Journal of the American Medical Informatics Association, 2023, **30(12)**, 2041–2049
https://doi.org/10.1093/jamia/ocad170
Advance access publication 28 August 2023
Review

Review

Federated and distributed learning applications for electronic health records and structured medical data: a scoping review

Siqi Li[1], Pinyan Liu[1], Gustavo G. Nascimento[2,3], Xinru Wang[1], Fabio Renato Manzolli Leite[2,3], Bibhas Chakraborty[1,4,5,6], Chuan Hong[6], Yilin Ning[1], Feng Xie[1,4], Zhen Ling Teo[7], Daniel Shu Wei Ting[1,7], Hamed Haddadi[8], Marcus Eng Hock Ong[4,9], Marco Aurélio Peres[2,3,4], Nan Liu[1,4,10,]*

[1]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore 169857, Singapore, [2]National Dental Research Institute Singapore, National Dental Centre Singapore, Singapore 168938, Singapore, [3]Oral Health Academic Clinical Programme, Duke-NUS Medical School, Singapore 169857, Singapore, [4]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore 169857, Singapore, [5]Department of Statistics and Data Science, National University of Singapore, Singapore 117546, Singapore, [6]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27708, United States, [7]Singapore National Eye Centre, Singapore, Singapore Eye Research Institute, Singapore 168751, Singapore, [8]Department of Computing, Imperial College London, London SW7 2AZ, England, United Kingdom, [9]Department of Emergency Medicine, Singapore General Hospital, Singapore 169608, Singapore, [10]Institute of Data Science, National University of Singapore, Singapore 117602, Singapore

***Corresponding author:** Nan Liu, Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore (liu.nan@duke-nus.edu.sg)

Author Contributions: S. Li and P. Liu contributed equally and are considered joint first authors of this work. M.A. Peres and N. Liu contributed equally and are considered joint senior authors of this work.

Abstract

Objectives: Federated learning (FL) has gained popularity in clinical research in recent years to facilitate privacy-preserving collaboration. Structured data, one of the most prevalent forms of clinical data, has experienced significant growth in volume concurrently, notably with the widespread adoption of electronic health records in clinical practice. This review examines FL applications on structured medical data, identifies contemporary limitations, and discusses potential innovations.

Materials and methods: We searched 5 databases, SCOPUS, MEDLINE, Web of Science, Embase, and CINAHL, to identify articles that applied FL to structured medical data and reported results following the PRISMA guidelines. Each selected publication was evaluated from 3 primary perspectives, including data quality, modeling strategies, and FL frameworks.

Results: Out of the 1193 papers screened, 34 met the inclusion criteria, with each article consisting of one or more studies that used FL to handle structured clinical/medical data. Of these, 24 utilized data acquired from electronic health records, with clinical predictions and association studies being the most common clinical research tasks that FL was applied to. Only one article exclusively explored the vertical FL setting, while the remaining 33 explored the horizontal FL setting, with only 14 discussing comparisons between single-site (local) and FL (global) analysis.

Conclusions: The existing FL applications on structured medical data lack sufficient evaluations of clinically meaningful benefits, particularly when compared to single-site analyses. Therefore, it is crucial for future FL applications to prioritize clinical motivations and develop designs and methodologies that can effectively support and aid clinical practice and research.

Key words: clinical decision-making; distributed algorithms; distributed learning; electronic health records; federated learning.

Introduction

The digitization of electronic health records (EHRs) has facilitated data analysis from multiple centers. This enables comparisons across populations and settings, as well as data combinations that enhance statistical power and generalizability.[1] Data sharing, a conventional approach in the healthcare industry for forming cross-regional partnerships, has proven beneficial for research reproducibility, cost-efficiency, redundancy prevention, and accelerating discovery and innovation.[2] However, such cooperation raises data privacy concerns[3],[4] and can be difficult due to privacy regulations like the European Union's General Data Protection Regulation,[5] making privacy-preserving techniques a critical area of interest in this field.

Federated learning (FL) has gained popularity in healthcare as a technique for maintaining privacy.[6] FL is a machine learning setting where multiple entities (clients) collaborate in solving a modeling problem, with each client's data stored locally and not exchanged or transferred.[7] Prior to the introduction of the term FL,[8] statisticians had been researching privacy-preserving statistical algorithms using terms such as

Received: May 19, 2023. **Revised:** July 19, 2023. **Editorial Decision:** August 12, 2023

"distributed learning"[9,10] or "distributed algorithms".[11,12] Despite differences in terminology, these algorithms all contribute to reducing barriers posed by privacy regulations across different countries and regions, enabling researchers to collaborate more effectively and efficiently. It is important to note that this article focuses only on privacy-preserving distributed learning and algorithms, and does not include broader definitions that may encompass methods for reducing computational costs.

Although there exist a number of FL applications in the medical field, most of these early adoptions have focused on unstructured data, particularly image data.[13] Structured data, which constitute a significant part of clinical data with the advent of large-scale EHRs, have been relatively underexplored in FL settings. FL for structured data differs from FL for image data in several ways, including sample sizes, data structures, modeling methodologies, research topics, and study designs.

Although several reviews have been conducted on FL applications in healthcare, they often discuss clinical data in general terms without delving into specific data types.[9,13,14] Furthermore, while some of these reviews address technical issues in-depth, they fall short in presenting their conclusions from the perspective of clinical applications. In response, we conducted a scoping review to summarize and examine current FL applications on structured clinical data, emphasizing the advantages of FL in clinical research. Our aim is to provide insights and suggestions for future FL applications in clinical decision-making.

Materials and methods

Search strategy and selection criteria

We conducted a review following the 2020 PRISMA[15] guidelines for systematic reviews. We searched for published articles employing FL frameworks to solve clinical/biomedical questions using structured data. We searched the SCOPUS, MEDLINE, Web of Science, Embase, and CINAHL databases for articles published before August 23, 2022, utilizing a combination of search terms, including "electronic health records," "EHR," "electronic medical records," "EMR," "registry/registries," "tabular," "federated learning," "distributed learning," and "distributed algorithms." A detailed search strategy is presented in eTable S1 of the Supplementary Material.

The final search was conducted on August 23, 2022. After removing duplicates, 2 reviewers (S.L. and P.L.) independently screened articles based on their titles and abstracts, with a third reviewer (F.X.) resolving any conflicts. The publications selected in the first round of screening underwent full-text examination to ensure they met the inclusion criteria: using structured data, employing FL, being a research article, having full text available, and addressing biomedical/clinical research data or questions.

Data extraction

We extracted information from the selected publications from 3 perspectives: data (cohort descriptive analysis, outcome, sample size per site and total, number of participating sites, data types, data public availability, and number of features), modeling (task and goal, modeling approach, hyperparameter methods, model performance metrics), and FL frameworks (unit of federation, participating countries/regions, FL structure, FL topology, one-shot or not, evaluation metrics, convergence analysis, solution for heterogeneity, FL and local model comparison, and code availability). This process was conducted collaboratively by multiple reviewers to ensure accuracy and consistency.

Results

Our search strategy yielded 1193 articles, and after removing 624 duplicate records, 569 articles were screened based on title and abstract. Sixty-two out of 569 articles were left for full-text screening, and 34 articles were included for this review. Figure 1 presents the PRISMA flow diagram, which contains detailed information about the selection process. One article may present more than one studies, depending on datasets, models, and FL frameworks applied. For instance, Halim et al[16] carried out 2 studies using the same dataset but different outcomes (1 binary and 1 multilabel); Sadilek et al[17] undertook 7 studies with 7 different datasets. Consequently, as shown and summarized in Table 1, the total number of studies for all 34 included papers is 72, and full details can be found in eTable S2 of the Supplementary Material.

General characteristics of included papers

Out of the 34 papers reviewed, 29 were published in 2020 or after, indicating a recent surge of interest in FL for clinical research with structured data. The total sample sizes reported for each study mainly consisted of unique patients with occasional overlaps, but the extent of the overlaps is unknown due to the lack of information. The total sample size varied widely, ranging from 141[33] to 4 408 710,[27] and the number of clients (participating sites) ranged from 2[29,37,38,40] to 314.[11] The participating clients used either artificially partitioned datasets or real isolated datasets. Most of the studies used the horizontal FL approach, where datasets share the same feature space but have different samples,[41] while only one article[39] assumed a vertical FL setting where all datasets share the same sample space but have different features.[41]

Dataset characteristics

Twenty-five out of 34 papers used structured data derived from EHRs, primarily containing patient demographics, vital signs, laboratory results, and other features commonly found in hospital records. Five articles[29,34,38,42,43] studied clinical cohorts that underwent long-term interventions or follow-up. Only 38.9% (28/72) of the studies provided descriptive analyses of their data, and 62.5% (45/72) used publicly accessible datasets. Five studies used datasets with more than 1000 features, while the remaining studies mostly (58.3%, or 42/72) used data with fewer than 41 features.

Modeling characteristics

In this review, we classified a study as "prediction" if its primary goal was to predict an outcome and report performance metrics, and a study as "association modeling" if its primary goal was to investigate the relationship between covariates and outcome(s) by reporting estimates of coefficients or odds ratios. These 2 types of studies differ in statistical inference, with the latter focusing on uncertainty measurements while the former does not. Another type of task is phenotyping, which is unsupervised and aimed at deriving research-grade phenotypes from clinical data.[44] Out of all 72 studies, 40

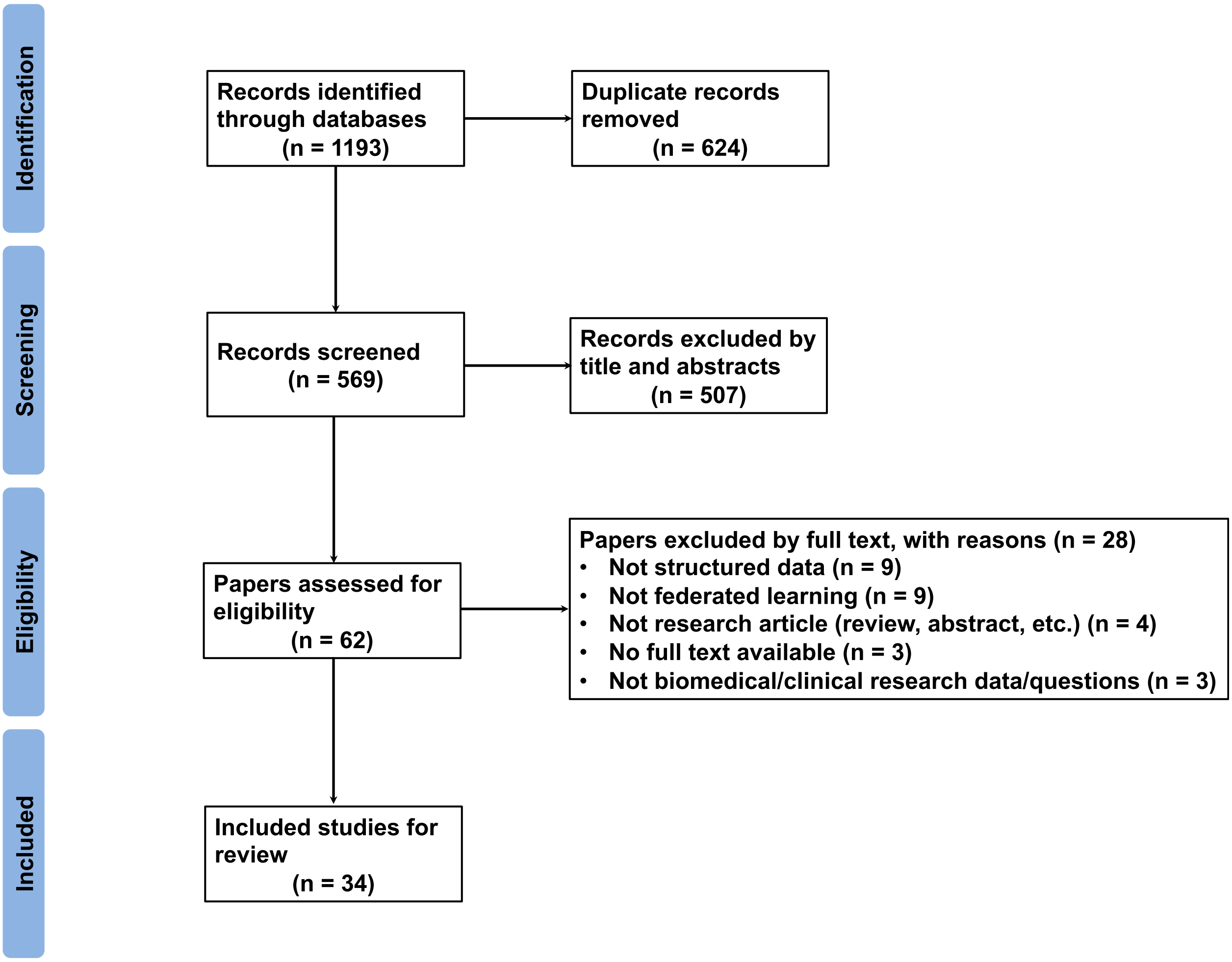

Identification
Screening
Eligibility
Included
Records identified through databases (n = 1193)
Duplicate records removed (n = 624)
Records screened (n = 569)
Records excluded by title and abstracts (n = 507)
Papers assessed for eligibility (n = 62)
Papers excluded by full text, with reasons (n = 28)
• Not structured data (n = 9)
• Not federated learning (n = 9)
• Not research article (review, abstract, etc.) (n = 4)
• No full text available (n = 3)
• Not biomedical/clinical research data/questions (n = 3)
Included studies for review (n = 34)

Figure 1. Preferred reporting items for systematic reviews (PRISMA) flow diagram.

(55.6%) performed prediction tasks, 20 (27.8%) investigated associations, and 9 (12.5%) conducted phenotyping. A majority of the studies (69.4%, or 50/72) investigated binary outcomes, and logistic regression was the most frequently used model (25.0%, or 18/72). Various types of neural networks were used in 26.4% (19/72) of the studies.

FL architectures

We categorized FL algorithms into one-shot and non-one-shot based on whether a FL algorithm required one or multiple rounds of communication (intermediate parameters needed to be transferred once or more times among participants). Our observations showed that 26.4% (19/72) of the studies used one-shot FL and 70.8% (51/72) used non-one-shot FL, with 74.5% (38/51) adopting a centralized FL topology and 25.5% (13/51) using a decentralized FL topology. We summarized the frameworks used in all studies and created a plot (Figure 2) to illustrate their applications and interdependencies. Of all studies, 27.8% (20/72) utilized FL frameworks that adopted one or more specific solutions to address data heterogeneity.

FL performance evaluation

Based on the articles reviewed for this study, FL frameworks were frequently evaluated for their computational performance, such as rounds of communications, number of iterations, and computation time, but less frequently on their performance compared to conventional single-site analyses. Only 41.2% (14/34) of the papers specifically discussed comparisons between local and FL models, and among these, 11 reported that FL models had some advantages over local models. Most papers compared local and FL models by directly comparing the same performance metrics achieved by each model on given testing sets, and only a few studies, such as Dayan et al,[19] reported generalizability of a model using average performance across different sites.

Discussion

The use of FL for structured medical data has been extensively explored, but certain aspects of such studies require special attention. Our review suggests that FL is sometimes applied without first assessing its actual benefits. A crucial consideration is whether FL can advance medical research objectives unattainable through single-site analyses. In this section, we provide a detailed discussion on when healthcare researchers should consider FL, additional precautions that should be taken, and technical details related specifically to clinical structured data.

Is FL necessary?

Before diving into technical details, we examine the necessity of FL for conducting research on structured medical data. We

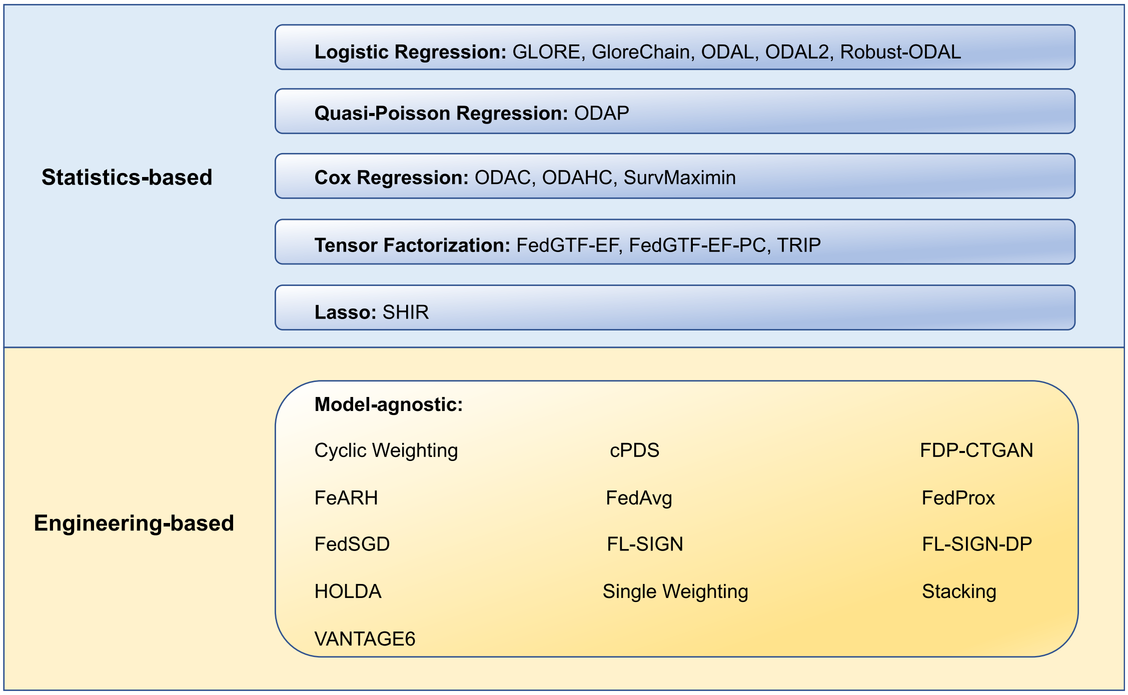
Table 1. Summary of information extraction table.

Data characteristics	No. of studies (out of 72)	Examples
Provide cohort descriptive analysis	28 (38.9%)	A summary statistics table describing patient population was provided in Ref.[18]
Patient cohort size (total)		
≤50 000	53 (73.6%)	The total patient cohort size used in Ref.[19] was 16 148
50 000-100 000	3 (4.2%)	The total patient cohort size used (in the first study) in Ref.[18] was 70 818
100 000	11 (15.3%)	The total patient cohort size used in Ref.[20] was 257 571
Not available	5 (6.9%)	No such details provided in Ref.[21]
No. clients		
≤5	35 (48.6%)	The number of clients in Ref.[22] was 8
6-35	20 (27.8%)	The number of clients in Ref.[23] was 20
35	9 (12.5%)	The number of clients in Ref.[24] was 314
Cross-device	4 (5.6%)	The first, second, fifth, and sixth studies in Ref.[17] used cross-device (cross-patient) setting
Not available	4 (5.6%)	No such details provided in Ref.[25]
No. features		
≤40	42 (58.3%)	The number of features in Ref.[26] was 23
41-1000	10 (13.9%)	The number of features in Ref.[27] was 85
1000	5 (6.9%)	The number of features in Ref.[22] was 2931
Not available	15 (20.8%)	No such details provided in Ref.[28]
Outcome		
Binary	50 (69.4%)	In-hospital mortality was used as outcome in Ref.[24]
Survival	5 (6.9%)	Cancer survival time was used as outcome in Ref.[29]
Other	16 (22.2%)	Frequency of serious adverse events (count) was used as outcome in Ref.[18]
Not available	1 (1.4%)	No such details provided in Ref.[25]
Data public availability		
Yes	45 (62.5%)	eICU data was used in Ref.[30]
No	27 (37.5%)	Data used in Ref.[11] is not publicly available
Model characteristics	**No. of studies (out of 72)**	**Examples**
Task		
Prediction	40 (55.6%)	Prediction of COVID mortality risk[1]
Association study	20 (27.8%)	Association between length of stay in COVID-19 patients with various patient characteristics[31]
Phenotyping	9 (12.5%)	Extraction of meaningful medical concepts[32]
Model misconduct detection	3 (4.2%)	A generalizable approach to identify model misconducts in FL[33]
Modeling approach		
Logistic regression	18 (25.0%)	Proposed and evaluated a one-shot distributed algorithm for logistic regression for heterogenous data[20]
Cox regression	6 (8.3%)	Proposed and evaluated a one-shot distributed algorithm for Cox regression[34]
Neural networks (NN)		
(1) General	10 (13.9%)	3-layer fully connected NN[22]
(2) Autoencoder	4 (5.6%)	5-layer fully connected denoising autoencoder[30]
(3) Perceptron	3 (4.2%)	3-layer MLP[26]
(4) Deep learning	2 (2.8%)	TabNet[35]
SVM	4 (5.6%)	Proposed and evaluated a FL framework for soft-margin l_1- regularized sparse SVM[36]
Tensor factorization	9 (12.5%)	Proposed and evaluated a federated tensor factorization method[37]
Other	15 (20.8%)	XGBoost[38]
Not available	1 (1.4%)	No such details provided in Ref.[25]
FL characteristics	**No. of studies (out of 72)**	**Examples**
Unit of federation		
Cross silo	64 (88.9%)	Three healthcare facilities within the OneFlorida Clinical Research Consortium[31]
Cross patient (device)	4 (5.6%)	Four studies reported cross patient results[17]
Both	3 (4.2%)	Three studies reported both cross patient and cross silo results[17]
Not available	1 (1.4%)	No such details provided in Ref.[25]
Participants		
Real isolated sites	37 (51.4%)	Data from 20 institutes across the globe[19]
Artificial partitions	32 (44.4%)	Ten sites formed by random splitting one dataset[36]
Not available	3 (4.2%)	No such details provided in Ref.[23]
FL topology		
Centralized	38 (52.8%)	FL-SIGN and FL-SIGN-DP[24]
Decentralized	13 (18.1%)	Decentralized stochastic gradient descent (DSGD) and tracking (DSGT)[23]
One-shot	19 (26.4%)	ODAL[11]
Not available	2 (2.8%)	No such details provided in Ref.[39]
Solution(s) for heterogenous data		
Yes	20 (27.8%)	Employed HinSAGE which introduces extra weight matrices for heterogeneous graph[16]
No	52 (72.2%)	Not available

(continued)

Table 1. (continued)

FL characteristics	No. of studies (out of 72)	Examples
Local vs FL comparison		
Yes	37 (51.4%)	Compared performances of local and FL models, where FL model outperformed local models[19]
No	33 (45.8%)	Not available
Not applicable	2 (2.8%)	Vertical instead of horizontal FL was used in Ref.[39]



Statistics-based

Logistic Regression: GLORE, GloreChain, ODAL, ODAL2, Robust-ODAL

Quasi-Poisson Regression: ODAP

Cox Regression: ODAC, ODAHC, SurvMaximin

Tensor Factorization: FedGTF-EF, FedGTF-EF-PC, TRIP

Lasso: SHIR

Engineering-based

Model-agnostic:

Cyclic Weighting, cPDS, FDP-CTGAN, FeARH, FedAvg, FedProx, FedSGD, FL-SIGN, FL-SIGN-DP, HOLDA, Single Weighting, Stacking, VANTAGE6

Figure 2. Visualization of FL frameworks utilized in the 34 papers included in this review, classified into 2 categories: statistics-based and engineering-based FL.

will begin by discussing why and how FL is anticipated to yield additional meaningful results beyond those produced by pre-existing local models. We will then discuss how data can affect the feasibility and efficacy of FL.

Cross-silo and cross-device are 2 distinct FL settings that can significantly impact the design of an FL algorithm.[45] Cross-silo design aims to facilitate collaboration among various organizations, such as hospitals and research institutions, while cross-device design enables collaboration among large populations of mobile devices.[45] When designing and implementing cross-silo FL, it is important to consider the presence of pre-existing local models, in addition to the technical distinctions highlighted in Wang et al.[45] This is particularly crucial given the vast amount of information available in large-scale EHRs today.

The comparison of the performance of FL models to local models has therefore become a natural point of discussion. Some papers that benchmark or develop FL frameworks in a broader context have highlighted the relevance and importance of such comparisons.[46,47] However, in this review, such comparisons were only observed in 14 out of 34 papers. For example, Cui et al[22] developed a new FL framework called FeARH and assessed its performance using artificially partitioned data. Although they provided a performance comparison of the central, FL baseline, and FeARH models, future researchers may find such demonstrations inadequate, as obtaining a ground truth central model is often impossible in real-world practice.

Unlike unstructured data, such as images, structured clinical data often contain features that can vary at the definition stage due to differences in clinical practice across institutions.[48,49] For example, diabetes is characterized by elevated levels of glucose in the bloodstream, and it can be diagnosed by fasting or random plasma glucose, each of which has a different cut-off point. Such inconsistency in disease diagnosis may lead to hidden heterogeneity, which might go undetected in downstream statistical analysis. Therefore, proper preparation and data harmonization across participating sites are required prior to implementing FL. The 4CE consortium,[50] which offers standardized patient-level EHRs that cover various aspects of COVID-19, such as epidemiology and pathophysiology,[51] is a notable example of data harmonization and standardization. Several clinically significant studies[51–53] were made possible by the consortium, one of which[1] being relevant to FL and included in this review. The existence and accomplishments of the 4CE consortium underscore the inevitability and significance of data harmonization for conducting clinically meaningful and trustworthy FL on structured clinical data. However, existing works have fallen short in addressing this problem, particularly those that rely on benchmark or artificially partitioned data.

Based on the findings of this review, we recommend that 2 broad goals should be achieved for FL applications with structured healthcare data. First, FL models should outperform locally developed models at least at one participating site. This can be demonstrated by improved accuracy in prediction tasks, narrower confidence intervals in effect size estimations, or the discovery of new phenotypes in phenotyping tasks. For instance, Dayan et al[19] demonstrated that their best FL model for predicting 24-h oxygen treatment for COVID-19 patients outperformed all 20 local models trained independently at each site. In another example, Kim et al[37] proposed a federated tensor factorization framework that successfully identified new phenotypes that were not captured in any of the local phenotyping analyses.[37] Second, for prediction tasks in particular, FL models should exhibit better stability and generalizability than at least one local model, as demonstrated by achieving lower performance variation than some local models. Dayan et al[19] demonstrated that the best FL model exhibited the highest level of generalizability among all local models, as measured by average area under the curve (AUC).

Technical details and challenges

FL algorithms: statistical versus engineering approaches

Both the statistics and engineering communities have conducted research on FL, with the significant distinction between them being the property of model agnosticism. Statistics-based FL algorithms usually involve model-specific statistical modeling, meaning that a single algorithm is typically applied to one type of model. For instance, as illustrated in Figure 2, a federated Cox regression might only be conducted using statistics-based FL methods such as ODAC,[34] ODACH,[43] and SurvMaximin.[1] By contrast, most engineering-based approaches have been developed in a model-agnostic manner, allowing for the use of a single FL framework for different machine learning models by employing the appropriate loss functions. For instance, in Sadilek et al,[17] FedAvg[8] has been applied to neural networks, logistic regression, and generalized linear models for binomial responses with log link.

To the best of our knowledge, the current literature lacks direct comparisons between statistics-based and engineering-based FL algorithms, and their precise advantages and disadvantages have not been thoroughly assessed. One apparent advantage of engineering-based methods is their model-agnostic property, which allows these algorithms to be directly applied to a wide range of commonly used models without the need for additional designs. Furthermore, engineering-based solutions might be more resilient to model misspecification, as most existing statistics-based FL algorithms focus on linear relationships that may not hold for real-world data.

Although statistics-based FL algorithms require one-to-one development and validation, they have an advantage over engineering-based FL in uncertainty measurements. In FL studies that aim to estimate parameters of interest such as the association between exposure and outcome, it is desirable to report the associated confidence intervals. Statistics-based FL methods have an advantage in estimating uncertainty measures in a distributed manner when the asymptotic distribution of an estimator is available, and a closed-form formula for the variance of estimated parameters exists.[31,34,43,54] In the absence of these conditions, bootstrap can serve as a flexible alternative for estimating standard errors for both statistics- and engineering-based FL. However, it is important to note that a simple bootstrap strategy may yield inconsistent results if the estimator is nonsmooth, necessitating additional measures to ensure effectiveness.[55] While only a few existing FL studies have examined the uncertainty of estimated associations, future investigations should consider these potential limitations.

Lastly, it is worth noting that the implementation of existing packages of engineering-based and statistics-based FL algorithms can present different levels of difficulty when applied to structured clinical data. The programming language difference between statistics-based and engineering-based FL algorithms (R or Python, etc.), can potentially pose challenges for users, particularly for biostatisticians and epidemiologists who commonly use R for data analysis. Additionally, engineering-based FL frameworks, being model agnostic, can be relatively more difficult to adapt, as many usage demos are applied to unstructured data, particularly images, making it inconvenient for users to directly apply them to their own models. Furthermore, the requirement for hyperparameter tuning in engineering-based methods can add complexity to the adoption process.

Statistical heterogeneity: challenges and potential benefits

Engineering-based FL typically defines statistical heterogeneity as scenarios where data are not independently and identically distributed (i.i.d.).[45,56] However, the statistical literature usually distinguishes between heterogeneity in covariate effects (conditional distribution of $Y|X$) and heterogeneity in covariates distributions ($P(X)$), as they have different impact on model building and inference.[57–59] The differing viewpoints[6,60,61] surrounding the impact of non-i.i.d. data on FL algorithms assuming i.i.d. may be explained by the failure to identify the source of heterogeneity. In this review, for instance, Vaid et al[26] found that FL models obtained via FedAvg outperformed local models when predicting COVID-19 outcomes, despite heterogeneity in patient demographics and outcome prevalence across sites. This result may be attributed to the homogeneity of the conditional distribution $Y|X$ across sites, even if $P(X)$ is not homogeneous. Unlike the covariate distribution $P(X)$, the conditional distribution $Y|X$ across different populations could be complex and difficult to specify correctly.[57] Given the complexity of real-world data, it is challenging to predict the suitability of classic i.i.d.-based FL frameworks without sufficient empirical evidence. Therefore, future researchers are recommended to benchmark these frameworks with heterogeneous data to assess their effectiveness.

It is also noteworthy that while heterogeneity is generally considered a challenge for supervised learning models, there is evidence to suggest that it can be beneficial for certain unsupervised tasks.[62] Specifically, the proposed federated clustering algorithm by Dennis et al[62] demonstrated improved separations on heterogeneously partitioned data when compared to i.i.d.-partitioned data, as evidenced by achieving a cost that was closer to the original oracle clustering.[62]

Convergence analysis for optimization

Almost all FL algorithms aim to estimate the parameters of interest by solving an optimization problem, making convergence crucial for ensuring accurate estimation. If an FL algorithm frequently fails to converge, it may not be suitable for real-world data. However, convergence analysis has not been adequately addressed in the current FL literature on structured healthcare data. Of the 19 papers in this review proposing new FL frameworks, only 5[23,24,32,36,37] discussed framework convergence. Stochastic gradient descent (SGD) is a popular choice for smooth optimization when the assumptions such as the existence of lower bounds, Lipschitz smoothness, and bounded variance are met.[63] However, SGD-based FL algorithms do not function well in nonsmooth cases, requiring the development of unique strategies as seen in some studies.[32,36]

Data privacy: differential privacy

Although FL enables model training without exchanging or sharing data, adversaries can still analyze the differences in related parameters trained and uploaded during the FL process to obtain private information.[64] As an example, Hitaj et al[65] proposed a generative adversarial networks (GAN)-based reconstruction attack against convolutional neural networks trained on image data, in which the trained generator successfully mimics the training samples.[66] To further address this data leakage issue, differential privacy (DP) techniques[67] have been integrated with FL frameworks to add artificially controlled noise[14] before, during, or after model training.[68] DP has also been combined with GANs to generate synthetic data that can be shared with collaborators. This approach is particularly useful for structured data,[25] since GANs have a greater flexibility in modeling distributions compared to their statistical counterparts.[69]

While only 5 articles[17,19,24,25,35] in this review addressed the integration of DP with FL, broader reviews of FL in computer science[68,70] have shown that it is a popular topic for general FL applications. As FL continues to be adopted for large-scale and widespread medical applications, the integration of FL with technologies such as DP is likely to become increasingly important in the future.

Standardized pipelines and domain-specific FL for structured medical data

Our review found that many existing FL studies lack sufficient detail on model selection and data preprocessing procedures. The unavailability of open-source codes further obstructs the reproducibility and future applications of FL. To tackle these issues, we recommend fostering close collaborations between FL researchers and healthcare professionals to create more standardized processes and assessment techniques in future studies. Additionally, we encourage future researchers to emphasize the importance of benchmark datasets and evaluations, in order to develop standardized, open-source pipelines that could accelerate and facilitate research in the field.

We also observed that all the reviewed papers have focused on modeling tasks, thereby overlooking the potential of FL to enhance nonmodeling aspects of healthcare research. A pertinent example is Zhou et al,[71] which proposed using an FL-based generative adversarial nets for missing value imputation. Although this example comes from a nonhealthcare domain, few studies have employed similar techniques to process healthcare data. Nonetheless, these methods hold promise if they can reduce bias in data analysis, ultimately leading to more robust and trustworthy healthcare findings.

Limitations

This review aims to provide a comprehensive overview of the applications of FL on structured clinical data. As such, we did not delve into detailed analyses of the mathematical and technological aspects of FL frameworks, especially those related to optimization.

Conclusion

The application of FL on structured medical data is still in its early stages. Most studies primarily focus on prediction tasks and often lack robust demonstrations of clinically significant results. Further exploration combining engineering- and statistics-based FL algorithms may present novel opportunities. Additionally, this review underscores the importance of establishing standardized methodologies and protocols, as well as promoting the release of open-source codes, to ensure reproducibility and transparency in future FL research in healthcare.

Author contributions

N.L. and S.L. conceived and designed the study. S.L. and P.L. performed literature search, abstract and full-text screening, and analyzed the data. G.G.N. and F.R.M.L. provided expertise on health services research and healthcare delivery. S.L., P.L., G.G.N., X.W., and N.L. drafted the manuscript. All authors commented on versions of the article. All authors read and provided comments or approved the final manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the Duke/Duke-NUS Collaboration grant. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Conflicts of interest

None declared.

Data availability

There are no new data associated with this article.

References

1. Wang X, Zhang HG, Xiong X, et al.; Consortium for Clinical Characterization of COVID-19 by EHR 4CE. SurvMaximin: robust federated approach to transporting survival risk prediction models. *J Biomed Inform*. 2022;134:104176.
2. van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health*. 2014;14:1144.
3. Antunes RS, André Da Costa C, Küderle A, et al. Federated learning for healthcare: systematic review and architecture proposal. *ACM Trans Intell Syst Technol*. 2022;13(4):1-23.

4. Nguyen DC, Pham Q-V, Pathirana PN, et al. Federated learning for smart healthcare: a survey. *ACM Comput Surv*. 2023;55(3):1-37.
5. Hoofnagle CJ, van der Sloot B, Borgesius FZ. The European Union general data protection regulation: what it is and what it means. *Inf Commun Technol Law*. 2019;28(1):65-98.
6. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3:1-7.
7. Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. 2021. http://arxiv.org/abs/1912.04977, October 31, 2022, preprint: not peer reviewed.
8. McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR; 2017:1273-1282. Accessed July 5, 2022. https://proceedings.mlr.press/v54/mcmahan17a.html
9. Kirienko M, Sollini M, Ninatti G, et al. Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI. *Eur J Nucl Med Mol Imaging*. 2021;48(12):3791-3804.
10. Jochems A, Deist TM, van Soest J, et al. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital – a real life proof of concept. *Radiother Oncol*. 2016;121(3):459-467.
11. Duan R, Boland MR, Moore JH, et al. ODAL: a one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pac Symp Biocomput*. 2019;24:30-41.
12. Gao Y, Liu W, Wang H, et al. A review of distributed statistical inference. *Stat Theory Relat Fields*. 2022;6(2):89-99.
13. Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. *PLoS Digit Health*. 2022;1(5):e0000033.
14. Shyu C-R, Putra KT, Chen H-C, et al. A systematic review of federated learning in the healthcare area: from the perspective of data properties and applications. *Appl Sci*. 2021;11(23):11191.
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg*. 2021;88:105906.
16. Halim SM, Khan L, Hamlen KW, et al. A federated approach for learning from electronic health records. In: *2022 IEEE 8th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. 2022:218-223. https://doi.org/10.1109/BigDataSecurityHPSCIDS54978.2022.00049
17. Sadilek A, Liu L, Nguyen D, et al. Privacy-first health research with federated learning. *NPJ Digit Med*. 2021;4(1):132-138.
18. Edmondson MJ, Luo C, Duan R, et al. An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes. *Sci Rep*. 2021;11(1):19647.
19. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;27(10):1735-1743.
20. Tong J, Duan R, Li R, et al. Robust-ODAL: learning from heterogeneous health systems without sharing patient-level data. *Pac Symp Biocomput Pac Symp Biocomput*. 2020;25:695-706.
21. Kavitha Bharathi S, Dhavamani M, Niranjan K. A federated learning based approach for heart disease prediction. In: *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*. 2022:1117-1121. https://doi.org/10.1109/ICCMC53470.2022.9754119
22. Cui J, Zhu H, Deng H, et al. FeARH: federated machine learning with anonymous random hybridization on electronic medical records. *J Biomed Inform*. 2021;117:103735.
23. Lu S, Zhang Y, Wang Y. Decentralized federated learning for electronic health records. In: *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. 2020:1-5. https://doi.org/10.1109/CISS48834.2020.1570617414
24. Kerkouche R, Ács G, Castelluccia C, et al. Privacy-preserving and bandwidth-efficient federated learning: an application to in-hospital mortality prediction. In: *Proceedings of the Conference on Health, Inference, and Learning*. ACM; 2021:25-35. https://doi.org/10.1145/3450439.3451859
25. Fang ML, Dhami DS, Kersting K. DP-CTGAN: differentially private medical data generation using CTGANs. In: Michalowski M, Abidi SSR, Abidi S, eds. *Artificial Intelligence in Medicine*. Springer International Publishing; 2022:178-88. https://doi.org/10.1007/978-3-031-09342-5_17
26. Vaid A, Jaladanki SK, Xu J, 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(1):e24207.
27. Fontana M, Naretto F, Monreale A. A new approach for cross-silo federated learning and its privacy risks. In: *2021 18th International Conference on Privacy, Security and Trust (PST)*. 2021:1-10. https://doi.org/10.1109/PST52912.2021.9647753
28. Choudhury O, Park Y, Salonidis T, et al. Predicting adverse drug reactions on distributed health data using federated learning. *AMIA Annu Symp Proc*. 2020;2019:313-322.
29. Geleijnse G, Chiang RC-J, Sieswerda M, et al. Prognostic factors analysis for oral cavity cancer survival in the Netherlands and Taiwan using a privacy-preserving federated infrastructure. *Sci Rep*. 2020;10(1):20526.
30. Huang L, Shea AL, Qian H, et al. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J Biomed Inform*. 2019;99:103291.
31. Edmondson MJ, Luo C, Nazmul Islam M, et al. Distributed Quasi-Poisson regression algorithm for modeling multi-site count outcomes in distributed data networks. *J Biomed Inform*. 2022;131:104097.
32. Ma J, Zhang Q, Lou J, et al. Communication efficient federated generalized tensor factorization for collaborative health data analytics. *Proc Int World-Wide Web Conf Int WWW Conf*. 2021;2021:171-182.
33. Kuo T-T, Pham A. Detecting model misconducts in decentralized healthcare federated learning. *Int J Med Inform*. 2021;158:104658.
34. Duan R, Luo C, Schuemie MJ, et al. Learning from local to global: an efficient distributed algorithm for modeling time-to-event data. *J Am Med Inform Assoc*. 2020;27(7):1028-1036.
35. Mehta J, Desai R, Mehta J, et al. Towards a multi-modular decentralized system for dealing with EHR data. In: *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2022:567-572. https://doi.org/10.1109/ICACCS54159.2022.9785302
36. Brisimi TS, Chen R, Mela T, et al. Federated learning of predictive models from federated electronic health records. *Int J Med Inform*. 2018;112:59-67.
37. Kim Y, Sun J, Yu H, et al. Federated tensor factorization for computational phenotyping. *KDD Proc Int Conf Knowl Discov Data Min*. 2017;2017:887-895.
38. Lopes RR, Mamprin M, Zelis JM, et al. Local and distributed machine learning for inter-hospital data utilization: an application for TAVI outcome prediction. *Front Cardiovasc Med*. 2021;8:787246.
39. Cha D, Sung M, Park Y-R. Implementing vertical federated learning using autoencoders: practical application, generalizability, and utility study. *JMIR Med Inform*. 2021;9(6):e26598.
40. Rajendran S, Obeid JS, Binol H, et al. Cloud-based federated learning implementation across medical centers. *JCO Clin Cancer Inform*. 2021;5:1-11.
41. Yang Q, Liu Y, Chen T, et al. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol*. 2019;10:1–19. doi:10.1145/3298981.
42. Duan R, Chen Z, Tong J, et al. Leverage real-world longitudinal data in large clinical research networks for Alzheimer's disease and

related dementia (ADRD). *AMIA Annu Symp Proc*. 2021;2020:393-401.

43. Luo C, Duan R, Naj AC, et al. ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. *Sci Rep*. 2022;12(1):6627.
44. Richesson RL, Sun J, Pathak J, et al. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med*. 2016;71:57-61.
45. Wang J, Charles Z, Xu Z, et al. A field guide to federated optimization. 2021. http://arxiv.org/abs/2107.06917, December 7, 2022, preprint: not peer reviewed.
46. Chai D, Wang L, Chen K, et al. FedEval: a benchmark system with a comprehensive evaluation model for federated learning. 2020. http://arxiv.org/abs/2011.09655, October 17, 2022, preprint: not peer reviewed.
47. Cho YJ, Jhunjhunwala D, Li T, et al. To federate or not to federate: incentivizing client participation in federated learning. In: *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*. 2022. https://openreview.net/forum?id=pG08eM0CQba
48. Rasmussen A, Ice JA, Li H, et al. Comparison of the American-European Consensus Group Sjögren's syndrome classification criteria to newly proposed American College of Rheumatology criteria in a large, carefully characterized SICCA cohort. *Ann Rheum Dis*. 2014;73(1):31-38.
49. Petersmann A, Müller-Wieland D, Müller UA, et al. Definition, classification and diagnosis of diabetes mellitus. *Exp Clin Endocrinol Diabetes*. 2019;127(S 01):S1-S7.
50. Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med*. 2020;3:109.
51. Weber GM, Hong C, Xia Z, et al.; Consortium for Clinical Characterization of COVID-19 by EHR (4CE). International comparisons of laboratory values from the 4CE collaborative to predict COVID-19 mortality. *NPJ Digit Med*. 2022;5(1):74.
52. Zhang HG, Dagliati A, Shakeri Hossein Abad Z, et al.; Consortium for Clinical Characterization of COVID-19 by EHR (4CE). International electronic health record-derived post-acute sequelae profiles of COVID-19 patients. *NPJ Digit Med*. 2022;5(1):81-11.
53. Klann JG, Weber GM, Estiri H, et al.; Consortium for Clinical Characterization of COVID-19 by EHR (4CE) (CONSORTIA AUTHOR). Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. *J Am Med Inform Assoc*. 2021;28(7):1411-1420.
54. Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: a communication-efficient and privacy-preserving distributed algorithm. *J Am Med Inform Assoc*. 2020;27(3):376-385.
55. Chakraborty B, Murphy S, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat Methods Med Res*. 2010;19(3):317-343.
56. Li T, Sahu AK, Talwalkar A, et al. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag*. 2020;37(3):50-60.
57. Liu M, Zhang Y, Liao KP, et al. Augmented transfer regression learning with semi-non-parametric nuisance models. 2022. http://arxiv.org/abs/2010.02521, July 3, 2023, preprint: not peer reviewed.
58. Gu T, Taylor JMG, Mukherjee B. A synthetic data integration framework to leverage external summary-level information from heterogeneous populations. *Biometrics* 2023. https://doi.org/10.1111/biom.13852
59. Liu M, Xia Y, Cho K, et al. Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *J Mach Learn Res*. 2021;22:126:5607-126:5632.
60. Li T, Sahu AK, Zaheer M, et al. Federated optimization in heterogeneous networks. 2020. http://arxiv.org/abs/1812.06127, September 20, 2022, preprint: not peer reviewed.
61. Zhao Y, Li M, Lai L, et al. Federated learning with non-IID data. 2018. https://doi.org/10.48550/arXiv.1806.00582, preprint: not peer reviewed.
62. Dennis DK, Li T, Smith V. Heterogeneity for the win: one-shot federated clustering. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR; 2021:2611-2620. Accessed December 9, 2022. https://proceedings.mlr.press/v139/dennis21a.html
63. Bernstein J, Zhao J, Azizzadenesheli K, et al. signSGD with majority vote is communication efficient and fault tolerant. 2019. doi:10.48550/arXiv.1810.05291
64. Wei K, Li J, Ding M, et al. Federated learning with differential privacy: algorithms and performance analysis. *IEEE Trans Inf Forensics Secur*. 2020;15:3454-3469. doi:10.1109/TIFS.2020.2988575
65. Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. Dallas, TX: ACM; 2017:603-618. doi:10.1145/3133956.3134012
66. Wang Z, Song M, Zhang Z, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning. In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*; 2019:2512-2520. doi:10.1109/INFOCOM.2019.8737416
67. Dwork C. Differential privacy. In: Bugliesi M, Preneel B, Sassone V, eds. *Automata, Languages and Programming*. Springer; 2006:1-12. https://doi.org/10.1007/11787006_1
68. Ouadrhiri AE, Abdelhadi A. Differential privacy for deep and federated learning: a survey. *IEEE Access*. 2022;10:22359-22380.
69. Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling tabular data using conditional GAN. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2019. Accessed November 14, 2022. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html
70. Choudhury O, Gkoulalas-Divanis A, Salonidis T, et al. Differential privacy-enabled federated learning for sensitive health data. 2020. https://doi.org/10.48550/arXiv.1910.02578, preprint: not peer reviewed.
71. Zhou X, Liu X, Lan G, et al. Federated conditional generative adversarial nets imputation method for air quality missing data. *Knowl-Based Syst*. 2021;228:107261.