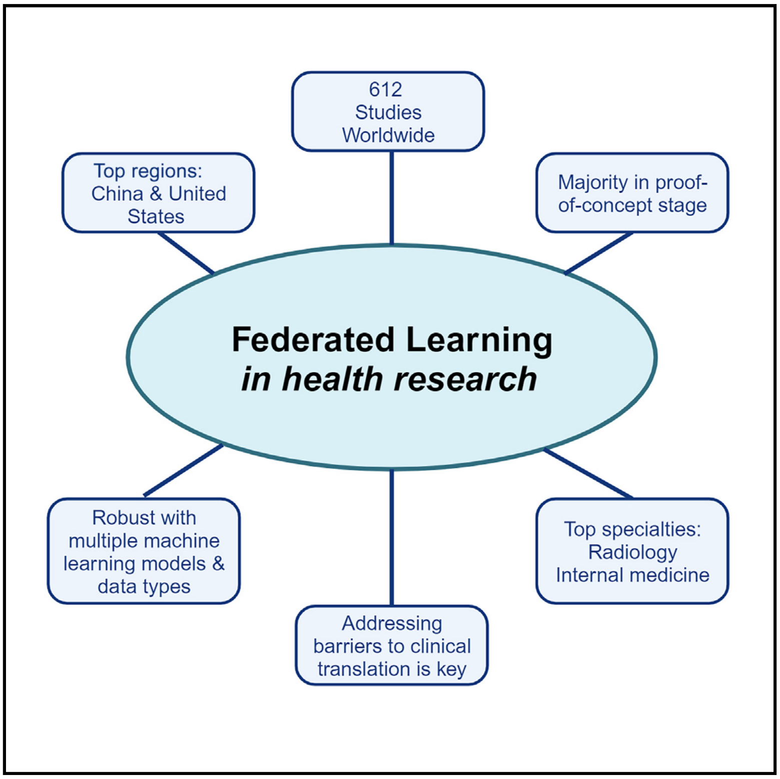
Cell Reports
Medicine

Article

Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture

Graphical abstract



Highlights

- This systematic review of federated learning (FL) in health includes 612 articles
- Only 5.2% are studies with real-life application of FL
- A variety of clinical specialties use FL, with radiology being the most common
- FL is compatible with various data types and machine learning models

Authors

Zhen Ling Teo, Liyuan Jin, Siqi Li, ...,
Yong Liu, Rick Siow Mong Goh,
Daniel Shu Wei Ting

Correspondence

daniel.ting@duke-nus.edu.sg

In brief

Teo et al. provides a comprehensive systematic review of federated learning (FL) in health and includes 612 articles. FL is compatible with multiple data types, machine learning models, and privacy-enhancing technologies. Addressing key barriers to clinical translation will enable FL to be a pivotal strategy for global health collaboration.

Teo et al., 2024, Cell Reports Medicine *5*, 101419
February 20, 2024 © 2024
https://doi.org/10.1016/j.xcrm.2024.101419

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture

Zhen Ling Teo,[1,2,6] Liyuan Jin,[2,3,6] Siqi Li,[2,3] Di Miao,[2,3] Xiaoman Zhang,[2,4] Wei Yan Ng,[1,2] Ting Fang Tan,[1,2] Deborah Meixuan Lee,[2,5] Kai Jie Chua,[1,2] John Heng,[1,2] Yong Liu,[4] Rick Siow Mong Goh,[4] and Daniel Shu Wei Ting[1,2,3,7,*]

[1]Singapore National Eye Centre, Singapore, Singapore
[2]Singapore Eye Research Institute, Singapore, Singapore
[3]Duke-NUS Medical School, Singapore, Singapore
[4]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[5]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore, Singapore
[6]These authors contributed equally
[7]Lead contact
*Correspondence: daniel.ting@duke-nus.edu.sg

https://doi.org/10.1016/j.xcrm.2024.101419

SUMMARY

Federated learning (FL) is a distributed machine learning framework that is gaining traction in view of increasing health data privacy protection needs. By conducting a systematic review of FL applications in healthcare, we identify relevant articles in scientific, engineering, and medical journals in English up to August 31st, 2023. Out of a total of 22,693 articles under review, 612 articles are included in the final analysis. The majority of articles are proof-of-concepts studies, and only 5.2% are studies with real-life application of FL. Radiology and internal medicine are the most common specialties involved in FL. FL is robust to a variety of machine learning models and data types, with neural networks and medical imaging being the most common, respectively. We highlight the need to address the barriers to clinical translation and to assess its real-world impact in this new digital data-driven healthcare scene.

INTRODUCTION

With the explosion of big data, rapid development in machine learning, and increasing global connectivity, collaborative training of machine learning models across different organizations and countries is at an unprecedented level.[1] The biggest concern with collaborative training in healthcare is data privacy concerns, which limit data sharing and the clinical implementation of what is technically possible.[2–4] Hence, there is increasing focus on privacy-preserving approaches such as federated learning (FL), blockchain technology, and generative adversarial networks. FL is a distributed machine learning framework introduced by Google in 2016[5] that allows for multi-party collaboration while preserving data privacy. It has been gaining traction in the medical field as an attractive privacy-enhanced alternative to traditional centralized training.[1,6] FL has been shown to remain robust to multiple data types including imaging, such as magnetic resonance imaging (MRI) for brain tumor segmentation[7]; chest X-ray for COVID-19 clinical outcome prediction[8]; electronic medical records for predicting hospitalization[9]; colored photographs, such as skin photographs in skin lesion diagnostics[10] and retinal fundus photographs[11]; and histology slides, such as in cancer diagnosis,[12] genomics,[13] and Internet of Medical Things.[14]

In the original FL framework, a centralized server broadcasts initial weights of the global model parameters to a set of selected participating site. Each participating site then trains a local model (shares the same model architecture with the global model) using its local data and sends updated model parameters to the centralized server. In this setup, raw data never leave the local device or site, thus addressing the shortcoming of traditional centralized learning where raw data are transferred to a central body (Figure 1). Once all participating sites have sent their updates, the server then aggregates them to update the global model weight. The updated global model is then broadcasted to a new set of participating sites for another round of local training processes. The process is repeated until the global model converges. The main advantage of FL is the generation of a higher-quality model by leveraging larger training datasets obtained from multiple sources, beyond what could have been achieved with the data of a single device or a system, while maintaining high levels of data privacy. The framework also minimizes data aggregation costs. In addition, it has been shown to remain robust in scenarios where clients have an uneven amount of data and non-independent and identically distributed (IID) data.[6] This makes FL an appealing machine learning subfield in the healthcare domain and especially advantageous in niche research areas where publicly available data are limited or restricted.

Despite its advantages, FL has yet to achieve widespread clinical adoption, and efforts to increase clinical translation are ongoing.[15] Aside from being a relatively newer privacy-preserving

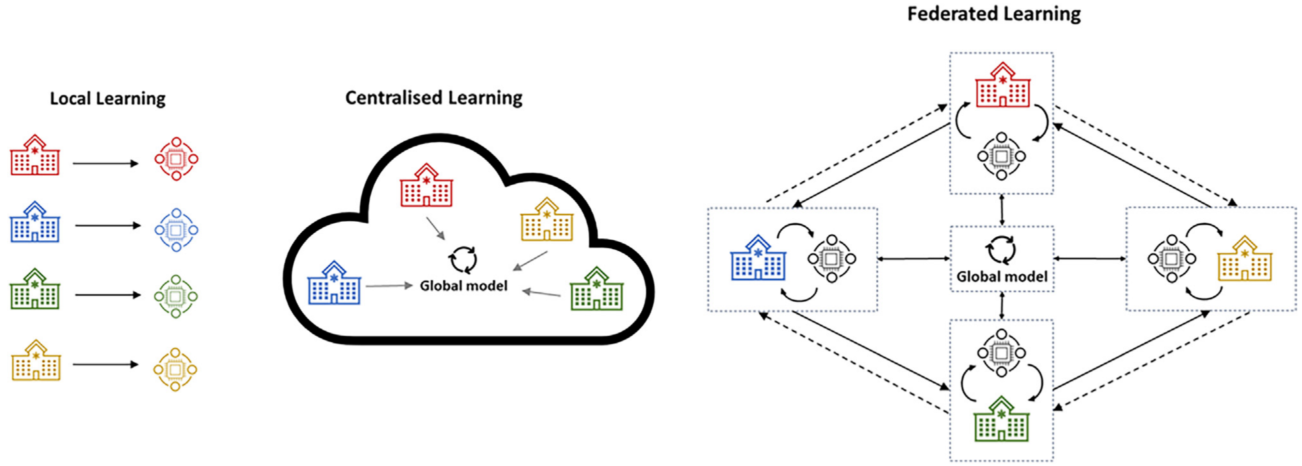


Figure 1. Federated learning in comparison to local learning and centralized learning

technology, research assessing the robustness of FL across a breadth of clinical domains and comparisons with existing machine learning frameworks are still ongoing.[16] In addition, while FL provides greater privacy protection, the sharing of model updates still represents a potential source of privacy compromise. This has led to the addition of further privacy mechanisms such as differential privacy and cryptographic methods (homomorphic encryption, secure multi-party computation, and blockchain) in newer FL models.

Hence, in this study, we perform a comprehensive systematic review on the current applications of FL in the healthcare domain, evaluate the barriers to widespread clinical adoption, and provide insights into future directions of FL-related health research.

Methods

Systematic review

We conducted a systematic review to determine the current applications of FL, in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses guidelines.[17] (Table S1). We aimed to identify all original research articles focusing on the use of FL in the healthcare domain. At the time of systematic review, there were not yet universal reporting guidelines for FL research, especially in the healthcare domain; thus, search definitions were kept broad to avoid omission of relevant articles.

Search strategy and selection criteria

A systematic search using PubMed, Medline, Web of Science, Scopus, Embase, Institute of Electrical and Electronics Engineers Xplore, ArXiv, Springerlink, CINAHL, ACM Digital Library, and Google Scholar was conducted. We included publications in the English language, from inception up to August 31st, 2023. Within medical databases such as PubMed (including Medline) and Embase, the search was conducted using Boolean operators "AND/OR" and a combination of the keywords "federated learning" and "decentralized machine learning" in the title, abstract, and medical subject headings (Table S2). The CINAHL database included nursing and allied health journals, and the search term "federated learning [all fields]" OR "decentralized machine learning [all fields]" was kept broad in view of a relatively smaller database.

Searches within non-medical databases such as Web of Science, Scopus, IEEE Xplore, ArXiv, ACM Digital Library, and Springer Link included a combination of terms such as "federated learning [title/abstract]" AND "healthcare [in text] OR "medical [in text]" (Table S2). In addition, we searched Google Scholar to identify original research articles located in gray literature. The following search term was used in Google Scholar database: "federated learning" OR "distributed machine learning" AND "health" OR "medical." Further literature search consisted of reviewing the reference lists of relevant articles. This adopted strategy identified all articles used in previous reviews.[18–20]

Inclusion and exclusion criteria for article selection

We included studies with the following criteria: (1) FL, (2) healthcare or medical related, (3) English language, and (4) original research articles proposing FL applications in the healthcare domain with or without prototype development. We excluded studies that were (1) duplicates, (2) surveys, opinions, editorial letters, book chapters, thesis, or conference proceedings, (3) not peer reviewed, and (4) did not have full text available. The article selection process is detailed in Figure 2. First, duplicate articles were removed with the use of Rayyan. Second, titles and abstracts were screened, and irrelevant articles were excluded. Third, the full texts of the remaining articles were assessed for eligibility. Based on the above criteria, two reviewers (Z.L.T. and L.J.) independently selected the studies for final inclusion. Disagreements between the two were resolved and adjudicated by the senior author (D.S.W.T.).

Data abstraction

A standardized data extraction spreadsheet using Microsoft Excel was used to collect information on each article. To ensure consistency in data abstraction, a data abstraction pilot was first performed by both reviewers. A list of studies consisting of 20% of all identified full-text reports was created with a computer-generated random sequence. Independent data charting was done by both reviewers for the randomly selected studies, and

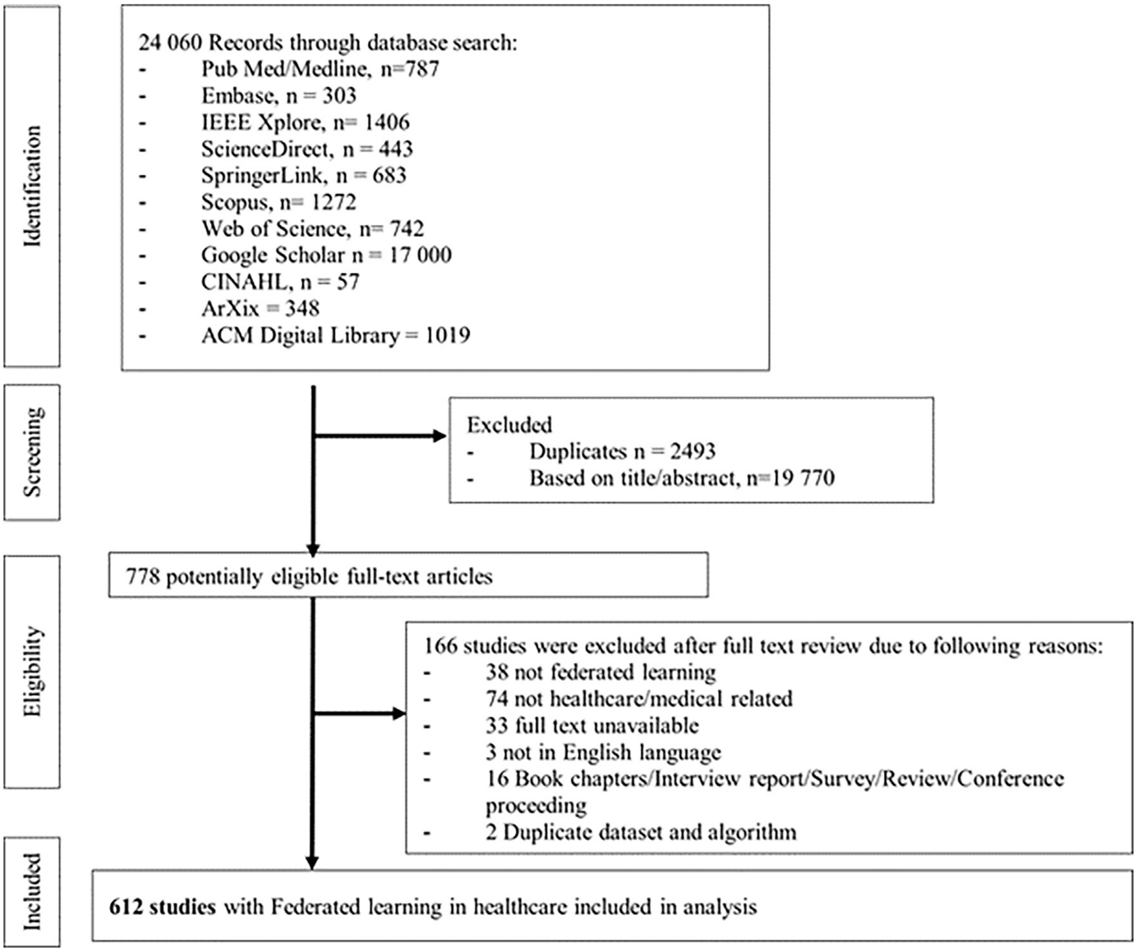

Identification
24 060 Records through database search:
- Pub Med/Medline, n=787
- Embase, n = 303
- IEEE Xplore, n= 1406
- ScienceDirect, n = 443
- SpringerLink, n = 683
- Scopus, n= 1272
- Web of Science, n= 742
- Google Scholar n = 17 000
- CINAHL, n = 57
- ArXix = 348
- ACM Digital Library = 1019
Screening
Excluded
- Duplicates n = 2493
- Based on title/abstract, n=19 770
Eligibility
778 potentially eligible full-text articles
166 studies were excluded after full text review due to following reasons:
- 38 not federated learning
- 74 not healthcare/medical related
- 33 full text unavailable
- 3 not in English language
- 16 Book chapters/Interview report/Survey/Review/Conference proceeding
- 2 Duplicate dataset and algorithm
Included
612 studies with Federated learning in healthcare included in analysis

Figure 2. Summary of article selection process

the results of data extraction were subsequently discussed. Disagreements were resolved by consensus between the two reviewers and, if required, adjudicated by the senior author (D.S.W.T.). For each included article, we extracted the following data: publication year, article type (technical design with or without prototype, real-life clinical application studies, FL adjuncts, reviews), data type utilized (imaging, electronic health records, histology slides, genomic or molecular data, colored photographs, Internet of Things), COVID-19 relevance, technical architecture information, and privacy mechanisms used. For studies with real-life clinical application of FL, we further extracted data on sample size, number of participating sites, levels of collaboration (international, regional, or local), clinical specialty, and the role of regulatory bodies involved, if any.

In detail, we defined an article to be of technical design type if it included a description of the FL model built on a framework with or without a prototype. Studies that artificially split the data into artificial silos to simulate a multi-site FL setting were considered to be of technical design with a prototype. Studies that applied FL to train a global model among multiple sites, without actual sharing or pooling of data, were considered as real-life clinical application studies. Studies that focused on testing adjuncts to an FL model such as proposing a new platform for FL, testing of verifiable credentials, or post-FL-processing modifications were included as FL adjuncts.

In this study, we defined FL collaboration as international if participating sites were from different countries/regions, regional if participating sites were from different city or states within a country, and local if participating sites were institutions or devices within a single city or state.

Technical architecture information included machine learning models, data partitioning, and communication architecture. Machine learning models were classified as neural networks (including artificial neural networks, recurrent neural networks, and convolutional neural networks), support vector machines, decision tree models, linear or logistic regression models, unsupervised, and others. FL models were also classified as horizontal, vertical, or transfer learning according to how the data are partitioned. In horizontal FL data partitioning, data of different parties (different sample) share the same feature space. In vertical FL data partitioning, datasets have similar sample space but differ in feature space. Lastly, federated transfer learning is used

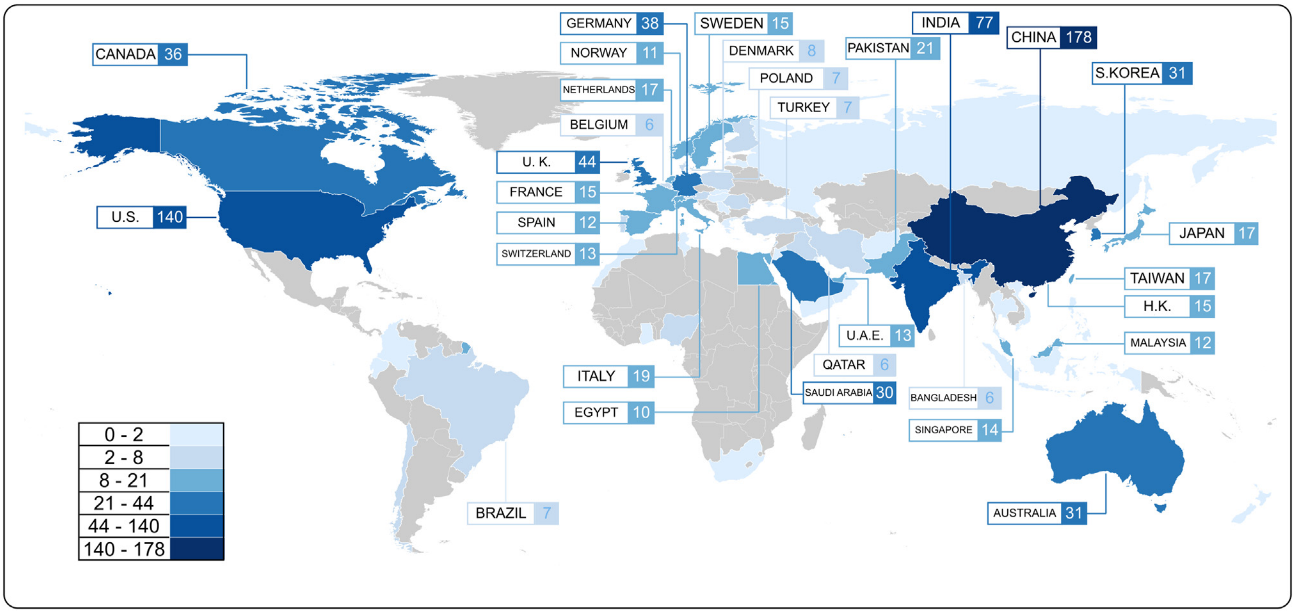


Figure 3. Global map of federated learning health-related research

when datasets differ in both sample and feature spaces, which is often seen in collaborative multi-institution health research. We also classified FL models according to the two major communication architecture types: centralized and decentralized. In a centralized communication architecture, a central server receives local model updates from each party, updates the parameters on the global model, and sends the training results back to each client. In a decentralized communication architecture, communications are performed among clients, with each client able to directly update the global model.

Data abstraction findings were categorized according to the article type and technical architecture. In view of heterogeneity between studies and the lack of pre-existing reporting guidelines for FL in health, findings are presented in a narrative review approach with descriptive statistics, and meta-analysis was deemed unsuitable. Furthermore, the majority of articles were published in engineering journals with a focus on technical design of FL models, thus precluding the assessment of risk of bias within and across included studies typically performed in a medical systematic review and meta-analysis.

RESULTS

Figure 2 shows the article selection process for studies included in the review. In brief, a total of 24,060 individual records were identified during the initial search, of which 778 articles were selected after title and abstract screening. After further reviewing the full text of these selected articles, 612 articles were included in the final analysis (Table S3).

Global trend of FL in health research

Across the world, FL research was conducted in 64 regions, with China and the US having the highest numbers of studies at 178 and 140, respectively (Figure 3). The majority of articles were published in multi-disciplinary scientific journals (43.6%, 267 out of 612), followed by engineering and computer science journals such as IEEE Xplore and ACM Digital Library (34.8%; 213 out of 612) and, lastly, medical journals (21.6%: 132 out of 612). All articles included were found to be published after 2018 with an exponential trend (Figure 4) in the number of relevant articles across the last few years, with 1 article (0.2% of all included articles in this analysis) published in 2018, 10 (1.6%) in 2019, 38 (6.2%) in 2020, 124 (20.3%) in 2021, 186 (14.1%) in 2022, and 253 (41.3%) published in the first three-quarters of 2023 alone. Even though the majority of articles were published after 2020, only a minority (16.0%) were COVID-19 related.

Out of all studies, only a minority of 5.2% (32 out of 612) were studies with real-life application of FL. A large majority of articles provided a technical FL framework with a prototype (65.0%, 398 out of 612) or without a prototype (17.0%, 104 out of 612). For the remaining articles, 3.8% (23 out of 612) were on FL adjuncts, and 9.0% (55 out of 612) were review papers on the use of FL in health.

Among all articles excluding reviews, a wide variety of data types were utilized for FL, as shown in an evidence map (Figure 5)—medical imaging data (including MRI, computed tomography, ultrasound, and X-rays) were the most common data type (41.7%), followed by clinical data and electronic medical records (23.7%), Internet of Things (13.6%), data from colored images (such as skin photographs and digital retinal photographs) (3.6%), histology slides (2.7%), genomic data (2.3%), or a combination of the above (6.8%). 1.6% of the studies did not provide information on the data type used. FL was conducted among a large variety of medical specialties, with radiology and internal medicine being the most common. Figure 6 shows the distribution of studies across data types and medical specialties in FL health research.

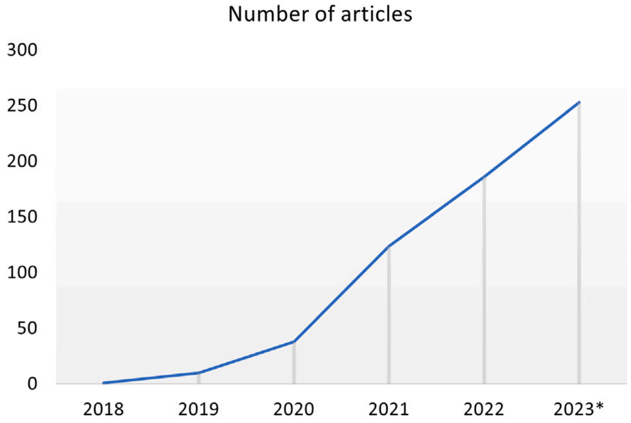


Figure 4. Trend of FL article by publication year
*Up to August 31st, 2023.

Technical architecture of studies

With regards to the technical architecture of the FL framework, a large variety of machine learning models were used across studies, excluding reviews (n = 577). Neural networks were the most popular machine learning model, with 76.3% (440 out of 57) of the included studies using some form of neural network in its FL model (Figure 7). Among neural networks, 67.7% used convolutional neural networks, 24.8% used artificial neural networks such as multi-layer perceptron and feedforward neural networks, and 7.5% used recurrent neural networks. Linear or logistic regression models were the second most common type, used in 46 out 577 (8.0%) studies. 3.5% of studies used decision trees, 3.5% used support vector machines, and 1.9% used other machine learning models including unsupervised methods. Out of the three forms of data partitioning in FL, horizontal data partitioning was the most common and was used in 511 studies; 23 studies used vertical data partitioning, and 24 studies used transfer learning (Figure 7). A large majority of the articles (83.7%, 483 out of 577) used a centralized communication architecture, while 10.6% (61 out of 577) were decentralized (Figure 7). Out of the 577 studies with at least a technical design, 160 (27.7%) included at least one privacy mechanism such as differential privacy (59 studies), homomorphic encryption (29 studies), secure multi-party computation (12 studies), or blockchain (60 studies).

While the majority of studies focused on presenting the technical details of FL, 34.3% (198 out of 557) of the studies compared the FL model with the centralized learning model. The majority of these studies found FL models to have comparable results with their centralized counterpart.

Studies with real-life application of FL

An additional analysis among the 32 studies with real-life application of FL was performed to evaluate the efficacy of FL in real-world settings and to provide greater insight into the clinical translation of FL. More than half of the studies involved international or regional collaboration. Regional collaboration was the most common, with 14 studies (43.8%) performing FL among sites from different cities and states. International collaboration was seen in 11 studies, and four studies featured local collaboration, while three studies kept details on geographical location of their participating sites private. The number of participating sites or devices in each study ranged from 2 to 314 and varied in technical architecture (Table S3). A wide variety of clinical specialties were involved in these FL studies, with the top three specialties involved being radiology followed by oncology and urology. Nearly all studies were intended for clinical application except for three studies with a pharmaceutical development application.[21–23]

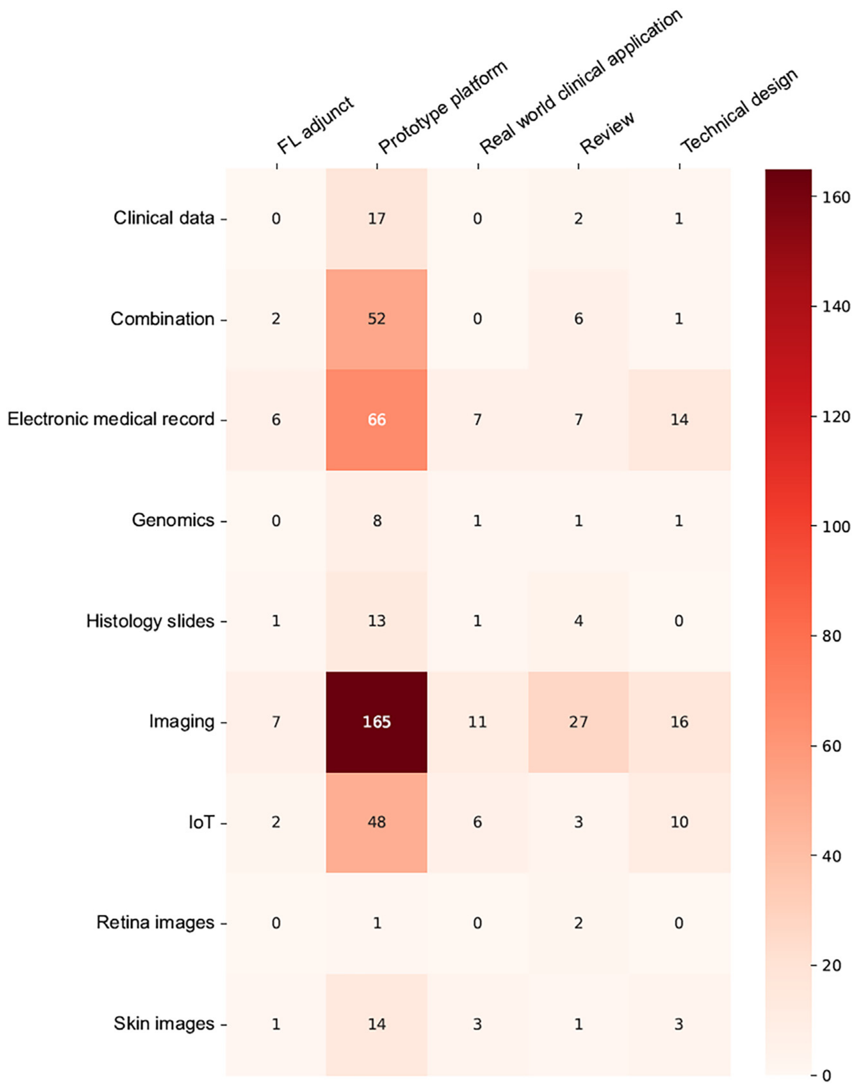
Neural network was the most common machine learning model (81%.3) among these studies with real-life application of FL, while the remaining minority used logistic regression and support vector machine. A variety of machine learning frameworks were used, with the most common being PyTorch (9 out of 32) followed by Tensorflow (8 out of 32). Three studies used Keras, two of which were used in conjunction with Tensorflow. Two studies that did not use neural networks used MATLAB. Ten studies did not name the machine learning framework used. Details on the cloud service were limited, with three studies using Google Cloud, three using Amazon Web Services, one using Microsoft Azure, and one using a private cloud service.

In most FL frameworks, especially that of a centralized communication architecture where a central server is involved, aggregation of model parameters should be supervised by regulatory authority. Within the federation, all clinical studies appointed a hospital trust network or a single institution leading the FL training as the regulatory authority. However, details on the regulatory role were limited, and none of the studies described inventorship contribution or intellectual property rights distribution between different participating sites.

DISCUSSION

This systematic review provides a summary of the current state of research on FL in healthcare. The majority of studies were found in engineering and science journals, with a minority of 21.6% found in medical journals. Importantly, only 5.2% of studies included real-world clinical application of FL, suggesting that the clinical use of FL in the healthcare domain is still in its relative infancy. This is supported by the exponential rise in health-related FL studies over the years (Figure 3), suggesting that FL in the medical domain is increasingly gaining popularity. COVID-19 provided a strong push toward digitalization of health data and demonstrated the utility of FL in a global health crisis. Early works by Dou et al.[24] and Bai et al.[25] showed that FL allowed for secure multi-site collaboration for COVID-19 research, especially in early stages where COVID-19 data were scarce. A few years after the onset of COVID-19, there remain an exponentially rising number of non-COVID-19-related FL studies, suggesting that the privacy-enhanced data-sharing strength of FL has widespread application beyond COVID-19. FL will likely remain relevant in the post-pandemic global health scene.

Our review showed that FL remained robust to a variety of machine learning models, with neural networks, especially convolutional neural networks, being the predominant model used. This correlates with the predominant data type used in current FL research, imaging, where deep convolutional neural networks



	FL adjunct	Prototype platform	Real world clinical application	Review	Technical design
Clinical data	0	17	0	2	1
Combination	2	52	0	6	1
Electronic medical record	6	66	7	7	14
Genomics	0	8	1	1	1
Histology slides	1	13	1	4	0
Imaging	7	165	11	27	16
IoT	2	48	6	3	10
Retina images	0	1	0	2	0
Skin images	1	14	3	1	3

Figure 5. Evidence map on the data variety used across article types

have been shown to given excellent performance.[26] In addition, the large variety of data types—both structured and unstructured data—successfully used in FL models is encouraging.

Key strengths and limitations

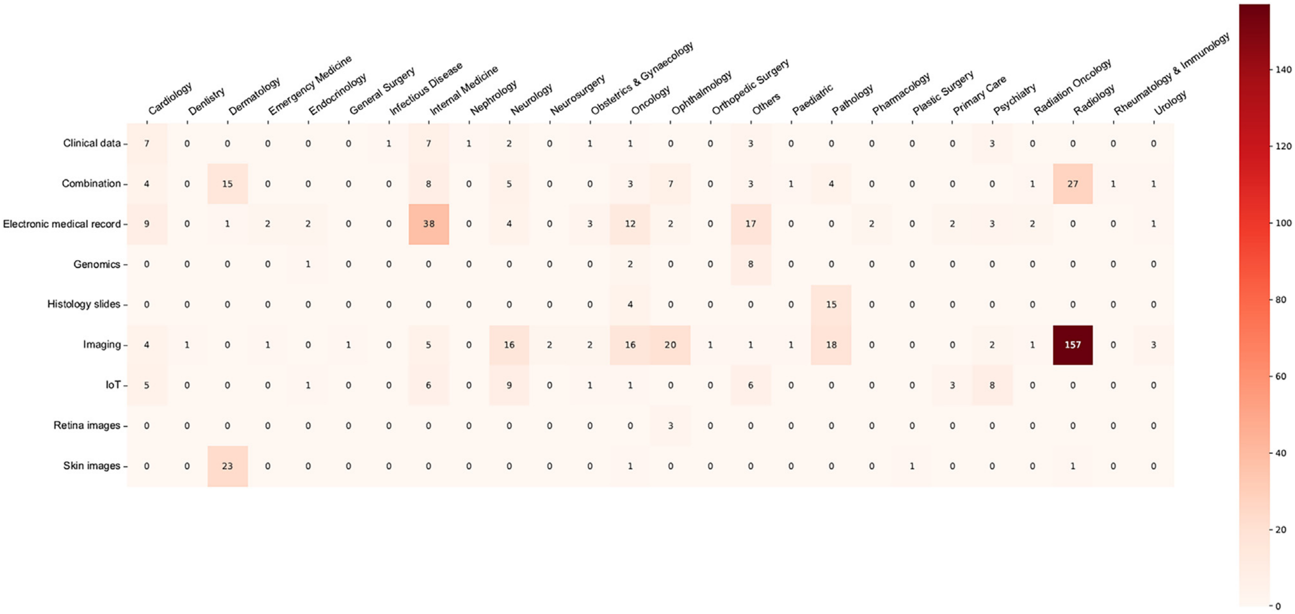
This review has a few strengths. First, it provides a comprehensive systematic review of FL with the inclusion of a large number of studies. Second, the database search included not only medical journals but spanned across engineering, scientific, and generic journals, where many FL models designed for healthcare were found. This is especially important given the relative infancy of FL, where technical papers are the large majority. Third, the classification of different studies allows for the identification of areas that require further research to increase the real-world clinical application of FL. Limitations of this review include the lack of standardized research reporting, especially on the technical components of the FL framework, thus limiting comparison across studies precluding useful meta-analysis. We attempted to reduce bias in the review process by having training and agreement between the two independent reviewers based on an initial pool of articles. Subsequently, a standardized data collection form was utilized with standardized definitions of terms, especially regarding technical architecture and results. In addition, non-peer-reviewed literature was excluded during the article selection process to increase robustness of the review amid balancing contemporary insights into the current state of healthcare-related FL.

Current barriers to clinical implementation and future directions

The limited number studies on real-world clinical application of FL suggests that clinical implementation of FL is not yet widespread. There are several barriers to clinical adoption of FL that must be addressed.

First, FL models are still at risk of privacy breaches. While raw data remain private in FL, model updates exchanged during the training process can still reveal sensitive health information. FL models are still susceptible to reconstruction attacks (where model parameters are used to infer original datasets) or membership inference attacks (where attackers used model predictions to infer if a specific individual data record was included in the training dataset). Countermeasures such as the use of differential privacy where noise is added to prevent inference attacks and homomorphic encryption to allow computation on encrypted data without decryption will confer higher levels of data security. While raw data are not transferred, FL is still susceptible to privacy leakages during the transfer of model updates and weights between nodes and servers. The addition of blockchain can further enhance security during the transfer of model updates governed by smart contracts, allowing for full decentralization and enhanced security. Beyond and prior to the medical setting, FL is widely used in other commercial sectors including banking and finance, manufacturing, energy industry, and ecommerce with guidelines on the use of machine learning models in trusted research environments[27] and rapid updates in the FL framework to enhance security and performance. These updated FL strategies such as defense to model-poisoning attacks and adversarial attack simulations can be applied to the medical domain as well.

Second, in addition to the use of non-IID health data, the inability to directly check and “clean” in an FL framework may



	Cardiology	Dentistry	Dermatology	Emergency Medicine	Endocrinology	General Surgery	Infectious Disease	Internal Medicine	Nephrology	Neurology	Neurosurgery	Obstetrics & Gynaecology	Oncology	Ophthalmology	Orthopedic Surgery	Others	Paediatric	Pathology	Pharmacology	Plastic Surgery	Primary Care	Psychiatry	Radiation Oncology	Radiology	Rheumatology & Immunology	Urology
Clinical data	7	0	0	0	0	0	1	7	1	2	0	1	1	0	0	3	0	0	0	0	0	3	0	0	0	0
Combination	4	0	15	0	0	0	0	8	0	5	0	0	3	7	0	3	1	4	0	0	0	0	1	27	1	1
Electronic medical record	9	0	1	2	2	0	0	38	0	4	0	3	12	2	0	17	0	0	2	0	2	3	2	0	0	1
Genomics	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	8	0	0	0	0	0	0	0	0	0	0
Histology slides	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	15	0	0	0	0	0	0	0	0
Imaging	4	1	0	1	0	1	0	5	0	16	2	2	16	20	1	1	1	18	0	0	0	2	1	157	0	3
IoT	5	0	0	0	1	0	0	6	0	9	0	1	1	0	0	6	0	0	0	0	3	8	0	0	0	0
Retina images	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
Skin images	0	0	23	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0

Figure 6. Evidence map on the distribution of article types across medical subspecialties

compromise model performance. It is known that quality of data—in the form of large, well-labeled, and diverse datasets—is a key determinant of machine learning model performance, especially in real-world settings.[28] Research to explore solutions such as data normalization before training or adapting the FL model to heterogeneous data is ongoing. Third, explainability of the FL model may be challenging due to the lack of direct access to original data. Explainability is key in medical diagnostics and is essential for clinician and patient acceptance of the FL model. Fourth, the FL framework requires that all local participating sites have the necessary infrastructure and computational capabilities for model training as well as access to the communication network for model update sharing. In an FL setting, the central server does not have control over individual nodes, which may go offline either intentionally or due to unreliable infrastructure and networks. This may be disruptive to the model training process. Several articles have suggested modifications to the FL framework such as the central server selecting clients based on availability and resources[29,30] or the use of blockchain for decentralization and avoidance of single-point failure problems.[31]

Fifth, self-sufficient sites with large amounts of diverse data may feel little incentive to participate in the federation. The ability to weigh contributions of each participating site and design a fair incentive mechanism will be required to attract participants with high-quality data to join the FL network. In addition, intellectual

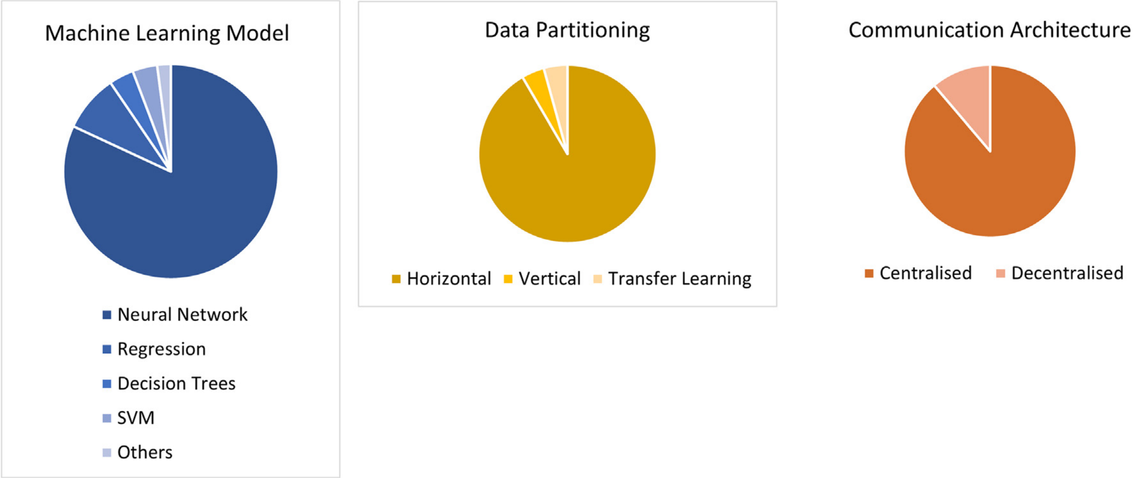


Figure 7. Technical architecture of studies

property considerations have not yet been well addressed. During the FL training process, technology transfer is inevitable, and intellectual property is important to determine ownership and further incentivize participation. For successful FL networks, we recommend that organizations identify the key regulatory body—which may be in the form of an established government regulatory authority, health or research institution boards/councils, or established academies or colleges of medicine and health. This regulatory body will provide leadership over all FL nodes, establish guidelines for standardized protocols and hardware/software infrastructures among nodes, and maintain regulatory oversight and adherence to ethical and legal requirements. The FL network should also engage in intellectual property distribution discussion early and ensure that intellectual property rights obtained are valid in all relevant participating territories, especially in regional or international collaborations.

Lastly, health economic analyses of FL are required for the acceptance of FL at the governance and policy levels.

Key future directions of FL-related research should aim to address these barriers to increase widespread clinical adoption of FL. Further research should focus on methods to (1) weigh inventorship contribution and address intellectual property issues to incentive high-quality data sharing and larger sites to join the federation; (2) boost privacy mechanisms such as the use of differential privacy, homomorphic encryption, secure multi-party computation, and blockchain; (3) explore explainability methods of FL models to increase acceptance from clinicians and patients; (4) develop standardized pipelines for data collection and feature extraction to improve data quality and model performance; (5) develop guidelines for participating nodes to acquire sufficient hardware and software infrastructure capabilities; and (6) perform health economics analyses to evaluate FL frameworks.

The gradual yet inevitable shift away from traditional data sharing has made it imperative for us to focus on privacy-preserving technologies such as FL. This systematic review provides a comprehensive overview of the current state of FL and acts as a foundation for future FL research. We strongly believe that FL holds great potential in this new era of data-driven digital health, and future works to address the barriers will make FL a pivotal strategy for global health collaboration.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2024.101419.

ACKNOWLEDGMENTS

This work was supported by the National Medical Research Council, Singapore (MOH-000655-00 and MOH-001014-00), Duke-NUS Medical School (Duke-NUS/RSF/2021/0018 and 05/FY2020/EX/15-A58), and the Agency for Science, Technology, and Research (A20H4g2141 and H20C6a0032).

AUTHOR CONTRIBUTIONS

Z.L.T., L.J., and D.S.W.T. contributed to the conceptualization and methodology design of the study. Z.L.T., L.J., W.Y.N., T.F.T., D.M.L., K.J.C., J.H., and X.Z. assisted in investigation, resources, and data curation. Z.L.T., S.L., and D.M. contributed to visualization of the work. Z.L.T., L.J., S.L., X.Z., W.Y.N., T.F.T., D.M.L., K.J.C., J.H., Y.L., R.S.M.G., and D.S.W.T. contributed to interpretation of data and manuscript preparation. Y.L., R.S.M.G., and D.S.W.T. provided supervision.

DECLARATION OF INTERESTS

D.S.W.T. holds a patent on a deep learning system for detection of retinal diseases and co-founded and holds equity in EyRIS Singapore.

Received: May 8, 2023
Revised: November 17, 2023
Accepted: January 18, 2024
Published: February 9, 2024

REFERENCES

1. Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R., and Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. *10*, 12598. https://doi.org/10.1038/s41598-020-69250-1.
2. Price, W.N., and Cohen, I.G. (2019). Privacy in the age of medical big data. Nat. Med. *25*, 37–43. https://doi.org/10.1038/s41591-018-0272-7.
3. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. *25*, 1337–1340. https://doi.org/10.1038/s41591-019-0548-6.
4. He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nat. Med. *25*, 30–36. https://doi.org/10.1038/s41591-018-0307-0.
5. McMahan HB, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-Efficient Learning of Deep Networks from Decentralized Data.arxiv Preprint at: Published online January 26, 2023. Accessed November 17, 2023. http://arxiv.org/abs/1602.05629
6. Sadilek, A., Liu, L., Nguyen, D., Kamruzzaman, M., Serghiou, S., Rader, B., Ingerman, A., Mellem, S., Kairouz, P., Nsoesie, E.O., et al. (2021). Privacy-first health research with federated learning. NPJ Digit. Med. *4*, 132. https://doi.org/10.1038/s41746-021-00489-2.
7. Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., and Bakas, S. (2019). Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. Brainlesion. *11383*, 92–104. https://doi.org/10.1007/978-3-030-11723-8_9.
8. Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.S., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nat. Med. *27*, 1735–1743. https://doi.org/10.1038/s41591-021-01506-3.

9. Brisimi, T.S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I.C., and Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. Int. J. Med. Inform. *112*, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007.

10. Hossen, M.N., Panneerselvam, V., Koundal, D., Ahmed, K., Bui, F.M., and Ibrahim, S.M. (2023). Federated Machine Learning for Detection of Skin Diseases and Enhancement of Internet of Medical Things (IoMT) Security. IEEE J. Biomed. Health Inform. *27*, 835–841. https://doi.org/10.1109/JBHI.2022.3149288.

11. Lu, C., Hanif, A., Singh, P., Chang, K., Coyner, A.S., Brown, J.M., Ostmo, S., Chan, R.V.P., Rubin, D., Chiang, M.F., et al. (2022). Federated Learning for Multicenter Collaboration in Ophthalmology: Improving Classification Performance in Retinopathy of Prematurity. Ophthalmol. Retina *6*, 657–663. https://doi.org/10.1016/j.oret.2022.02.015.

12. Adnan, M., Kalra, S., Cresswell, J.C., Taylor, G.W., and Tizhoosh, H.R. (2022). Federated learning and differential privacy for medical image analysis. Sci. Rep. *12*, 1953. https://doi.org/10.1038/s41598-022-05539-7.

13. Wang, Q., and Zhou, Y. (2022). FedSPL: federated self-paced learning for privacy-preserving disease diagnosis. Brief. Bioinform. *23*, bbab498. https://doi.org/10.1093/bib/bbab498.

14. Arikumar, K.S., Prathiba, S.B., Alazab, M., Gadekallu, T.R., Pandya, S., Khan, J.M., and Moorthy, R.S. (2022). FL-PMI: Federated Learning-Based Person Movement Identification through Wearable Devices in Smart Healthcare Systems. Sensors *22*, 1377. https://doi.org/10.3390/s22041377.

15. Zhang, A., Xing, L., Zou, J., and Wu, J.C. (2022). Shifting machine learning for healthcare from development to deployment and from models to data. Nat. Biomed. Eng. *6*, 1330–1345. https://doi.org/10.1038/s41551-022-00898-y.

16. Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N.A., 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.

17. Moher, D., Liberati, A., Tetzlaff, J., and Altman, D.G.; PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ *339*, b2535. https://doi.org/10.1136/bmj.b2535.

18. Crowson, M.G., Moukheiber, D., Arévalo, A.R., Lam, B.D., Mantena, S., Rana, A., Goss, D., Bates, D.W., and Celi, L.A. (2022). A systematic review of federated learning applications for biomedical data. PLOS Digit. Health *1*, e0000033. https://doi.org/10.1371/journal.pdig.0000033.

19. Prayitno, Shyu, C.R., Putra, K.T., Chen, H.C., Tsai, Y.Y., Hossain, K.S.M.T., Jiang, W., and Shae, Z.Y. (2021). A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. Appl. Sci. *11*, 11191. https://doi.org/10.3390/app112311191.

20. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. NPJ Digit. Med. *3*, 119. https://doi.org/10.1038/s41746-020-00323-1.

21. Kanani, P., Marathe, V.J., Peterson, D., Harpaz, R., and Bright, S. (2021). Private Cross-Silo Federated Learning for Extracting Vaccine Adverse Event Mentions. Preprint at arxiv, Published online. https://doi.org/10.48550/ARXIV.2103.07491.

22. Heyndrickx W, Mervin L, Morawietz T, N. Sturm, L. Friedrich, A. Zalewski, A. Pentina, L. Humbeck, M. Oldenhof, R. Niwayama et al. MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information. J. Chem. Inf. Model.. Published online August 29, 2023. https://doi.org/10.1021/acs.jcim.3c00799.

23. Heyndrickx, W., Arany, A., Simm, J., Pentina, A., Sturm, N., Humbeck, L., Mervin, L., Zalewski, A., Oldenhof, M., Schmidtke, P., et al. (2023). Conformal efficiency as a metric for comparative model assessment befitting federated learning. Artif. Intell. Life Sci. *3*, 100070. https://doi.org/10.1016/j.ailsci.2023.100070.

24. Dou, Q., So, T.Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., Li, Z., Si, W., Lee, H.H.C., Yu, K., et al. (2021). Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. NPJ Digit. Med. *4*, 60. https://doi.org/10.1038/s41746-021-00431-6.

25. Bai, X., Wang, H., Ma, L., Xu, Y., Gan, J., Fan, Z., Yang, F., Ma, K., Yang, J., Bai, S., et al. (2021). Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. Nat. Mach. Intell. *3*, 1081–1089. https://doi.org/10.1038/s42256-021-00421-z.

26. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S.W., Karthikesalingam, A., King, D., Ashrafian, H., and Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit. Med. *4*, 65. https://doi.org/10.1038/s41746-021-00438-z.

27. Jefferson, E., Liley, J., Malone, M., Reel, S., Crespi-Boixader, A., Kerasidou, X., Tava, F., McCarthy, A., Preen, R., Blanco-Justicia, A., et al. (2022). GRAIMATTER Green Paper: Recommendations for Disclosure Control of Trained Machine Learning (ML) Models from Trusted Research Environments (TREs). https://doi.org/10.5281/zenodo.7089491.

28. Nguyen, T.V., Dakka, M.A., Diakiw, S.M., VerMilyea, M.D., Perugini, M., Hall, J.M.M., and Perugini, D. (2022). A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data. Sci. Rep. *12*, 8888. https://doi.org/10.1038/s41598-022-12833-x.

29. Nikolaidis, F., Symeonides, M., and Trihinas, D. (2023). Towards Efficient Resource Allocation for Federated Learning in Virtualized Managed Environments. Future Internet *15*, 261. https://doi.org/10.3390/fi15080261.

30. Huang, W., Li, T., Wang, D., Du, S., Zhang, J., and Huang, T. (2022). Fairness and accuracy in horizontal federated learning. Inf. Sci. *589*, 170–185. https://doi.org/10.1016/j.ins.2021.12.102.

31. Liu, W., He, Y., Wang, X., Duan, Z., Liang, W., and Liu, Y. (2023). BFG: privacy protection framework for internet of medical things based on blockchain and federated learning. Connect. Sci. *35*, 2199951. https://doi.org/10.1080/09540091.2023.2199951.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Rayyan Systematic Review platform	Rayyan AI	https://www.rayyan.ai/
Python 3.7	Python	www.python.org
QGIS	QGIS	https://plugins.qgis.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, A/Professor Daniel Shu Wei Ting (daniel.ting@duke-nus.edu.sg).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All papers included in this systematic review are detailed in supplementary material and is publicly available as of the date of publication.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

No experimental model was used in this study.

METHOD DETAILS

This systematic review was performed in accordance to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines. A systematic search using PubMed, Medline, Web of Science, Scopus, Embase, Institute of Electrical and Electronics Engineers Xplore, ArXiv, Springerlink, CINAHL, ACM Digital Library and Google Scholar was conducted. We included studies with the following criteria: 1) federated learning; 2) healthcare or medical related 3) English language 4) original research articles proposing FL applications in the healthcare domain with or without prototype development. We excluded studies which 1) were duplicates; 2) surveys, opinions, editorial letters, book chapters, thesis, conference proceedings 3) not peer-reviewed 4) did not have full text available.

QUANTIFICATION AND STATISTICAL ANALYSIS

Rayyan AI software was used for the removal of duplicate articles during the systematic review. Python 3.7.0 software and QGIS software was used for the generation of evidence map and global map figures (Figure 3; Figure 5; Figure 6). Descriptive statistics were performed using Microsoft Excel.

Cell Reports Medicine, Volume *5*

Supplemental information

Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture

Zhen Ling Teo, Liyuan Jin, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, Ting Fang Tan, Deborah Meixuan Lee, Kai Jie Chua, John Heng, Yong Liu, Rick Siow Mong Goh, and Daniel Shu Wei Ting

Supplementary Table 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement, related to systematic review search process in Figure 2

The PRISMA guidelines(1) was used for the reporting of this systematic review and meta-analysis. The 27-item checklist is as follows:

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	3-4
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	5-6
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5-6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	7
Eligibility criteria	5-6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	8-9
Information sources	5-6	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	7-9
Search	5-6	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	7-8
Study selection	5-7	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	7-9

Data collection process	6-7	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	8-9
Data items	6-7	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7-10
Risk of bias in individual studies	8	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	-
Summary measures	8	State the principal summary measures (e.g., risk ratio, difference in means).	-
Synthesis of results	8	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	-
Risk of bias across studies	13	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	-
Additional analyses	13	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	10
RESULTS			
Study selection	9	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	11
Study characteristics	9	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	11
Risk of bias within studies	9-10	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	-
Results of individual studies	9-10	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	11-12
Synthesis of results	9-12	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	-
Risk of bias across studies	9-10	Present results of any assessment of risk of bias across studies (see Item 15).	-
Additional analysis	11-12	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	-
DISCUSSION			

Summary of evidence	12-13	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	14-16
Limitations	13-14	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	15-16
Conclusions	16-17	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	14-19
FUNDING			
Funding	18	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	21

Supplementary Table 2: Electronic Database Search Strategy, related to the systematic review search process in Figure 2

Advanced search options:

Database	**Search Term**
PubMed/ Medline	(Federated Learning[Title/Abstract]) OR (Decentralised machine learning[Title/Abstract])
Embase	Federated learning [Title/abstract] AND (health or medical) [All fields]
IEEEXplore	(Federated learning OR decentralised machine learning) AND (healthcare OR medical)
ScienceDirect	Title/Abstract/Author-specified keywords (Federated learning) AND Any field: (health OR medical)
SpringerLink	"Federated learning" AND (health OR medical)
ArXiv	Federated learning [Abstract] AND (health OR medical) [All fields]
Scopus	ABS-Title-Key (Federated learning) OR all text (health Or medical)
Web of Science	Federated learning (abstract) and Health or Medical (All fields)
CINAHL	Federated learning OR Decentralised machine learning [All fields]
Google Scholar	(Federated learning) OR Decentralised machine learning AND (health OR medical)
ACM Digital Library	[Title: federated learning] AND [Abstract: health or medicine or medical]