BMC Infectious Diseases

RESEARCH

Open Access

Federated learning as a smart tool for research on infectious diseases

Laura C. Zwiers[1,2*], Diederick E. Grobbee[1,2], Alicia Uijl[1,3,4] and David S. Y. Ong[1,2,5]

Abstract

Background The use of real-world data has become increasingly popular, also in the field of infectious disease (ID), particularly since the COVID-19 pandemic emerged. While much useful data for research is being collected, these data are generally stored across different sources. Privacy concerns limit the possibility to store the data centrally, thereby also limiting the possibility of fully leveraging the potential power of combined data. Federated learning (FL) has been suggested to overcome privacy issues by making it possible to perform research on data from various sources without those data leaving local servers. In this review, we discuss existing applications of FL in ID research, as well as the most relevant opportunities and challenges of this method.

Methods References for this review were identified through searches of MEDLINE/PubMed, Google Scholar, Embase and Scopus until July 2023. We searched for studies using FL in different applications related to ID.

Results Thirty references were included and divided into four sub-topics: disease screening, prediction of clinical outcomes, infection epidemiology, and vaccine research. Most research was related to COVID-19. In all studies, FL achieved good accuracy when predicting diseases and outcomes, also in comparison to non-federated methods. However, most studies did not make use of real-world federated data, but rather showed the potential of FL by using data that was manually partitioned.

Conclusions FL is a promising methodology which allows using data from several sources, potentially generating stronger and more generalisable results. However, further exploration of FL application possibilities in ID research is needed.

Keywords Infection, Vaccine, Federated learning, Big data, Machine learning, AI

Introduction

Over the past decade, there has been a vast increase in the use of real-world data in infectious disease (ID) research, including the use of machine learning (ML) and artificial intelligence (AI). Advantages of these techniques are the potential of personalized medicine, providing up-to-date and real-time predictions, and assisting in reducing human errors in clinical practice [1, 2].

Although very promising, the use of data in health research has its challenges, which include privacy legislation and intellectual property concerns. Data anonymization is often performed to protect privacy, but this is not always sufficient. For instance, models have been developed that can re-identify individuals based on a small set

*Correspondence:
Laura C. Zwiers
laura.zwiers@juliusclinical.com
[1] Julius Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
[2] Julius Clinical, Zeist, The Netherlands
[3] Department of Cardiology, Amsterdam University Medical Centers, Amsterdam Cardiovascular Sciences, University of Amsterdam, Amsterdam, The Netherlands
[4] Division of Cardiology, Department of Medicine, Karolinska Institutet, Stockholm, Sweden
[5] Department of Medical Microbiology and Infection Control, Franciscus Gasthuis & Vlietland, Rotterdam, The Netherlands

of demographic attributes [3] and face recognition software was found to have the ability to identify anonymous individuals from magnetic resonance imaging scans of the face [4]. Moreover, access to health data is strictly regulated, for instance through the Health Insurance Portability and Accountability Act (HIPAA) [5] and the General Data Protection Regulation (GDPR) in Europe [6]. Regulations like these have made data transactions and multi-institutional collaborations more difficult [7]. Furthermore, as collecting high-quality health data is a large investment and data are not generally collected for the sake of research only, researchers and institutions are often hesitant to share their data to not lose ownership [8]. Healthcare data are therefore often fragmented, such that hospitals may only access the data collected inhouse, and population registries cannot be shared across country borders [9].

The traditional approach of performing analyses using data from different sources involves pooling all data into a central server. Federated learning (FL) uses a different approach. With FL, analyses are performed locally and only the outputs are shared. A central server pools local model outputs to create a combined model, which is shared with local servers again for further optimization (Fig. 1). The FL method, which was introduced by Google in 2006 [10], overcomes the main privacy issues related to multi-institutional research. It was initially introduced for mobile device research, but quickly became of interest in healthcare, because it allowed local data holders to keep their own privacy policies and to control data access, while enabling researchers to analyze the data [8].

So far, FL has most often been applied in the fields of oncology and radiology, with most research employing imaging data [11]. During the coronavirus disease 2019 (COVID-19) pandemic, however, researchers recognized the usefulness of FL for research on infections and vaccines. This review provides an overview of existing applications of FL in ID research, which are mostly related to COVID-19. The advantages and challenges of the methodology and opportunities for future research are outlined.

Methods

References for this review were identified through searches of PubMed/MEDLINE [12], Google Scholar, Embase and Scopus [13] by using a combination of keywords related to FL ('federated learning', 'distributed learning', 'federated machine learning', 'federated deep learning', 'federated model') and IDs ('infectious diseases', 'infection', 'vaccines', 'virus', 'bacteria', 'parasite', 'fungi'). By including these four databases in the review, we aimed to achieve broad coverage of all relevant literature. That

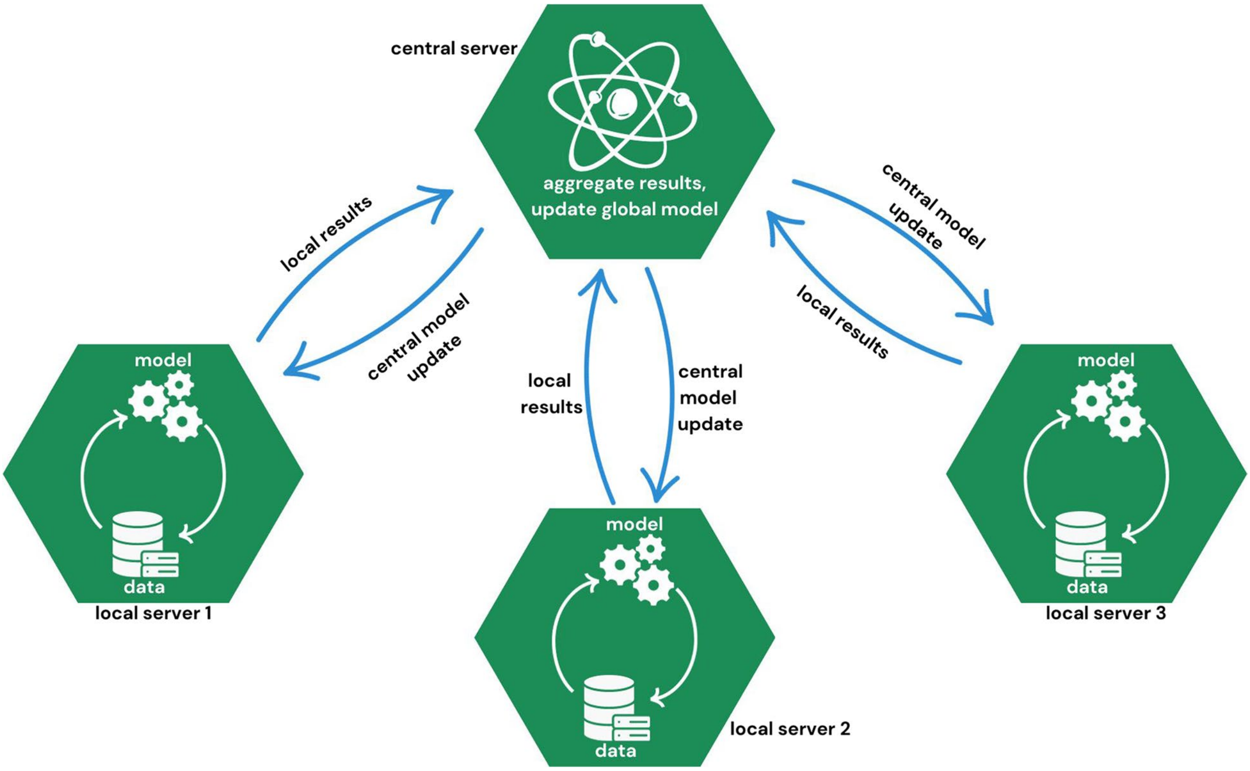


Fig. 1 Illustration of a FL architecture

is, PubMed/MEDLINE is the most commonly used database of medical literature, but it has been suggested that Embase should be used as a supplement to PubMed/MEDLINE in reviews [14]. Since FL is a statistical method that is often used in ML research, we also searched beyond the medical literature. This was done by including Google Scholar and Scopus in the search strategy, with the aim of covering a broader range of literature, and to also incorporate potentially relevant grey literature in the search [15]. An overview of the exact search terms used in PubMed/MEDLINE, Embase, and Scopus is presented in Additional Box 1 (Additional file 1).

Titles, abstracts, and full texts retrieved from PubMed/MEDLINE and Embase were screened for relevance by two reviewers. In case of conflicts, reviewers discussed the eligibility for inclusion of the article until consensus was reached. The additional explorative search in Google Scholar and Scopus was performed by one reviewer. Articles were included if they (1) discussed a clinical application of FL in the ID field, (2) had a clear description of the data and results, (3) were published open access, or through a subscription available to the authors, and (4) were published before July 2023. Articles were excluded if they (1) did not clearly describe the used data, (2) did not describe a clinical application of FL in the ID field, or (3) focused on methodology rather than a clinical application.

Results

Selection of references

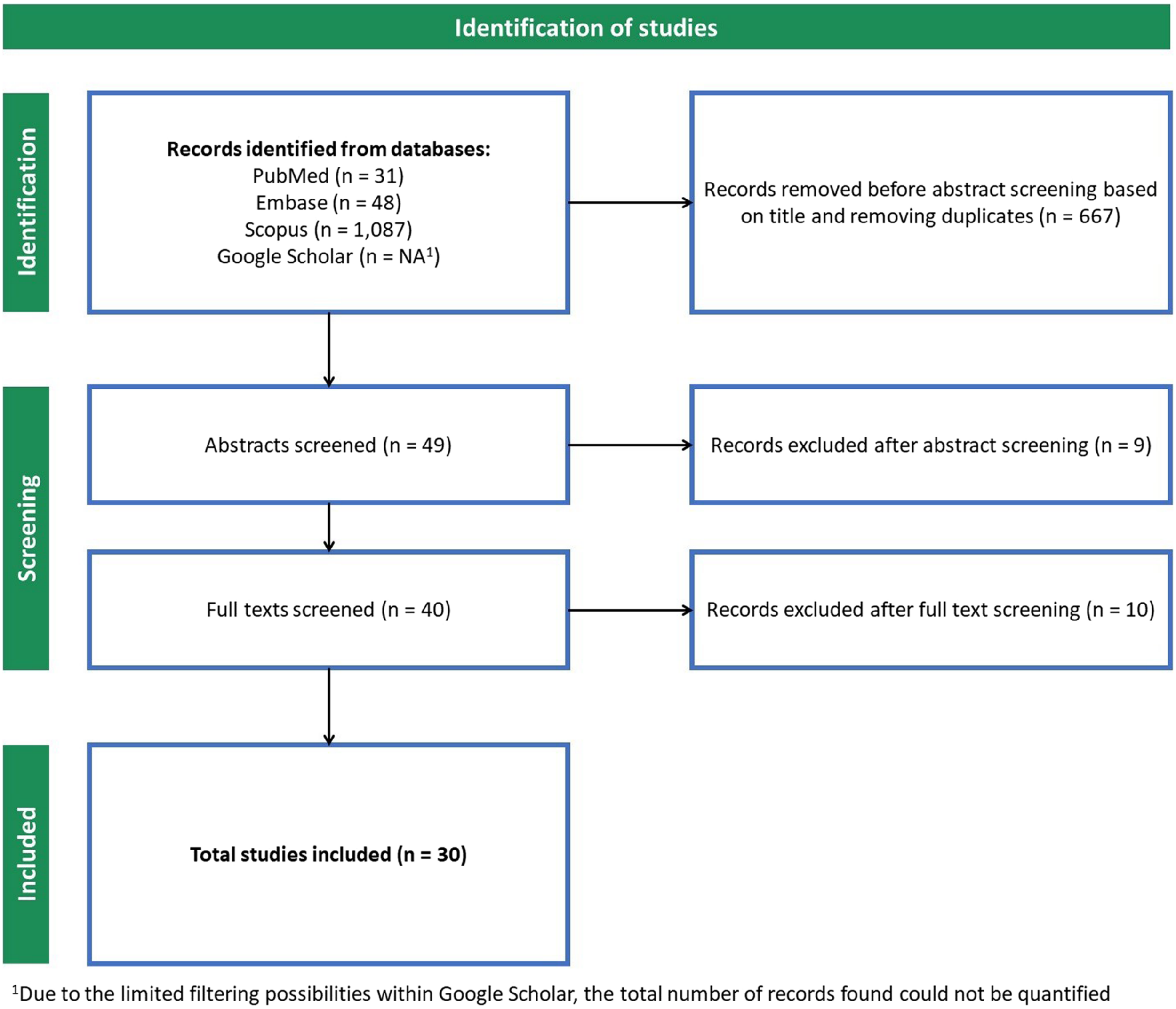
Following searches within the various databases and screening on title, 49 abstracts were screened, after which 9 articles were excluded due to inaccessibility of the full text, lack of FL methodology introduced, or because no clinical application of FL in ID research was presented (Fig. 2). Full texts of the 40 remaining articles were then screened, after which 10 were excluded due to several reasons. A list of the excluded articles and the reason for exclusion is presented in Tables A1 and A2 of Additional file 1. The 30 remaining articles were included in this review and are listed in Table 1. We could identify four sub-topics across the articles, namely disease screening ($n=21$), prediction of clinical outcomes ($n=3$), infection epidemiology ($n=4$), and vaccines ($n=2$). Figure 3 provides a graphical overview of the identified papers. The following sections outline the main findings per sub-topic.

Disease screening

A total of 21 studies reported on the use of FL in disease screening, of which the vast majority ($n=18$) propose the use of FL for analyzing Chest X-ray (CXR) and computed tomography (CT) images [16–33]. Although RT-PCR tests on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) remain the standard for diagnosing COVID-19, screening through the automatic assessment of chest X-ray (CXR) and computed tomography (CT) images has been suggested and investigated in several studies because of its wide availability and time-efficiency [16]. Since the development of models for disease screening requires many cases to achieve proper accuracy, FL was proposed as a tool for analyzing CXR and CT images from several sites in a single model to obtain sufficient cases [16–33]. Across these studies, which span over 100,000 medical images from a variety of sources, FL was shown to achieve proper accuracy, with accuracy measures often higher than 90%. Moreover, FL was shown to perform much better compared to local modelling in most cases. As an example, Dayan et al. found that FL achieved 16–18% better accuracy compared to models developed on single-site data [37]. Of note, most research in this context was not performed on real-world federated data, but rather on data that was manually partitioned for testing FL methodologies. For instance, Feki et al. performed their experiment using one larger dataset of CXR images, which they manually split into four parts of equal size [17]. Therefore, not all results may be directly generalizable to clinical practice.

Soltan et al. also used FL for COVID-19 screening, but their input data consisted of clinical data routinely collected within the first hour after the patient arrived at the hospital, namely vital signs, full blood count, liver function, urea and electrolytes, and C-reactive protein [34]. Their FL framework was applied in emergency departments across three hospital groups in the UK and different statistical models were applied in the federated setting. Overall, federated models achieve higher accuracy compared to local models, with accuracy up to 84%.

Another possible application of FL in disease screening relates to the analysis of SARS-CoV-2 spike sequences to identify and classify virus variants without the need for sharing the genome sequencing data. Chourasia et al. [35] presented a FL methodology for variant classification based on spike sequences, which could allow for combining data from a variety of sources to study variants. Their method achieved an overall accuracy of 93% with regard to predictive performance using only spike sequences rather than a more elaborate and expensive standard approach using full-length genome sequences. Sim et al. [36] also introduced a variant classification algorithm based on FL, using genetic data from eight different COVID-19 strains. They performed different simulations where data were distributed across multiple data providers. Their FL methodology included private training at each data provider, which preserved privacy. The FL framework achieved almost similar accuracy (i.e.


Identification of studies
Identification
Screening
Included
Records identified from databases:
PubMed (n = 31)
Embase (n = 48)
Scopus (n = 1,087)
Google Scholar (n = NA¹)
Records removed before abstract screening based on title and removing duplicates (n = 667)
Abstracts screened (n = 49)
Records excluded after abstract screening (n = 9)
Full texts screened (n = 40)
Records excluded after full text screening (n = 10)
Total studies included (n = 30)

[1] Due to the limited filtering possibilities within Google Scholar, the total number of records found could not be quantified

Fig. 2 Flowchart describing the search process and selection of references

a classification Area Under the Receiver Operator Characteristic (ROC) curve (AUC) of 0.99) as was achieved using centralized training.

Prediction of clinical outcomes

Several studies have used FL for predicting clinical outcomes following SARS-CoV-2 infection. Many Clinical Decision Support (CDS) models were developed during the pandemic for triaging patients, most of which were only trained and validated on local data. This lacked diversity, which is why Dayan et al. [37] and Vaid et al. [38] developed CDS models using FL. Their models employed real-world data from over 16,000 and 4,000 patients from 20 and 5 different hospitals, respectively, and achieved better accuracy than models trained on single-site data.

Hoyos et al. [39] presented an application of FL for supporting clinical decision-making in dengue virus infections and proposed different FL tools for the diagnosis and treatment of dengue. Using two datasets of around 400 observations, their application outperformed local models in terms of accuracy, sensitivity, and specificity.

Monitoring infection epidemiology

During the COVID-19 pandemic, many models were used for gaining insight into virus spread and predicting infection rates. To circumvent privacy issues, FL was proposed as a method for developing these models. Kumaresan et al. [40] developed a FL model using mobility data from mobile devices to predict infection rates and availability of hospital beds by employing a SEIR (Susceptible-Exposed-Infected-Recovered) model. Samuel et al. [41] employed FL in a SEIR model

Table 1 Overview of studies using Federated Learning in infectious disease research

Study	Topic	Aim	Type of data/level of federation + number of sites	Type of data split	Results and conclusion
Chowdhury et al., 2023 [16]	Disease screening	Distinguishing COVID-19 from other causes of viral pneumonia and healthy individuals using CXR images	1823 CXR images from various publicly available datasets	Data manually split by authors	FL achieved almost perfect accuracy (global accuracy of 99.6%)
Feki et al., 2021 [17]	Disease screening	COVID-19 screening using CXR images. Comparing FL to centralized model	216 CXR images from various publicly available datasets	Data manually split into four sites by the authors	A comparison between FL and centralized modelling across different data properties resulted in nearly identical accuracy, sensitivity, and specificity
Abdul Salam et al., 2021 [18]	Disease screening	COVID-19 screening using CXR images. Comparing FL model to traditional model	Two open-source datasets with a total of 5,144 CXR images	Data manually split by the authors	FL achieved better predictive accuracy than traditional deep learning
Zhang et al., 2021 [19]	Disease screening	Propose a FL methodology to screen for COVID-19 pneumonia from CXRs	Publicly available dataset of 3,600 CXR images	Data manually split into 100 sites by the authors	The proposed methodology had better accuracy (94.45%) and privacy preservation compared to centralized models
Zhang et al., 2021 [20]	Disease screening	Propose a FL methodology for COVID-19 screening using both CXR and CT images	Publicly available datasets of 2,960 CXR and 746 CT images	Data manually split into three sites by the authors	The proposed methodology achieved proper accuracy in comparison to other FL methods
Li et al., 2022 [21]	Disease screening	Propose a FL methodology for COVID-19 screening from CXRs	Dataset of CXR images	Data manually split by the authors	The proposed methodology outperformed other FL methods and had an accuracy of up to 95.0%
Ho et al., 2022 [22]	Disease screening	Propose a FL methodology for COVID-19 screening from CXRs and symptom data	Various publicly available datasets with 15,153 CXR images in total, and one symptom dataset with 5,434 observations	Data manually split into three sites by the authors	The proposed methodology achieved good accuracy (up to 96.7%) and higher privacy compared to other models
Malik et al., 2023 [23]	Disease screening	Propose a FL methodology for distinguishing COVID-19 from four distinct chest disorders from CXRs	Various publicly available datasets with over 100,000 CXR images combined	Data manually split into three sites by the authors	Proposed model achieved almost perfect accuracy (98.5%) and outperforms other approaches while protecting privacy
Dou et al., 2021 [24]	Disease screening	COVID-19 image interpretation from CT images	Model development on 75 CT images from three hospitals, external validation on 57 CT images from four other hospitals	Each hospital was considered a site	The model achieved high performance, with an AUC of 0.953 in internal validation, and an AUC between 0.812 and 0.957 in different external validation settings

Table 1 (continued)

Study	Topic	Aim	Type of data/level of federation + number of sites	Type of data split	Results and conclusion
Florescu et al., 2022 [25]	Disease screening	COVID-19 screening using CT images	Various datasets of 2230 CT images combined, some of which were publicly available	Data manually split into three sites by the authors	In internal validation, FL performed slightly worse compared to centralized modelling (AUC of 0.838 vs. 0.939), while performance was nearly identical in external validation (0.793 vs. 0.790 AUC)
Kumar et al., 2021 [26]	Disease screening	COVID-19 screening using CT images	34,006 CT images from three different hospitals and one third party dataset for external validation	Each hospital was considered a site	FL was shown to achieve similar accuracy as local models for a comparable task with an accuracy of 98.7% and sensitivity of 98.0%
Durga and Poovammal, 2022 [27]	Disease screening	Propose a FL methodology for COVID-19 screening using CT images	Three real-world datasets of over 34,000 CT images combined from various hospitals	Each dataset was considered a separate site	The proposed model outperformed several existing models in terms of accuracy, precision, recall, and specificity. The average accuracy of the FL method was 98.5%
Malik et al., 2023 [28]	Disease screening	Propose a FL methodology for COVID-19 screening from CT images	Five publicly available datasets of CT images	A subset of 540 CT images of each dataset was considered a separate site	The proposed model outperformed local models with up to 99.0% accuracy
Peng et al., 2022 [29]	Disease screening	Apply different FL methodologies for COVID-19 screening from CXRs	Five real-world datasets of CXRs from 42 hospitals in Europe and the US; three of the datasets were used for model development and two for external validation	Each dataset was considered a separate site	FL models tended to achieve higher accuracy compared to local models
Wang et al., 2022 [30]	Disease screening	Propose a FL methodology for COVID-19 screening from CT images	1313 CT images of normal chests, 1316 of novel coronavirus and 1171 of COVID-19	Data manually split into 20 sites by the authors	The proposed method achieved almost similar accuracy as centralized training (94.7% vs. 96.3%) while preserving privacy
Liu et al., 2020 [31]	Disease screening	Apply different FL techniques to COVID-19 screening from CXRs	Open access dataset of 15,282 CXRs	Data manually split into five sites by the authors	Across different techniques, a sensitivity of over 0.90 was achieved in internal validation and over 0.86 in external validation
Qayyum et al., 2022 [32]	Disease screening	Propose a novel methodology for COVID-19 screening from CXRs and chest ultrasounds	1,564 CXRs and 545 chest ultrasounds	Data manually split by the authors	The proposed methodology outperformed other FL techniques

Table 1 (continued)

Study	Topic	Aim	Type of data/level of federation + number of sites	Type of data split	Results and conclusion
Kandati and Gadekallu, 2023 [33]	Disease screening	Propose a novel methodology for COVID-19 screening from CT images	317 CT images of COVID-19, normal, and viral pneumonia	Data manually split into 10 sites by the authors	Proposed method achieved over 90% accuracy
Soltan et al., 2023 [34]	Disease screening	COVID-19 screening across three hospital groups, using data on vital signs and laboratory values	Vital signs and laboratory values of patients with an unscheduled acute emergency care admission across three hospital groups in the UK	Each hospital group was considered a site	Federated models achieved higher accuracy compared to local models, with accuracy up to 84.4%
Chourasia et al., 2023 [35]	Disease screening	Classify different COVID-19 variants using spike proteins	Publicly available data of 9,000 spike protein sequences, with nine different variants	Data manually split into three sites by the authors	FL method achieved a good performance with an accuracy of 93% and AUC of 0.961
Sim et al., 2023 [36]	Disease screening	Classification of COVID-19 strains	16,000 samples of COVID-19, of which 2,000 per variant for model development, and 4,000 samples for external validation	Data manually split into different sites	FL achieved almost similar accuracy as centralized training (98.4% vs. 98.6%) while preserving privacy
Dayan et al., 2021 [37]	Clinical outcome prediction	Development of a CDS model that can predict oxygen requirements and mortality for COVID-19 patients using FL	CXRs, laboratory values, and vital signs from 16,148 patients from 20 sites across the world	Data stemmed from 20 sites across the world	FL model achieved 16% better AUC and 38% better generalizability compared to models trained on single-site data for the prediction of 24-h oxygen treatment. For the prediction of 72-h oxygen treatment, the improvements were 18% and 34%, respectively
Vaid et al., 2021 [38]	Clinical outcome prediction	Predict 7-day mortality in hospitalized patients with COVID-19	EHR data from 4,029 COVID-19 positive patients from five different hospitals	Data stemmed from five different hospitals, which were all separate sites	FL models achieved AUC between 0.694 and 0.836 and outperformed the models developed at individual hospitals in most cases
Hoyos et al., 2023 [39]	Clinical outcome prediction	Develop FL approached for the prediction of mortality and prescription of treatment in severe dengue	Two datasets of 400 and 398 observations from dengue endemic regions with variables related to severe dengue and treatment	Both datasets were considered as separate sites	FL model (with two clients) had accuracy of 0.96, which was higher than the accuracy of local models (0.87 and 0.86, respectively)
Kumaresan et al., 2022 [40]	Infection epidemiology	Predict COVID-19 infection rate and the availability of beds using SEIR models	Google data on mobility and COVID-19 in the US	Data manually federated at state level, with five states included in the experiment	The FL model was able to predict infection rates and bed availability with good accuracy
Samuel et al., 2023 [41]	Infection epidemiology	Analyzing the COVID-19 pandemic through an extended SEIR model	Simulated data that should represent the data on patients from different centers for disease control	Data manually split into sites by the authors	The proposed methodology could effectively analyze the pandemic and was robust against privacy attacks

Table 1 (continued)

Study	Topic	Aim	Type of data/level of federation + number of sites	Type of data split	Results and conclusion
Chen et al., 2021 [42]	Infection epidemiology	Create a COVID-19 vulnerability map based on data from self-reporting apps	Simulated data based on a publicly available dataset, which represent data from different self-reporting apps	Data manually split into different sites (by neighborhood) by the authors	The proposed method worked well on simulated data
Pang et al., 2021 [43]	Infection epidemiology	Develop a city Digital Twin to analyze COVID-19 response plans	Publicly available datasets of daily epidemic data	Data manually split into different sites by the authors	The work showed potential of using FL and Digital Twin combined on real data
Wang et al., 2023 [44]	Vaccine research	Predict COVID-19 vaccination side effects from EHR data	EHR data of 6,526 patients from health insurance claims	Data naturally split at state level	The FL model was effective at predicting side effects
Kanani et al., 2021 [45]	Vaccine research	Vaccine adverse event detection from the VAERS	17,841 publicly available VAERS reports	Data manually split (by manufacturer) by the authors	FL is a promising method for privacy-preserving adverse event detection

AUC Area under the receiver operating characteristics (ROC) curve, *FL* Federated learning, *CXR* Chest X-ray, *CT* Computed Tomography, *EHR* Electronic Health Record, *SEIR* Susceptible- Exposed-Infected-Recovered, *VAERS* Vaccine Adverse Event Reporting System

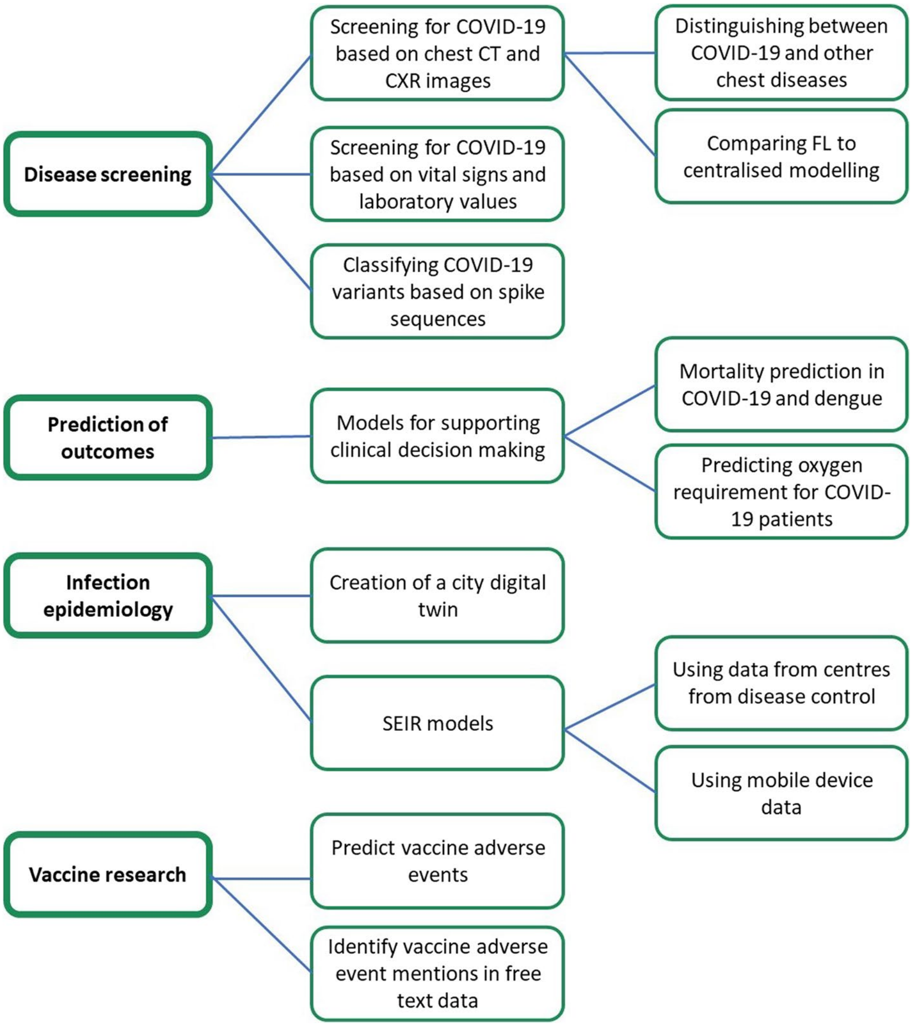

Disease screening
Screening for COVID-19 based on chest CT and CXR images
Distinguishing between COVID-19 and other chest diseases
Comparing FL to centralised modelling
Screening for COVID-19 based on vital signs and laboratory values
Classifying COVID-19 variants based on spike sequences
Prediction of outcomes
Models for supporting clinical decision making
Mortality prediction in COVID-19 and dengue
Predicting oxygen requirement for COVID-19 patients
Infection epidemiology
Creation of a city digital twin
SEIR models
Using data from centres from disease control
Using mobile device data
Vaccine research
Predict vaccine adverse events
Identify vaccine adverse event mentions in free text data

Fig. 3 Overview of main FL applications in ID scientific literature

for analyzing the COVID-19 pandemic using data from different centers for disease control. Chen et al. [42] provided a proof-of-concept for using FL to create a vulnerability map using data from different self-reporting apps. Such a map could be used for identifying high-risk areas, and combining different apps creates a more diverse sample of the population. These studies were all performed on data that were manually split into different sites by the author.

The dynamics of a virus infection can also be modelled using a city Digital Twin, which essentially is a virtual replica of a city or region that provides an insight into factors that influence virus spread. Pang et al. [43] proposed a FL solution for developing a Digital Twin that can be used for response plan management and analyzing the effect of different measures to reduce virus spread.

Vaccine research

The COVID-19 pandemic led to the development of many vaccines and the implementation of large-scale vaccination campaigns. However, with the many vaccines also came the risk of the occurrence of side effects. FL is of interest in studying rare vaccine side effects, as usually only few cases are reported at individual institutions. At the same time, vaccine records are very privacy-sensitive and cannot easily be shared across sites in order to create a dataset with sufficient events. Wang et al. [44] used electronic health record (EHR) data of 6,526 patients from insurance claims across the United States, which were federated on state-level, to predict vaccine side effects. Their FL model performed well in predicting adverse events in comparison to centralized training.

Vaccine adverse events are often recorded as free text in patient records. ML methods that can identify mentions of adverse events from the free text could be developed, but these need large and heterogeneous datasets to perform well. Kanani et al. [45] applied FL for vaccine adverse event detection using data from the Vaccine Adverse Event Reporting System (VAERS), federated on vaccine-manufacturer level. The case study demonstrated that FL is a promising approach for identifying mentions of vaccine adverse events from these data, which could be useful for achieving more complete overviews of the adverse events that occur during vaccination campaigns.

Discussion

This review summarized the findings of studies on applications of FL for disease screening, prediction of clinical outcomes, infection epidemiology, and vaccine research, of which most were related to COVID-19. The literature search did not yield many studies related to other IDs. In fact, only one study on another ID (i.e., dengue fever) fulfilled inclusion criteria for this review. More research on the use and added value of FL applications for assessing many other types of infections is warranted to capture the needs and challenges associated with various IDs.

The main promise of FL is ensuring data privacy and allowing data holders to have their own governance in place while creating the possibility to be part of multi-institutional analyses. Models for disease detection or prediction of clinical outcomes [16–27, 35, 37–39] can only achieve sufficient accuracy when there are sufficient observations. During the early phase of an emerging ID, data at individual institutions are often of insufficient size to achieve robust findings. FL allows for a faster establishment of more accurate detection and prediction models. This promise also holds for vaccine research because adverse events are rare.

FL also provides more flexibility for local calibration of models. For instance, Li et al. [46] introduced domain adaptation techniques that help improve overall accuracy of federated learning applications. The possibility of local calibration could be especially interesting in settings where heterogeneous epidemiological settings apply, as the final predictive global model following the FL process is iteratively trained on the different individual sites. FL also allows for more equitable research. By combining data from different sources, a more representative source population can be achieved [10]. For instance, specific subgroups may be over- or underrepresented in single-institution research [10], and single-institution models may therefore be vulnerable to bias [16]. FL has been renowned for its potential to include data from underrepresented populations and from hospitals with fewer resources for participation in research [8, 11].

Monitoring of infection epidemiology can be supported by location data from mobile devices [40] or through self-reporting of symptoms [42]. Such examples relate closely to the more classic FL applications outside of healthcare (e.g. [47]), as the methodology was first developed for analysis of mobile device data. A related application is that of using data from wearable devices, which are increasingly used for monitoring health and disease. During the COVID-19 pandemic, many researchers used wearable devices for infection surveillance [48, 49], contract tracing [48] and monitoring behavioral changes [50]. Collecting data from all wearable devices centrally is considered a threat to user privacy, which is why the use of FL in research concerning such data has been suggested [51, 52].

Challenges in the use of FL

Although FL could provide a solution to many issues related to the use of real-world data, various methodological and practical challenges are associated with its use. The following paragraphs discuss some of the most relevant challenges when applying FL.

Heterogeneous data

Data heterogeneity is a serious issue in the application of FL in health research. Issues with data heterogeneity have been argued to deteriorate the performance of FL models [53], and some common FL algorithms may perform poorly in case of highly heterogeneous data [53, 54]. Within ID applications, there are several ways in which data heterogeneity can be a challenge. With EHR data, static variables as gender and age are generally stored similarly across sites, but hospital visit data often consist of different codes. When applying FL to data from various laboratories, issues may also arise, as different sites make use of different software and instruments, amongst others. Imaging data are much less heterogeneous, which is why these have been used more often in FL research [11, 44].

Clinical interpretability

FL studies tend to focus on complex prediction tasks using ML algorithms, which are not easily understandable for clinicians. ML algorithms are often referred to as "black boxes", with the precise computations underlying predictions unknown [55, 56], while some understanding of the underlying model is warranted when the model may be used for clinical decision making. Clinicians are more interested in knowing why the model produces certain outputs, but algorithms from most FL research only focus on maximizing predictive performance. The need for interpretable research and simple clinical scores and predictor variables is often overlooked [39].

In addition to models being difficult to comprehend, especially by clinicians who are the main users, the data used in current FL applications in ID might be a limitation for generalizability of the study findings. Indeed, numerous studies included in this review make use of simulated or manually partitioned data, which do not fully capture the complexities of real-world data. The findings of these studies may therefore not precisely reflect clinical practice, as it remains unclear whether similar predictive performance would be achieved on real-world federated data.

External validity

Apart from the difficulties with interpreting ML output, there is also an issue with external validity of many existing FL applications. External validation is a crucial step, but cannot be done for many FL studies, because data and code are not shared. This does not only limit the possibilities for validation, but also reduces the likelihood of more innovative models being developed [11].

The lack of external validation is also a problem because FL research often does not employ a sufficiently diverse sample. While FL makes it possible to use data from a wide variety of sources, researchers often resort to only using a selective pool of databases in their analyses (e.g. only data from New York hospitals in the study by Vaid et al. [38]), which limits generalizability.

Privacy-accuracy trade-off

Although FL has been praised for its ability to ensure privacy while incorporating data from different institutions, some privacy issues remain. Specifically, it is still possible to learn about sensitive data through local model outputs [11, 54]. Differential privacy methods have been introduced to overcome this issue by adding artificial noise to outputs [57], but these methods come at the cost of decreased accuracy. A privacy-accuracy trade-off therefore still exists [11, 54].

Infrastructure issues

A final important challenge in the application of FL in ID research relates to infrastructure issues. These stem from, for instance, different institutions not having the software in place for local model running and updating. Moreover, the heterogeneity in systems may complicate the application of FL, for instance because some systems may be able to handle more complex tasks than others [54]. Communicating between sites for updating may also pose a challenge, as communication may be slower than computation, which could lead to very long runtimes [54]. In order to overcome these issues, institutions must invest in infrastructures that allow for performing FL research. For instance, fit-for-purpose common data models need to be developed and data need to be locally transformed to align with these common data models [58]. In addition, information technology infrastructures of institutions need to be able to run complex analyses that are required for FL.

Future perspectives

Despite its challenges, FL has the potential to provide a solution to many current challenges in the use of data for ID research. For the study of rare infections, or at the early stages of an outbreak, the method allows for developing accurate models much faster, which could have significant impact in a next pandemic. Early in the COVID-19 pandemic it was already noted that, while incredibly many studies on the virus were being conducted, the availability of robust data was limited [59]. This highlights the opportunities of using FL in such situations. Moreover, in case of very privacy-sensitive data, such as mobility data and vaccine records, FL provides a way of leveraging much data without harming privacy.

Another potentially interesting application of FL is the study of antimicrobial resistance (AMR). Limited data quantity at individual institutions makes it challenging to study different aspects of AMR [60], which suggests that combining input data from multiple different local data sources without the need for data extraction to preserve privacy can help increase the study sample size more easily. Data on AMR are already routinely collected by many microbiological laboratories and research institutions, which means that FL could be applied if the appropriate infrastructure would be implemented. It is often insufficient to merely study incidence and prevalence of AMR, as patient and disease risk factors related to AMR are more relevant to investigate. The assessment of risk factors of AMR might be more feasible in a FL setting compared to a centralized setting, as FL helps protect privacy. Also for last-resort antimicrobial therapies, which are rarely used due to antimicrobial stewardship principles, a FL approach to pool from multiple data sources might help to achieve sufficient statistical power to address questions related to effects and adverse effects of these last-resort antibiotics. Recently, FL has also been suggested to be of interest for automated infection surveillance, including for the surveillance of healthcare associated infections, because of the potential for a significant reduction in the local surveillance burden [58]. FL-based automated surveillance could provide the advantages of standardization and upscaling due to a centralized approach, while it also allows for real-time access to local data with flexibility in algorithms on a local level, such that optimal local implementation of infection surveillance could be achieved.

Current FL applications have mostly been proof of concepts and need to be improved. Many did not make use of real-world, multi-institutional data. Moreover, applications that incorporate real-world data generally do not use very diverse data. To achieve the generalizability that FL allows for, researchers should aim to employ data from a wider variety of sources. Furthermore, more open availability of study code and test data would allow for applying existing models in more diverse populations. The latter would not only aid in external validation, but could also incentivize researchers to keep improving existing models.

Conclusions

FL is a promising methodology for many applications in the context of ID research, but some challenges need to be addressed. Examples of real-world applications in the field are limited, hence the true possibilities of the method are under-investigated. More research on additional applications and methodologies for using FL in ID research is therefore warranted.

Abbreviations

AMR	Antimicrobial resistance
AUC	Area under the receiver operating characteristics curve
AI	Artificial intelligence
CDS	Clinical Decision Support
COVID-19	Coronavirus disease 2019
CT	Computed Tomography
CXR	Chest X-ray
EHR	Electronic Health Record
FL	Federated learning
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
ID	Infectious disease
ML	Machine learning
ROC	Receiver Operator Characteristic
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SEIR	Susceptible- Exposed-Infected-Recovered
VAERS	Vaccine Adverse Event Reporting System

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-024-10230-5.

Supplementary Material 1.

Acknowledgements

Not applicable.

Clinical trial

Clinical trial number: not applicable.

Authors' contributions

LCZ and DSYO contributed to the conception and design of the review. LCZ performed the literature search and performed the abstract and full-text screening process together with DSYO. LCZ wrote the first draft versions of this manuscript under the supervision of DSYO. DEG, AU and DSYO revised the subsequent manuscript versions critically for important intellectual content, and all authors approved the final manuscript version to be submitted.

Funding

No funding.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

LCZ, DSYO and DEG work (part-time) for Julius Clinical Ltd., which is an Academic Contract Research Organization that receives financial support from Moderna Ltd. to conduct research on the cardiovascular safety of COVID-19 vaccination. Moderna had no role in the current paper. AU has no conflict of interest.

Received: 7 May 2024 Accepted: 14 November 2024
Published online: 21 November 2024

References

1. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. JAMA. 2013;309:1351–2. https://doi.org/10.1001/jama.2013.393.
2. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230–43. https://doi.org/10.1136/svn-2017-000101.
3. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019;10:3069. https://doi.org/10.1038/s41467-019-10933-3.
4. Schwarz CG, Kremers WK, Therneau TM, Sharp RR, Gunter JL, Vemuri P, et al. Identification of Anonymous MRI Research Participants with Face-Recognition Software. N Engl J Med. 2019;381:1684–6. https://doi.org/10.1056/NEJMc1908881.
5. Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC 2022. https://www.cdc.gov/phlp/publications/topic/hipaa.html. Accessed 3 Jan 2024.
6. Voigt P, Von Dem Bussche A. The EU General Data Protection Regulation (GDPR). Cham: Springer International Publishing; 2017. https://doi.org/10.1007/978-3-319-57959-7.
7. Prayitno, Shyu C-R, Putra KT, Chen H-C, Tsai Y-Y, Hossain KSMT, et al. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. Appl Sci 2021;11:11191. https://doi.org/10.3390/app112311191.
8. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. Npj Digit Med. 2020;3:1–7. https://doi.org/10.1038/s41746-020-00323-1.
9. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. J Healthc Inform Res. 2021;5:1–19. https://doi.org/10.1007/s41666-020-00082-4.
10. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh A, Zhu J, editors. Proc 20th Int. Conf Artif Intell Stat. vol. 54, PMLR; 2017, p. 1273–82.
11. Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health. 2022;1: e0000033. https://doi.org/10.1371/journal.pdig.0000033.
12. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021;50:D20–6. https://doi.org/10.1093/nar/gkab1112.
13. Schotten M, el Aisati M, Meester WJN, Steiginga S, Ross CA. A Brief History of Scopus: The World's Largest Abstract and Citation Database of Scientific Literature. Res. Anal.: Auerbach Publications; 2017.
14. Frandsen TF, Eriksen MB, Hammer DMG, Christensen JB, Wallin JA. Using Embase as a supplement to PubMed in Cochrane reviews differed across fields. J Clin Epidemiol. 2021;133:24–31. https://doi.org/10.1016/j.jclinepi.2020.12.022.
15. Haddaway NR, Collins AM, Coughlin D, Kirk S. The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching.

PLoS ONE. 2015;10: e0138237. https://doi.org/10.1371/journal.pone.0138237.

16. Chowdhury D, Banerjee S, Sannigrahi M, Chakraborty A, Das A, Dey A, et al. Federated learning based Covid-19 detection. Expert Syst. 2023;40: e13173. https://doi.org/10.1111/exsy.13173.
17. Feki I, Ammar S, Kessentini Y, Muhammad K. Federated learning for COVID-19 screening from Chest X-ray images. Appl Soft Comput. 2021;106: 107330. https://doi.org/10.1016/j.asoc.2021.107330.
18. Abdul Salam M, Taha S, Ramadan M. COVID-19 detection using federated machine learning. PLoS ONE. 2021;16: e0252573. https://doi.org/10.1371/journal.pone.0252573.
19. Zhang L, Shen B, Barnawi A, Xi S, Kumar N, Wu Y. FedDPGAN: Federated Differentially Private Generative Adversarial Networks Framework for the Detection of COVID-19 Pneumonia. Inf Syst Front J Res Innov. 2021;23:1403–15. https://doi.org/10.1007/s10796-021-10144-6.
20. Zhang W, Zhou T, Lu Q, Wang X, Zhu C, Sun H, et al. Dynamic-Fusion-Based Federated Learning for COVID-19 Detection. IEEE Internet Things J. 2021;8:15884–91. https://doi.org/10.1109/JIOT.2021.3056185.
21. Li Z, Xu X, Cao X, Liu W, Zhang Y, Chen D, et al. Integrated CNN and Federated Learning for COVID-19 Detection on Chest X-Ray Images. IEEE/ACM Trans Comput Biol Bioinform 2022;PP. https://doi.org/10.1109/TCBB.2022.3184319.
22. Ho T-T, Tran K-D, Huang Y. FedSGDCOVID: Federated SGD COVID-19 Detection under Local Differential Privacy Using Chest X-ray Images and Symptom Information. Sensors. 2022;22:3728. https://doi.org/10.3390/s22103728.
23. Malik H, Naeem A, Naqvi RA, Loh W-K. DMFL_Net: A Federated Learning-Based Framework for the Classification of COVID-19 from Multiple Chest Diseases Using X-rays. Sensors. 2023;23:743. https://doi.org/10.3390/s23020743.
24. Dou Q, So TY, Jiang M, Liu Q, Vardhanabhuti V, Kaissis G, et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. Npj Digit Med. 2021;4:1–11. https://doi.org/10.1038/s41746-021-00431-6.
25. Florescu LM, Streba CT, Şerbănescu M-S, Mămuleanu M, Florescu DN, Teică RV, et al. Federated Learning Approach with Pre-Trained Deep Learning Models for COVID-19 Detection from Unsegmented CT images. Life. 2022;12:958. https://doi.org/10.3390/life12070958.
26. Kumar R, Khan AA, Kumar J, Zakria, Golilarz NA, Zhang S, et al. Blockchain-Federated-Learning and Deep Learning Models for COVID-19 Detection Using CT Imaging. IEEE Sens J 2021;21:16301–14. https://doi.org/10.1109/JSEN.2021.3076767.
27. Durga R, Poovammal E. FLED-Block: Federated Learning Ensembled Deep Learning Blockchain Model for COVID-19 Prediction. Front Public Health. 2022;10: 892499. https://doi.org/10.3389/fpubh.2022.892499.
28. Malik H, Anees T, Naeem A, Naqvi RA, Loh W-K. Blockchain-Federated and Deep-Learning-Based Ensembling of Capsule Network with Incremental Extreme Learning Machines for Classification of COVID-19 Using CT Scans. Bioeng Basel Switz. 2023;10:203. https://doi.org/10.3390/bioengineering10020203.
29. Peng L, Luo G, Walker A, Zaiman Z, Jones EK, Gupta H, et al. Evaluation of federated learning variations for COVID-19 diagnosis using chest radiographs from 42 US and European hospitals. J Am Med Inform Assoc JAMIA. 2022;30:54–63. https://doi.org/10.1093/jamia/ocac188.
30. Wang Z, Cai L, Zhang X, Choi C, Su X. A COVID-19 Auxiliary Diagnosis Based on Federated Learning and Blockchain. Comput Math Methods Med. 2022;2022:7078764. https://doi.org/10.1155/2022/7078764.
31. Liu B, Yan B, Zhou Y, Yang Y, Zhang Y. Experiments of Federated Learning for COVID-19 Chest X-ray Images 2020. https://doi.org/10.48550/arXiv.2007.05592.
32. Qayyum A, Ahmad K, Ahsan MA, Al-Fuqaha A, Qadir J. Collaborative Federated Learning for Healthcare: Multi-Modal COVID-19 Diagnosis at the Edge. IEEE Open J Comput Soc. 2022;3:172–84. https://doi.org/10.1109/OJCS.2022.3206407.
33. Kandati DR, Gadekallu TR. Federated Learning Approach for Early Detection of Chest Lesion Caused by COVID-19 Infection Using Particle Swarm Optimization. Electronics. 2023;12:710. https://doi.org/10.3390/electronics12030710.
34. Soltan AAS, Thakur A, Yang J, Chauhan A, D'Cruz LG, Dickson P, et al. Scalable federated learning for emergency care using low cost microcomputing: Real-world, privacy preserving development and evaluation of a COVID-19 screening test in UK hospitals 2023:2023.05.05.23289554. https://doi.org/10.1101/2023.05.05.23289554.
35. Chourasia P, Murad T, Tayebi Z, Ali S, Khan IU, Patterson M. Efficient classification of sars-cov-2 spike sequences using federated learning. In Annual International Conference on Information Management and Big Data. Cham: Springer Nature Switzerland. 2023;80–96.
36. Sim JJ, Zhou W, Chan FM, Annamalai MSMS, Deng X, Tan BHM, et al. CoVnita, an end-to-end privacy-preserving framework for SARS-CoV-2 classification. Sci Rep. 2023;13:7461. https://doi.org/10.1038/s41598-023-34535-8.
37. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021;27:1735–43. https://doi.org/10.1038/s41591-021-01506-3.
38. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach. JMIR Med Inform. 2021;9: e24207. https://doi.org/10.2196/24207.
39. Hoyos W, Aguilar J, Toro M. Federated learning approaches for fuzzy cognitive maps to support clinical decision-making in dengue. Eng Appl Artif Intell. 2023;123: 106371. https://doi.org/10.1016/j.engappai.2023.106371.
40. Kumaresan M, Kumar MS, Muthukumar N. Analysis of mobility based COVID-19 epidemic model using Federated Multitask Learning. Math Biosci Eng MBE. 2022;19:9983–10005. https://doi.org/10.3934/mbe.2022466.
41. Samuel O, Omojo AB, Onuja AM, Sunday Y, Tiwari P, Gupta D, et al. IoMT: A COVID-19 Healthcare System Driven by Federated Learning and Blockchain. IEEE J Biomed Health Inform. 2023;27:823–34. https://doi.org/10.1109/JBHI.2022.3143576.
42. Chen JJ, Chen R, Zhang X, Privacy PMA, Framework PFL, for COVID-19 Vulnerability Map Construction. ICC,. IEEE Int. Conf Commun. 2021;2021:1–6. https://doi.org/10.1109/ICC42927.2021.9500975.
43. Pang J, Huang Y, Xie Z, Li J, Cai Z. Collaborative city digital twin for the COVID-19 pandemic: A federated learning solution. Tsinghua Sci Technol 2021;26:759–71. https://doi.org/10.26599/TST.2021.9010026.
44. Wang J, Qian C, Cui S, Glass L, Ma F. Towards Federated COVID-19 Vaccine Side Effect Prediction. In: Amini M-R, Canu S, Fischer A, Guns T, Kralj Novak P, Tsoumakas G, editors. Mach. Learn. Knowl. Discov. Databases, Cham: Springer Nature Switzerland; 2023, p. 437–52. https://doi.org/10.1007/978-3-031-26422-1_27.
45. Kanani P, Marathe VJ, Peterson D, Harpaz R, Bright S. Private Cross-Silo Federated Learning for Extracting Vaccine Adverse Event Mentions. In: Kamp M, Koprinska I, Bibal A, Bouadi T, Frénay B, Galárraga L, et al., editors. Mach. Learn. Princ. Pract. Knowl. Discov. Databases, Cham: Springer International Publishing; 2021, p. 490–505. https://doi.org/10.1007/978-3-030-93733-1_37.
46. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med Image Anal 2020;65:101765. https://doi.org/10.1016/j.media.2020.101765.
47. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proc 20th Int Conf Artif Intell Stat. PMLR. 2017;54:1273–82.
48. Channa A, Popescu N, Skibinska J, Burget R. The Rise of Wearable Devices during the COVID-19 Pandemic: A Systematic Review. Sensors. 2021;21:5787. https://doi.org/10.3390/s21175787.
49. Mitratza M, Goodale BM, Shagadatova A, Kovacevic V, van de Wijgert J, Brakenhoff TB, et al. The performance of wearable sensors in the detection of SARS-CoV-2 infection: a systematic review. Lancet Digit Health. 2022;4:e370–83. https://doi.org/10.1016/S2589-7500(22)00019-X.
50. Sun S, Folarin AA, Ranjan Y, Rashid Z, Conde P, Stewart C, et al. Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19. J Med Internet Res. 2020;22: e19992. https://doi.org/10.2196/19992.
51. Khan LU, Saad W, Han Z, Hossain E, Hong CS. Federated Learning for Internet of Things: Recent Advances, Taxonomy, and Open Challenges. IEEE Commun Surv Tutor. 2021;23:1759–99. https://doi.org/10.1109/COMST.2021.3090430.
52. Zhang T, Gao L, He C, Zhang M, Krishnamachari B, Avestimehr AS. Federated Learning for the Internet of Things: Applications, Challenges, and Opportunities. IEEE Internet Things Mag. 2022;5:24–9. https://doi.org/10.1109/IOTM.004.2100182.

53. Zhu H, Xu J, Liu S, Jin Y. Federated learning on non-IID data: A survey. Neurocomputing. 2021;465:371–90. https://doi.org/10.1016/j.neucom.2021.07.098.
54. Li T, Sahu AK, Talwalkar A, Smith V. Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Process Mag. 2020;37:50–60. https://doi.org/10.1109/MSP.2020.2975749.
55. Baselli G, Codari M, Sardanelli F. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? Eur Radiol Exp. 2020;4:30. https://doi.org/10.1186/s41747-020-00159-0.
56. Medicine TLR. Opening the black box of machine learning. Lancet Respir Med. 2018;6:801. https://doi.org/10.1016/S2213-2600(18)30425-9.
57. 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. https://doi.org/10.1109/TIFS.2020.2988575.
58. van Rooden SM, van der Werff SD, van Mourik MSM, Lomholt F, Møller KL, Valk S, et al. Federated systems for automated infection surveillance: a perspective. Antimicrob Resist Infect Control. 2024;13:113. https://doi.org/10.1186/s13756-024-01464-8.
59. Weiner DL, Balasubramaniam V, Shah SI, Javier JR. COVID-19 impact on research, lessons learned from COVID-19 research, implications for pediatric research. Pediatr Res. 2020;88:148–50. https://doi.org/10.1038/s41390-020-1006-3.
60. Anahtar MN, Yang JH, Kanjilal S. Applications of Machine Learning to the Problem of Antimicrobial Resistance: an Emerging Model for Translational Research. J Clin Microbiol n.d.;59:e01260–20. https://doi.org/10.1128/JCM.01260-20.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.