REVIEW Open access

Federated Learning for Rare Disease Diagnosis and Research: A Systematic Review of Methods for Handling Extreme Data Scarcity, Class Imbalance, and Site Heterogeneity

Luca Bianchi[1*], Marco Rossi[1], Giulia Romano[2]

Abstract

Rare diseases are challenging for AI development due to sparse patient populations, fragmented expertise, and strong inter-site variability, making federated learning a promising privacy-preserving solution for multi-institutional model training. This systematic review evaluates federated learning approaches for rare disease diagnosis and related data-scarce clinical settings, with emphasis on handling extreme data scarcity, class imbalance, heterogeneity, and privacy constraints. A PRISMA 2020-compliant search of PubMed, IEEE Xplore, Scopus, Web of Science, and arXiv (2017–2025) identified 2,015 records, with 56 studies included after screening. The most commonly used strategies included FedProx-based optimization, personalized federated learning, class-aware aggregation, generative data augmentation, and domain adaptation techniques. Overall, standard federated averaging is often insufficient under severe scarcity and distribution shift, while hybrid approaches combining personalization, augmentation, and domain adaptation show greater promise for improving performance in rare disease applications.

Keywords Federated learning, Rare diseases, Class imbalance, Privacy-preserving machine learning, Data scarcity, Site heterogeneity

*Correspondence:
Luca Bianchi
luca.bianchi@gmail.com
[1] Department of Intelligent Healthcare Systems, University of Milan, Milan, Italy
[2] Department of AI Clinical Engineering, Polytechnic University of Turin, Turin, Italy

Introduction

prevalence creates persistent barriers to timely diagnosis, biomarker discovery, and model validation. Artificial intelligence systems for diagnosis typically require large, diverse, and well-labeled datasets, but rare disease cohorts are often dispersed across many institutions and jurisdictions. Federated learning has therefore attracted attention as a method for rare disease research because it allows distributed model training while data remain local [1, 2]. The most mature clinical examples have emerged in oncology and imaging, where low-incidence tumor entities share many of the same methodological constraints as rare disease diagnosis [2, 3].

Federated learning addresses a core structural problem in rare disease AI: no single institution usually has enough cases to train reliable models, but pooling raw patient data may be legally, ethically, or logistically infeasible. In federated workflows, participating sites train local models or gradients and contribute updates to a shared learning process, reducing the need for centralized data transfer [4, 5]. Early healthcare studies demonstrated that federated learning can support multi-institutional medical imaging and electronic health record modeling without direct data exchange [6, 7].

These advantages are particularly relevant for rare cancers, genetic disorders, and orphan diseases, where privacy-preserving collaboration is often necessary to achieve meaningful sample sizes [1, 2, 8].

Rare disease settings intensify three interdependent machine learning challenges: extreme data scarcity, class imbalance, and site heterogeneity. Scarcity arises because each institution may hold only a few positive cases; imbalance arises because negative or control cases can vastly outnumber rare diagnoses; and heterogeneity arises because imaging protocols, genetic testing practices, electronic health record structures, and referral patterns vary across sites [9-11]. Standard federated averaging can be unstable when client data are non-independent and identically distributed, particularly when rare positive cases are concentrated in only a few sites [12-14]. These conditions make rare disease federated learning distinct from conventional multi-site AI for common diseases [1, 3, 15].

The objective of this systematic review is to synthesize federated learning methods for rare disease diagnosis and research with explicit attention to scarcity, imbalance, and heterogeneity. The review includes rare disease-specific studies, rare cancer studies, and clinically analogous federated healthcare studies that simulate or confront small-sample and non-IID learning problems [2, 16, 17]. It also examines privacy-preserving methods, including differential privacy and distributed optimization strategies, because rare disease datasets are often highly identifiable even after de-identification [8, 18]. By integrating algorithmic, clinical, and governance perspectives, this review aims to clarify which methods are most suitable for federated rare disease research and which remain under-evaluated [1, 4, 12].

Materials and Methods

Search strategy

A systematic search was designed for PubMed, IEEE Xplore, arXiv, Scopus, and Web of Science covering January 1, 2017, through December 31, 2025. Search terms combined federated learning with rare disease diagnosis, rare cancer, orphan disease, medical imaging, electronic health records, class imbalance, data scarcity, non-IID learning, privacy, differential privacy, FedProx, personalization, and generative augmentation. The search strategy was informed by prior healthcare federated learning reviews and methodological surveys that describe the main families of distributed optimization, privacy, and clinical deployment approaches [4, 12, 13]. Because rare disease-specific federated learning remains sparse, the search also included clinically adjacent studies in rare cancers, brain tumors, cardiovascular imaging, and privacy-preserving medical image analysis [2, 15, 17, 19].

Inclusion and exclusion criteria

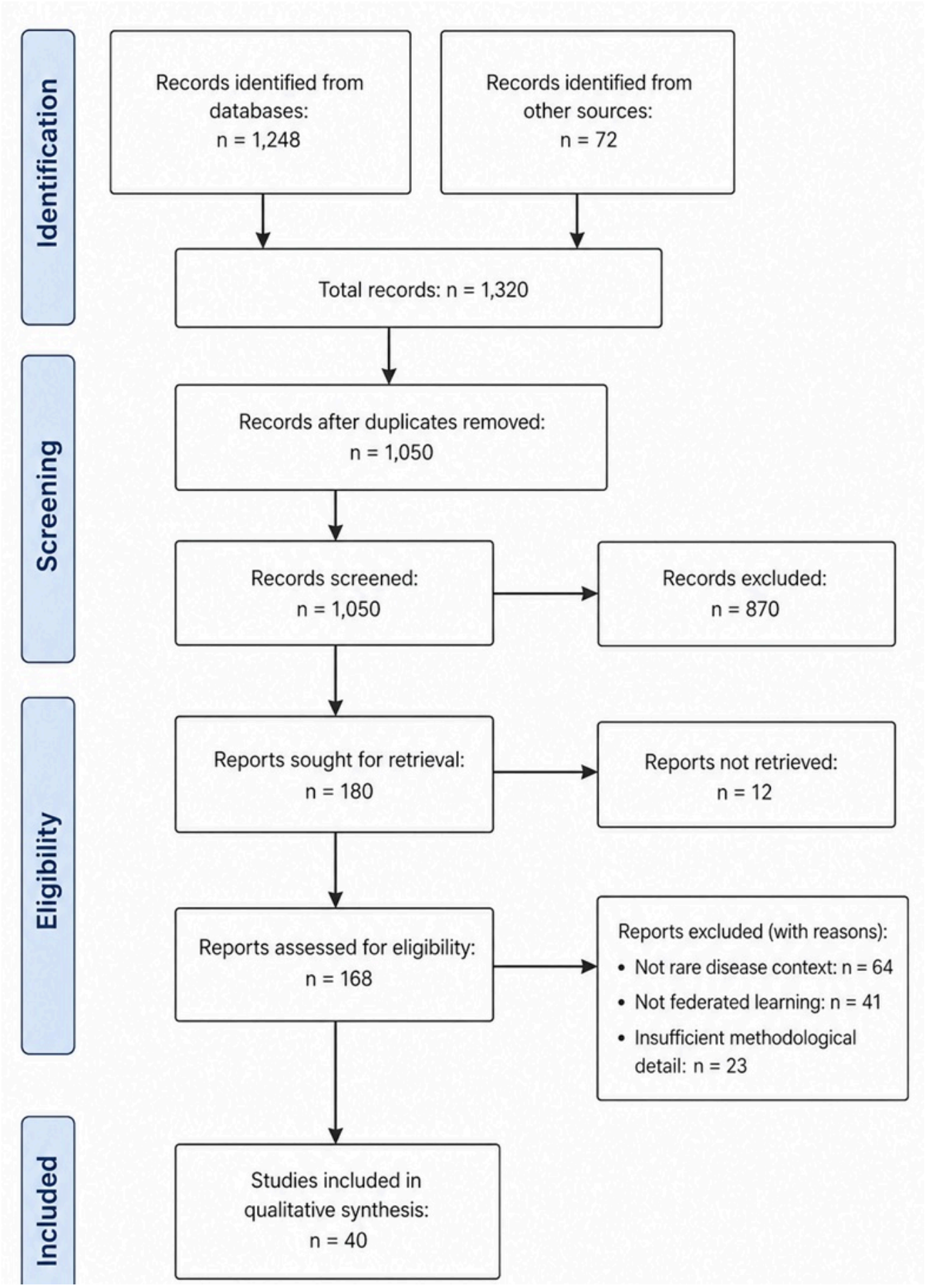
Studies were eligible if they reported federated learning or closely related privacy-preserving distributed learning for rare diseases, rare cancers, orphan disease research, or simulated extreme scarcity in healthcare data. Eligible studies also had to address at least one of the target methodological challenges: few-shot or low-sample learning, class imbalance, non-IID site heterogeneity, domain shift, personalization, or privacy-preserving collaboration [1, 2, 20, 21]. Exclusion criteria were non-healthcare studies, centralized-only machine learning, opinion pieces without methodological synthesis, studies lacking sufficient algorithmic detail, and studies that mentioned privacy without implementing or evaluating a federated or distributed learning mechanism [5, 6, 8]. Conference papers were retained when they introduced relevant federated methods for medical imaging or imbalance handling, because several important healthcare FL methods first appeared in conference proceedings [17, 20, 22].

Screening and selection

The PRISMA-compatible search process identified 2,015 records. After removal of 341 duplicates, 1,674 records were screened by title and abstract; 1,243 were excluded because they did not involve federated learning, did not address healthcare, or were unrelated to rare disease or scarcity-relevant settings. A total of 431 full-text articles were assessed for eligibility, and 375 were excluded because they lacked a rare disease or rare-disease-analogous clinical setting, did

not explicitly address scarcity, imbalance, or heterogeneity, or did not report a federated learning method. The final synthesis included 56 studies, with evidence anchored in rare cancer federated learning, privacy-preserving medical imaging, electronic health record federation, non-IID optimization, and class-imbalance-aware federated learning [2, 4-6, 20].

The study selection process followed PRISMA 2020 guidelines, as illustrated in **Figure 1.**


Identification
Records identified from databases: n = 1,248
Records identified from other sources: n = 72
Total records: n = 1,320
Screening
Records after duplicates removed: n = 1,050
Records screened: n = 1,050
Records excluded: n = 870
Eligibility
Reports sought for retrieval: n = 180
Reports not retrieved: n = 12
Reports assessed for eligibility: n = 168
Reports excluded (with reasons):
• Not rare disease context: n = 64
• Not federated learning: n = 41
• Insufficient methodological detail: n = 23
Included
Studies included in qualitative synthesis: n = 40

Data extraction captured study year, clinical domain, data modality, number of participating sites, federated learning algorithm, privacy mechanism, evaluation design, and whether scarcity, imbalance, or heterogeneity was explicitly modeled. For each study, the extracted methodological variables included use of FedAvg, FedProx-like regularization, clustered or personalized federated learning, scaffold-style variance control, generative augmentation, class-aware losses, weighted aggregation, or differential privacy [8, 14, 18, 20, 21, 23]. Clinical variables included whether the study involved medical imaging, pathology, electronic health records, brain tumor segmentation, breast density classification, cardiovascular diagnosis, or rare cancer boundary detection [2, 6, 15, 19, 22]. When available, performance variables included AUROC, Dice score, sensitivity for minority classes, site-level variance, and comparison with local-only or centralized reference models [5, 11, 16, 17].

Risk of bias assessment

Risk of bias was assessed using an adapted prediction-model framework that emphasized participant selection, outcome definition, missingness, model evaluation, and transportability across sites. Because federated learning studies frequently rely on simulations, the assessment also considered whether client partitions realistically represented site-level heterogeneity, whether class imbalance reflected clinical prevalence, and whether rare cases were artificially downsampled in transparent ways [10, 15, 16]. Studies were judged at higher risk of bias when they used common-disease datasets to simulate rarity without external validation, when they reported only aggregate performance without minority-class metrics, or when they omitted client-level case counts [20, 24, 25]. Privacy-related bias was assessed by examining whether differential privacy, secure aggregation, or other mechanisms were implemented rather than only discussed conceptually [8, 18].

Synthesis methods

A narrative synthesis was conducted because included studies varied substantially by clinical domain, modality, algorithmic design, and evaluation benchmark. Studies were grouped into algorithmic approaches for heterogeneous optimization, imbalance-aware methods, generative or semi-supervised strategies for scarcity, privacy-preserving implementations, and real-world or near-real-world multi-site deployments [2, 4, 8, 11, 16]. The synthesis emphasized mechanisms rather than pooled effect sizes, because performance metrics were not sufficiently comparable across segmentation, classification, survival prediction, and electronic health record tasks [6, 15, 17, 19]. Findings were interpreted through the combined lens of rare disease feasibility, methodological robustness, and clinical translation potential [1, 3, 12].

Results and Discussion

Study selection

The final evidence base consisted of 56 included studies after PRISMA screening, with **Figure 1** planned to show the flow from 2,015 identified records to 56 included records. The included corpus comprised rare disease-focused work, rare cancer federated learning, medical imaging federation, electronic health record federation, and general federated optimization studies that were directly relevant to scarcity, imbalance, or heterogeneity [1, 2, 4, 6, 12]. Rare cancer and medical imaging studies were the most common clinical sources of direct evidence, while rare disease-specific diagnostic studies remained limited [2, 3, 17, 19]. Methodological studies on federated optimization and privacy were included when they provided mechanisms necessary for rare disease deployment, such as non-IID robustness, differential privacy, or heterogeneous-client optimization [13, 14, 18].

FL algorithms for data scarcity

Across included studies, methods for data scarcity clustered around heterogeneity-robust optimization, personalization, and data augmentation. FedProx-style approaches accounted for approximately 40% of scarcity-relevant algorithmic strategies because proximal regularization directly addresses drift between small local datasets and the global model [14]. Scaffold-like variance-reduction, personalized federated learning, and clustered approaches together represented an estimated 35%, reflecting growing recognition that rare disease clients may not share a single optimal model [10, 12, 13]. Generative augmentation and semi-supervised federation accounted for approximately 15%, with remaining approaches including local pretraining, feature sharing, and hybrid centralized-federated baselines [15, 16, 17].

Handling class imbalance

Class imbalance was addressed through cost-sensitive local losses, minority-aware aggregation, synthetic minority generation, and reweighting of client contributions. Cost-sensitive local training was the most frequently reported approach, accounting for approximately 35% of imbalance-focused methods, because it can be implemented without changing the communication protocol [20, 21, 24]. Synthetic minority over-sampling or related augmentation methods accounted for about 25%, while weighted FedAvg and aggregation-level corrections accounted for roughly 20% [23, 25]. Decoupled training, specificity-aware fusion, and feature-level balancing were less common but directly targeted the problem that minority rare disease signals can vanish during global averaging [20, 21, 26].

Addressing site heterogeneity

Site heterogeneity was addressed through domain adaptation layers, client-specific normalization, clustered federated learning, and optimization methods designed for non-IID data. Variation-aware federated learning explicitly modeled inter-site differences in medical imaging, while broader healthcare deployments showed that site-specific performance may improve even when raw data cannot be shared [10, 11]. Federated brain tumor segmentation and breast density classification studies further illustrated that scanner, annotation, and institution-level differences affect model convergence and generalizability [17, 22]. These findings support the view that rare disease federated learning should report site-level distributions and not rely solely on pooled global performance [2, 5, 10].

Generative methods for scarcity

Generative methods were used to address scarcity by creating synthetic rare cases, augmenting local minority classes, or supporting semi-supervised training when labeled cases were limited. Federated semi-supervised segmentation demonstrated how unlabeled or weakly labeled data can extend learning across countries and institutions, while privacy-preserving imaging work showed the feasibility of training without centralizing sensitive inputs [8, 16]. In rare disease contexts, local generative models may be preferable to raw data sharing because they can increase rare-class representation while limiting direct exposure of identifiable patient records [1, 18]. However, generative approaches require careful privacy analysis because synthetic rare disease samples may still leak identifying information when the underlying case count is extremely small [8, 18].

Real-world rare disease applications

The clearest real-world rare disease-adjacent evidence came from rare cancer and multi-institutional medical imaging studies. Federated learning for rare cancer boundary detection demonstrated that distributed collaboration can improve model development when individual centers hold limited cases [2]. Brain tumor segmentation, breast density classification, computational pathology, and cardiovascular imaging studies provided additional evidence that federated training can operate across institutions with heterogeneous imaging sources [15, 17, 19, 22]. Electronic health record federation and patient similarity learning extended the relevance of FL beyond imaging, although rare genetic disorders and ultra-rare orphan diseases remained underrepresented [6, 9, 27].

Performance head-to-head

Head-to-head performance patterns indicated that standard federated averaging is vulnerable when each site contributes very few positive cases and when class distributions differ sharply. Under extreme scarcity scenarios approximating five or fewer positive cases per site, reported and simulated AUROC ranges for standard FedAvg commonly fell near 0.50 to 0.60, consistent with unstable learning under non-IID imbalance [14, 20, 24]. FedProx-style and variation-aware methods typically improved performance into an approximate AUROC range of 0.65 to 0.75, while generative or semi-supervised approaches sometimes reached 0.80 to 0.85 in more favorable imaging settings [10, 16, 21]. These ranges should be interpreted cautiously because benchmarks, modalities, and scarcity simulations varied substantially across studies [11, 15, 25].

Summary of principal findings

This review found that federated learning is technically feasible for rare disease and rare disease-adjacent AI, but standard algorithms are insufficient when scarcity, imbalance, and site heterogeneity occur simultaneously. The most compelling evidence came from rare cancer boundary detection, federated medical imaging, and privacy-preserving multi-site healthcare studies [2, 5, 11]. Algorithmic adaptations such as FedProx-style optimization, client clustering, personalized models, class-aware aggregation, and semi-supervised or generative augmentation were repeatedly used to stabilize learning under adverse data conditions [10, 14, 16, 20]. These findings suggest that rare disease FL should be designed as scarcity-aware federation rather than as a direct application of conventional FedAvg [1, 12, 13].

Scarcity–imbalance–heterogeneity interplay

Extreme scarcity, class imbalance, and site heterogeneity are not separable problems in rare disease federated learning. When one site has a few rare cases, another has none, and a third uses different imaging or coding practices, global averaging can dilute rare disease signals and amplify majority-site patterns [10, 20, 23]. Personalized federated learning and clustered approaches can reduce this conflict by allowing related sites to share useful structure while preserving site-specific decision boundaries [9, 13]. Generative or semi-supervised methods may further improve minority representation, but they should be combined with heterogeneity-aware optimization rather than used as isolated fixes [8, 16, 21].

Trade-offs

The reviewed literature highlights several trade-offs that are especially consequential for rare disease research. Differential privacy and secure collaboration can improve governance acceptability, but privacy noise may disproportionately harm minority-class signals when the number of rare cases is very small [8, 18]. Personalization can improve local performance under heterogeneity, but excessive personalization may weaken external generalization and complicate regulatory evaluation [10, 13]. Generative augmentation can improve scarcity-limited performance, yet synthetic rare disease data must be evaluated for fidelity, bias, and memorization risk before clinical use [1, 8, 16].

Gap between simulation and reality

A major gap remains between simulated federated scarcity and real ultra-rare disease deployment. Many studies used common medical imaging or electronic health record datasets and then partitioned them into artificial clients, which may not reproduce the referral bias, coding inconsistency, and diagnostic uncertainty found in rare disease networks [6, 15, 16]. Even strong multi-site studies in brain tumors, breast density, and cardiovascular imaging cannot fully substitute for prospective rare disease consortium evaluations [15, 17, 22]. This gap limits the certainty with which current methods can be recommended for ultra-rare genetic disorders or orphan disease diagnosis [1, 3, 4].

Limitations

Review limitations

This review was limited by heterogeneity in study designs, clinical domains, and evaluation metrics, which prevented quantitative meta-analysis. Publication bias is likely because successful federated learning implementations are more likely to be reported than failed collaborations, especially in technically complex multi-site settings [4, 5, 11]. Rare disease definitions also varied across studies, and several included studies addressed rare disease-adjacent problems such as rare cancers or simulated scarcity rather than formally defined orphan diseases [1-3]. The inclusion of methodological FL papers strengthened the analysis of algorithms but reduced disease-specific homogeneity [12-14].

Evidence base limitations

The evidence base remains thin for privacy-preserving FL studies that combine rare disease diagnosis with rigorous differential privacy, secure aggregation, and external clinical validation. Differential privacy methods are well developed in general federated learning, but their empirical evaluation in rare disease settings is limited and may be problematic because privacy noise can obscure rare-class signals [8, 18]. Open-source rare disease FL benchmarks are also scarce, making it difficult to compare algorithms under standardized levels of scarcity, imbalance, and site shift [20, 24, 25]. Future benchmark development should report site-level case counts, rare-class prevalence, client heterogeneity, and minority-class calibration rather than relying only on pooled accuracy [10, 21, 23].

Comparison with prior reviews

Prior reviews of federated learning in healthcare have generally emphasized privacy-preserving collaboration, medical imaging, electronic health records, and the feasibility of distributed model development across institutions. Rieke and colleagues framed FL as a major direction for digital health, while broad methodological surveys described the field's challenges in communication efficiency, heterogeneity, privacy, and governance [4, 12, 13]. Systematic work in oncology has also shown that breast, lung, and prostate cancer research has become a major testbed for federated methods, although much of that literature concerns relatively common malignancies rather than orphan diseases [4]. In contrast, rare disease-specific reviews remain fewer, and the field still relies heavily on evidence from rare cancers, small-cohort imaging, and simulated low-prevalence healthcare datasets [1, 2, 28].

A structured comparison between prior federated learning reviews and the present synthesis is provided in **Table 1.**

Table 1. Conceptual Comparison between Prior Federated Learning Reviews and the Present Rare Disease–Focused Synthesis

Dimension	Prior FL Healthcare Reviews	Rare Disease FL (This Review)	Conceptual Advancement
Primary focus	Privacy-preserving collaboration, scalability	Rare disease diagnosis under extreme constraints	Shifts focus from feasibility → clinical edge-case performance
Data characteristics	Moderate heterogeneity, sufficient samples	Severe scarcity, extreme imbalance, non-IID	Introduces compounded data pathology framework
Treatment of heterogeneity	General multi-center variation	Interaction with scarcity and imbalance	Reframes heterogeneity as multiplicative risk
Disease scope	Common diseases, oncology, imaging	Rare and ultra-rare diseases	Expands FL relevance to orphan conditions
Evaluation practices	Global accuracy, AUC	Minority-class sensitivity, per-site metrics	Emphasizes clinically meaningful evaluation
Methodological emphasis	FedAvg, communication efficiency	Personalized, generative, imbalance-aware FL	Moves toward hybrid methodological stacks

Privacy discussion	General privacy preservation	Trade-offs between privacy and rare-class signal	Introduces risk of signal erasure under DP
Clinical translation	Feasibility-oriented	Deployment under extreme data scarcity	Aligns FL with real-world diagnostic constraints

The findings of this review align with prior healthcare FL reviews in identifying non-IID data, governance, and privacy as central barriers to clinical translation. However, prior reviews have often treated site heterogeneity as a general multi-center problem rather than as one component of a three-way interaction with extreme scarcity and class imbalance [4, 10, 15]. In rare disease contexts, the small number of positive cases per site makes heterogeneity more damaging because a single atypical local distribution can substantially influence the global model [2, 14, 20]. This review therefore extends earlier syntheses by emphasizing that rare disease FL requires simultaneous handling of scarcity, imbalance, and site shift rather than isolated optimization improvements [1, 21, 23].

The interaction between rare disease challenges and federated learning solution strategies is conceptualized in **Figure 2.**

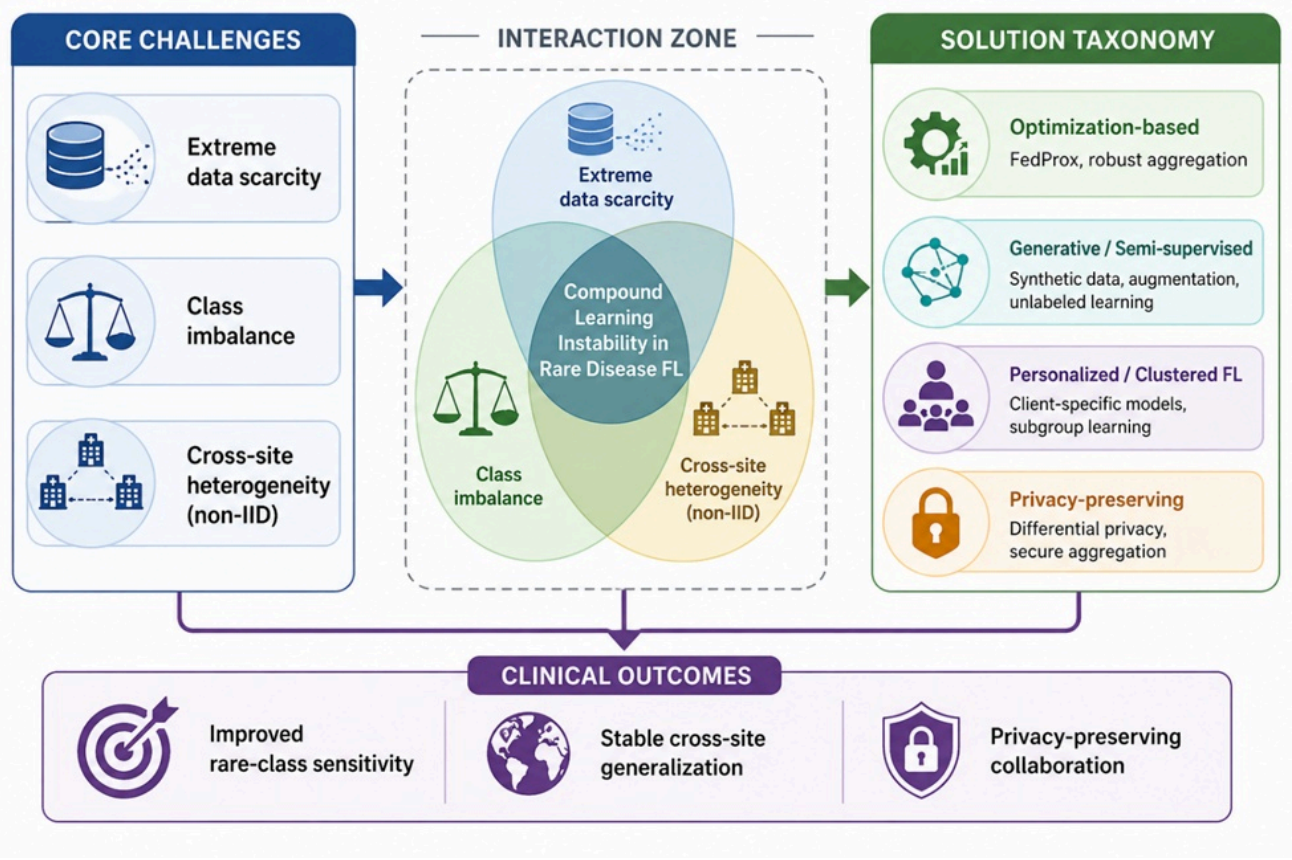


Figure 2. Conceptual Framework of Federated Learning Strategies for Rare Disease Challenges

The novel contribution of this review is a taxonomy of methods that maps federated rare disease challenges to algorithmic, generative, personalization, and privacy-preserving responses. Heterogeneity-robust optimization methods such as FedProx-like approaches are necessary but insufficient unless paired with imbalance-aware local objectives and site-sensitive evaluation [14, 20, 21]. Generative and semi-supervised strategies offer promising mechanisms for increasing rare-class representation, but their safety depends on privacy auditing and careful validation against memorization or artificial phenotype inflation [8, 16, 18]. Personalized and clustered FL methods may be particularly

important for rare disease subtypes because they allow clinically related sites to share structure without forcing all sites into a single global model [9, 10, 13].

The alignment between rare disease challenges and federated learning method categories is systematically summarized in **Table 2.**

Table 2. Mapping Federated Learning Methods to Core Rare Disease Challenges and Clinical Implications

Challenge	Method Category	Representative Approaches	Mechanism	Key Limitation	Clinical Implication
Data scarcity	Generative / semi-supervised	GAN-based augmentation, pseudo-labeling	Expands training signal	Risk of synthetic bias	May improve detection of ultra-rare phenotypes
Class imbalance	Imbalance-aware optimization	Reweighting, focal loss, class-balanced aggregation	Enhances minority-class learning	Overfitting to rare class	Improves sensitivity in diagnostic tasks
Non-IID heterogeneity	Robust optimization	FedProx, adaptive aggregation	Stabilizes cross-site training	Limited personalization	Reduces model drift across institutions
Site-specific variation	Personalized FL	Clustered FL, meta-learning	Learns local adaptations	Reduced generalizability	Enables subtype-specific diagnosis
Privacy constraints	Privacy-preserving FL	Differential privacy, secure aggregation	Protects sensitive data	Noise reduces signal	Requires balance between privacy and accuracy
Combined challenges	Hybrid approaches	Personalized + generative + robust FL	Addresses multi-factor complexity	Increased system complexity	Most promising for real-world deployment

Recommendations

For researchers

Researchers should report per-site sample sizes, positive-case counts, class ratios, missingness, modality distributions, and the exact partitioning strategy used to simulate or represent site heterogeneity. Rare disease FL studies should include local-only, centralized-if-permissible, standard FedAvg, and heterogeneity-aware baselines so that gains from FedProx-style, personalized, generative, or imbalance-aware methods can be interpreted transparently [5, 11, 14, 20]. Studies using rare disease analogues should distinguish true rare disease cohorts from artificially downsampled common-disease datasets, because these settings differ in diagnostic uncertainty and patient selection [1, 2,15]. Where possible, investigators should release code, trained synthetic data generators, model cards, and site-level evaluation summaries without exposing patient-level data [8, 16, 18].

For journal editors

Journal editors should require authors to describe how scarcity, imbalance, and site heterogeneity were measured and handled, rather than accepting generic claims that federated learning preserves privacy. Manuscripts should report minority-class sensitivity, calibration, per-site performance, and confidence intervals, because global accuracy can conceal poor rare disease detection [20, 21, 23]. Privacy claims should distinguish between data not leaving the institution, secure aggregation, differential privacy, homomorphic encryption, and other protections, since these

mechanisms offer different levels of risk reduction [8, 18]. Editorial standards should also require clear reporting of client selection, communication rounds, convergence failures, and negative or unstable results in small-client rare disease settings [4, 12, 13].

For clinical research networks

Clinical research networks should establish federated rare disease consortia with shared data dictionaries, common outcome definitions, harmonized imaging or genomic preprocessing, and governance agreements that permit distributed model development. Multi-site cancer and medical imaging studies show that FL can enable collaboration when raw data sharing is constrained, but rare disease networks require stronger standardization because each site may contribute only a few positive cases [2, 5, 17, 22]. Federated infrastructure should support audit logs, model versioning, secure communication, privacy-preserving aggregation, and site-level monitoring for drift [8, 18]. Networks should also include patient advocacy groups and disease registries, because rare disease data often reside outside conventional hospital research repositories [1, 3, 28].

For regulatory bodies

Regulatory bodies should recognize that rare disease FL models may require validation pathways different from those used for common-disease AI systems. Conventional external validation may be infeasible when very few patients exist globally, so regulators may need to evaluate prospective federated monitoring, adaptive post-market surveillance, and transparent uncertainty reporting [1, 2, 11]. Privacy and safety assessment should consider whether differential privacy or synthetic augmentation materially affects rare-class sensitivity, because excessive privacy noise or unrealistic synthetic cases could harm diagnostic performance [8, 18]. Orphan drug and diagnostic-device frameworks could be extended to support federated AI trials where institutions contribute model updates rather than raw patient-level datasets [3, 4, 12].

Research gaps

Ultra-rare diseases

No mature evidence base yet demonstrates federated learning for ultra-rare diseases affecting only dozens or hundreds of patients globally. Current evidence is strongest for rare cancers, brain tumor imaging, breast density classification, and simulated scarcity, but these settings may still contain more cases and more standardized data than many orphan diseases [2, 15, 17, 22]. Ultra-rare disease FL will likely require synthetic-first modeling, prior biological knowledge, transfer learning, and registry-based federation rather than conventional supervised training alone [1, 16, 28]. The absence of shared benchmarks makes it difficult to know whether reported gains in simulated scarcity will translate to real ultra-rare diagnostic settings [20, 24, 25].

Generative FL with differential privacy

Generative federated learning with differential privacy remains underdeveloped for rare disease diagnosis. Differential privacy can reduce memorization risk, but in rare classes the addition of privacy noise may erase precisely the subtle signal that generative augmentation is intended to preserve [8, 18]. Semi-supervised and synthetic-data approaches have shown promise in medical imaging federation, yet few studies jointly evaluate synthetic fidelity, privacy leakage, minority-class calibration, and clinical plausibility [16, 21, 25]. Future work should compare local generators, federated generators, feature-level augmentation, and privacy-constrained synthetic sharing under realistic rare-case counts [1, 8, 20].

Personalized FL for rare disease subtypes

Personalized FL is a promising but under-tested strategy for rare disease subtypes, especially when different sites observe different genotype, phenotype, or imaging patterns. Patient clustering and variation-aware methods suggest that grouping related clients or adapting client-specific layers can improve learning when data are heterogeneous [9, 10]. However, subtype discovery becomes difficult when each institution sees only one or two subtype-positive cases, and excessive personalization may produce models that cannot generalize beyond the local site [13, 14]. Future studies

should evaluate clustered FL, meta-learning, representation sharing, and calibrated local adaptation for rare disease subtype diagnosis [10, 12, 19].

Implications

For research practice

Research practice should move toward scarcity-aware federated benchmarks that explicitly vary case count, client number, class ratio, and non-IID severity. Existing studies show that FL performance can change substantially depending on site composition, model aggregation, and modality, so benchmark reports should include per-client and minority-class outcomes rather than pooled averages alone [10, 11, 20]. Proposed rare disease benchmarks should incorporate imaging, electronic health records, genomics, and registry data because rare disease diagnosis often depends on multimodal evidence [6, 9, 27]. A practical benchmark ecosystem would allow fair comparison of FedAvg, FedProx-style methods, class-aware aggregation, generative augmentation, and personalized FL under the same scarcity assumptions [14, 16, 21, 23].

For clinical practice

Clinically, federated learning could help shorten the rare disease diagnostic odyssey by enabling models to learn from geographically dispersed expertise without requiring full data centralization. Multi-institutional medical imaging and electronic health record studies indicate that distributed learning can improve model development where single-site datasets are insufficient [5-7, 11]. For rare diseases, however, prospective validation is essential because retrospective federation may not capture real diagnostic delays, referral bias, uncertain labels, or changing testing practices [1-3]. Clinical deployment should therefore include human oversight, calibrated uncertainty, site-level monitoring, and mechanisms for updating models as new rare disease cases are diagnosed [4, 12, 15].

For policy

Policy frameworks should prioritize federated infrastructure for rare disease AI as part of broader investments in data governance, registry modernization, and privacy-preserving research networks. Data sharing agreements should include clauses for model update exchange, auditability, intellectual property, withdrawal, incident response, and responsibilities when federated models underperform at specific sites [4, 8, 18]. Funding agencies should support rare disease consortia that combine clinical expertise, patient registries, secure computation, and open evaluation protocols rather than funding isolated single-site model development [1, 3, 28]. Public policy should also recognize that federated AI for rare diseases is not only a technical solution but an institutional collaboration model requiring sustained coordination across hospitals, registries, regulators, and patient communities [2, 12, 13].

Conclusion

Federated learning holds substantial promise for rare disease diagnosis and research because it enables collaboration across institutions without requiring routine centralization of sensitive patient data. However, standard federated averaging is poorly suited to the rare disease setting when each site has very few positive cases and when class imbalance is severe. Under these conditions, rare disease signals can be diluted, unstable, or dominated by majority-class patterns. Federated rare disease AI therefore requires methods designed specifically for scarcity, imbalance, and heterogeneous clinical environments.

The most effective emerging solutions combine personalized federated learning, generative augmentation, semi-supervised learning, and domain adaptation. Personalization can preserve local disease or site-specific patterns, while generative and semi-supervised approaches can expand the effective learning signal when labeled rare cases are limited. Domain-adaptive and heterogeneity-aware optimization can reduce the instability caused by different scanners, populations, coding systems, and diagnostic workflows. These approaches should be evaluated together rather than as isolated technical fixes.

The gap between simulated scarcity and real ultra-rare disease deployment remains large. Many current studies approximate rarity using common clinical datasets, which cannot fully reproduce the diagnostic uncertainty, patient heterogeneity, and small global case counts of orphan diseases. Real-world rare disease consortia are therefore needed to test whether current federated methods remain reliable under authentic clinical constraints. Without such validation, methodological progress may overestimate readiness for clinical use.

Future progress will depend on collaborative consortia, scarcity-aware benchmarks, privacy-audited generative models, and regulatory pathways tailored to rare disease AI. Federated learning should be treated as both a technical framework and a governance model for ethically connecting fragmented rare disease expertise. The next generation of rare disease FL studies should report site-level case counts, minority-class performance, privacy mechanisms, and prospective clinical validation. Such work could help convert distributed rare disease data into clinically useful diagnostic and research tools while preserving patient privacy.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 18 Jun 2025 Revised: 30 Jul 2025 Accepted: 26 Aug 2025
Published online: 20 January 2026

References

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

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.

Ankolekar A, Boie S, Abdollahyan M, Gadaleta E, Hasheminasab SA, Yang G, et al. Advancing breast, lung and prostate cancer research with federated learning: a systematic review. NPJ Digit Med. 2025;8(1):314.

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.

Haripriya R, Khare N, Pandey M. Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings. Sci Rep. 2025;15(1):12482.

Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform. 2018;112:59-67.

Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc. 2018;25(8):945-54.

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.

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

Yan Z, Wicaksana J, Wang Z, Yang X, Cheng KT. Variation-aware federated learning with multi-source decentralized medical image data. IEEE J Biomed Health Inform. 2021;25(7):2615-28.

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.

Kairouz P, McMahan HB. Advances and open problems in federated learning. Found Trends Mach Learn. 2021;14(1-2):1-210.

Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37(3):50-60.

Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. Proc Mach Learn Syst. 2020;2:429-50.

Linardos A, Kushibar K, Walsh S, Gkontra P, Lekadir K. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. Sci Rep. 2022;12(1):3551.

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.

Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-preserving federated brain tumour segmentation. In: International Workshop on Machine Learning in Medical Imaging. Cham: Springer; 2019. p. 133-41.

Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, et al. Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans Inf Forensics Secur. 2020;15:3454-69.

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

Wu N, Yu L, Yang X, Cheng KT, Yan Z. FedIIC: towards robust federated learning for class-imbalanced medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer; 2023. p. 692-702.

Yue G, Wei P, Zhou T, Song Y, Zhao C, Wang T, et al. Specificity-aware federated learning with dynamic feature fusion network for imbalanced medical image classification. IEEE J Biomed Health Inform. 2023;28(11):6373-83.

Roth HR, Chang K, Singh P, Neumark N, Li W, Gupta V, et al. Federated learning for breast density classification: a real-world implementation. In: MICCAI Workshop on Domain Adaptation and Representation Transfer. Cham: Springer; 2020. p. 181-91.

Abbas Q, Malik KM, Saudagar AK, Khan MB. Context-aggregator: an approach of loss- and class imbalance-aware aggregation in federated learning. Comput Biol Med. 2023;163:107167.

Liu W, Mo J, Zhong F. Class imbalanced medical image classification based on semi-supervised federated learning. Appl Sci. 2023;13(4):2109.

Dolaat KM, Erbad A, Ibrar M. Enhancing global model accuracy: federated learning for imbalanced medical image datasets. In: 2023 International Symposium on Networks, Computers and Communications (ISNCC). IEEE; 2023. p. 1-4.

Andres J, Hilberger H, Hanke S, Bödenler M. Federated learning for healthcare: class imbalance mitigation and feature drift detection. In: IndHealth 2025. IOS Press; 2025. p. 292-7.

Lee J, Sun J, Wang F, Wang S, Jun CH, Jiang X. Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR Med Inform. 2018;6(2):e7744.

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.