24].

Communication metrics

Communication evaluation should measure rounds to convergence, bytes transmitted per hospital per round, server-side aggregation latency, dropout tolerance, and the feasibility of training across time zones. SMPC-specific metrics should include the number of cryptographic shares, the number of aggregation servers, synchronization overhead, and the effect of secure aggregation on end-to-end training time. Cross-national FL studies and federated medical imaging benchmarks indicate that a model can be scientifically promising yet operationally impractical if communication costs exceed institutional capacity [9, 11, 25]. For a 50-hospital federation, communication metrics must therefore be reported alongside privacy and utility metrics rather than treated as implementation details [6, 24].

Limitations

Technical limitations

The main technical limitation is that DP noise may overwhelm the rare-class signal when individual hospitals have fewer than ten positive cases. SMPC also adds latency and bandwidth requirements, making training slower and more complex than standard FedAvg or centralized learning. Medical FL studies show that site heterogeneity, non-identical data distributions, and model instability remain difficult even before formal privacy and cryptographic layers are added [9, 19, 24]. The proposed framework therefore improves privacy architecture but does not remove the underlying statistical challenge of learning from extremely sparse rare disease data [5, 14].

Clinical limitations

Clinical implementation depends on label quality, diagnostic consistency, and harmonized phenotype definitions across hospitals. Rare disease labels may be inconsistent because diagnosis can depend on genetic testing availability, specialist review, imaging interpretation, or evolving disease classifications. Federated rare disease research frameworks highlight the importance of governance and privacy-by-design, but they also imply that

technical privacy protections cannot compensate for poor clinical annotation or weak institutional coordination [6, 26, 27]. Continuous data quality monitoring, clinician review, and cross-site adjudication are therefore necessary for the framework to support safe rare disease detection [4, 20].

Conclusion

Federated learning with differential privacy and secure multi-party computation offers a coherent systems architecture for rare disease detection across 50 or more international hospitals. The framework keeps raw patient data local while enabling hospitals to contribute to a shared model through protected updates and encrypted aggregation. It is designed for settings where data scarcity, legal constraints, and clinical urgency make conventional centralized learning impractical.

The key advantage of the framework is that it combines complementary protections rather than relying on a single privacy mechanism. Federated learning addresses data locality, differential privacy bounds patient-level influence, and secure multi-party computation protects the aggregation layer from plaintext exposure. Together, these components enable multi-continental collaboration while reducing the privacy risks associated with rare and potentially identifiable clinical phenotypes.

The framework also has important limitations. Communication costs may be high, secure aggregation can introduce latency, and differential privacy can reduce sensitivity for rare classes when data are extremely sparse. Legal harmonisation across jurisdictions remains difficult, and technical safeguards must be accompanied by clear governance, clinical validation, and ongoing data quality monitoring.

Future implementation should occur through rare disease consortia that already coordinate specialist hospitals, registries, and clinical expertise. Networks such as undiagnosed disease programs and European rare disease collaborations could use standardized privacy frameworks to make federated AI development more reproducible, auditable, and ethically defensible. The central goal is not to replace clinical expertise, but to create a secure learning infrastructure that helps rare disease knowledge move across borders without forcing patient data to do the same.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 Sep 2025 Revised: 21 Nov 2025 Accepted: 27 Jan 2026
Published online: 20 July 2026

References

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: International MICCAI Brainlesion Workshop; 2018 Sep 16; Granada, Spain. Cham: Springer; 2018. p. 92-104.

Guo P, Wang P, Zhou J, Jiang S, Patel VM. Multi-institutional collaborations for improving deep learning-based magnetic

resonance image reconstruction using federated learning. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit; 2021; Nashville, TN. Piscataway (NJ): IEEE; 2021. p. 2423-32.

Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res. 2021;5(1):1-9.

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.

Wang J, Ma F. Federated learning for rare disease detection: a survey. Rare Dis Orphan Drugs J. 2023;2(4):N-A.

Süwer S, Ullah MS, Probul N, Maier A, Baumbach J. Privacy-by-design with federated learning will drive future rare disease research. J Neuromuscul Dis. 2026;13(1):6-19.

Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-11.

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.

Yang D, Xu Z, Li W, Myronenko A, Roth HR, Harmon S, et al. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Med Image Anal. 2021;70:101992.

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.

Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021;27(10):1735-43.

Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm learning for decentralized and confidential clinical machine learning. Nature. 2021;594(7862):265-70.

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.

Koutsoubis N, Waqas A, Yilmaz Y, Ramachandran RP, Schabath MB, Rasool G. Privacy-preserving federated learning and uncertainty quantification in medical imaging. Radiol Artif Intell. 2025;7(4):e240637.

Onireti MY, Shukla RM, Das T. Splitting smarter: differential privacy for secure healthcare federated learning. Sci Rep. 2025;15(1):43625.

Zheng L, Cao Y, Yoshikawa M, Zhang X, Liu Y, Chen H, et al. Sensitivity-aware differential privacy for federated medical imaging. Sensors. 2025;25(9):2847.

Taiello R, Cansiz S, Vesin M, Montagner A, Andrearczyk V, Depeursinge A, et al. Enhancing privacy in federated learning: secure aggregation for real-world healthcare applications. In: Int Conf Med Image Comput Comput Assist Interv; 2024 Oct 7; Marrakesh, Morocco. Cham: Springer; 2024. p. 204-14.

Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research. Phys Med Biol. 2022;67(20):204002.

Manthe M, Duffner S, Lartizien C. Federated brain tumor segmentation: an extensive benchmark. Med Image Anal. 2024;97:103270.

Zenk M, Baid U, Pati S, Sheller M, Xu Z, Harmon S, et al. Towards fair decentralized benchmarking of healthcare AI algorithms with the Federated Tumor Segmentation (FeTS) challenge. Nat Commun. 2025;16(1):6274.

Tuladhar A, Tyagi L, Souza R, Forkert ND. Federated learning using variable local training for brain tumor segmentation. In: International MICCAI Brainlesion Workshop; 2021 Sep 27; Strasbourg, France. Cham: Springer; 2021. p. 392-404.

Wicaksana J, Yan Z, Zhang D, Yang X, Cheng KT, Wu M, et al. Fedmix: mixed supervised federated learning for medical image segmentation. IEEE Trans Med Imaging. 2022;42(7):1955-68.

Wu Y, Zeng D, Wang Z, Shi Y, Hu J. Distributed contrastive learning for medical image segmentation. Med Image Anal. 2022;81:102564.

Guan H, Yap PT, Bozoki A, Liu M. Federated learning for medical image analysis: a survey. Pattern Recognit. 2024;151:110424.

Pati S, Baid U, Zenk M, Edwards B, Sheller M, Wang SH, et al. The federated tumor segmentation (FeTS) challenge. arXiv. 2021;arXiv:2105.05874.

Montalvo N, Requena F, Capriotti E, Rausell A. Federated learning for the pathogenicity annotation of genetic variants in multi-site clinical settings. Bioinformatics. 2025;41(10):btaf523.

Meduri K, Nadella GS, Yadulla AR, Reddy PK, Venkata SK, Kumar PS, et al. Leveraging federated learning for privacy-preserving analysis of multi-institutional electronic health records in rare disease research. J Econ Technol. 2025;3:177-89.

Wen J, Li X, Ye X, Li X, Mao H. A highly generalized federated learning algorithm for brain tumor segmentation. Sci Rep. 2025;15(1):21053.

Alphonse S, Mathew F, Dhanush K, Dinesh V. Federated learning with integrated attention multiscale model for brain tumor segmentation. Sci Rep. 2025;15(1):11889.