HHS Public Access

Author manuscript

Adv Kidney Dis Health. Author manuscript; available in PMC 2023 May 24.

Published in final edited form as:

Adv Kidney Dis Health. 2023 January ; 30(1): 4–16. doi:10.1053/j.akdh.2022.11.007.

Federated Learning in Health care Using Structured Medical Data

Wonsuk Oh,

Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY

Girish N. Nadkarni

Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; Division of Data-Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

The success of machine learning-based studies is largely subjected to accessing a large amount of data. However, accessing such data is typically not feasible within a single health system/hospital. Although multicenter studies are the most effective way to access a vast amount of data, sharing data outside the institutes involves legal, business, and technical challenges. Federated learning (FL) is a newly proposed machine learning framework for multicenter studies, tackling data-sharing issues across participant institutes. The promise of FL is simple. FL facilitates multicenter studies without losing data access control and allows the construction of a global model by aggregating local models trained from participant institutes. This article reviewed recently published studies that utilized FL in clinical studies with structured medical data. In addition, challenges and open questions in FL in clinical studies with structured medical data were discussed.

Recent advanced machine learning (ML) methods, with widespread adaptation of electronic health records (EHR), facilitate opening a new era of personalized medicine.[1,2] Personalized medicine is emerging clinical practices and studies using a broad range of patient characteristics to identify subtypes of disease at a granular level and guide the best treatment strategies. Personalized medicine can improve clinical outcomes as most clinically challenging diseases are not single entries and are treated differently. For instance, acute kidney injury (AKI)[3,4] consists of 3 subtypes: prerenal AKI, intrinsic acute kidney diseases, and acute postrenal obstructive nephropathy. Each subtype has its own risk of adverse outcomes and requires different treatment strategies. The diagnosis, prognosis, and treatment

---

Address correspondence to Girish N. Nadkarni, MD, MPH, Barbara T. Murphy Division of Nephrology, Director, The Charles Bronfman Institute of Personalized Medicine, System Chief, Division of Data-Driven and Digital Medicine, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1243, New York, NY 10029 girish.nadkarni@mountsinai.org.

Financial Disclosure: The authors declare no competing nonfinancial interests, but the following competing financial interests: G.N.N. is a founder of Renalytix, Pensieve, and Verici; provides consultancy services to AstraZeneca, Reata, Renalytix, Siemens Healthineer, and Variant Bio; and serves as a scientific advisory board member for Renalytix and Pensieve. He also has equity in Renalytix, Pensieve, and Verici. W.O. has no relevant financial disclosures.

recommendation are challenging problems with conventional linear models. Many ML methods[1] naturally deal with these subtypes by considering latent interactions among features and demonstrating a superior numerical performance compared to linear models in various clinical studies, including chronic kidney disease[5-7] and acute kidney failure.[8-11]

The size and diversity of data largely determine the success of ML-driven studies.[12] Data from a single institute may not be sufficient to apply ML methods. This is because ML methods can address subtypes only if the data contain subtypes with enough observations. However, each subtype will not likely secure sufficient observations from a single institute due to demographic, socioeconomic, and geographic biases. In addition, ML methods are susceptible to overfitting issues. The only way to overcome this challenge without losing granularity is to access more extensive and representative data.

Multicenter studies are a promising solution to address these issues. Multicenter studies typically pool the data from multiple institutes, potentially across different regions. These data may be less biased and more representative of the population and eventually a substantial number of observations. However, sharing medical data outside each institute poses legal, business, and technical challenges. First, medical data contain sensitive private health information. Therefore, sharing this privacy data is regulated by the Health Insurance Portability and Accountability Act[13] from the United States or the General Data Protection Regulation[14] from the European Union. Second, medical data are of significant business value, so health care providers are likely reluctant to share data outside the institute. Finally, patients' privacy is still endangered even with the cutting-edge privacy mechanisms, including deidentification and randomization.

Federated learning (FL)[15] is a recently proposed framework to address potential data leakage issues in multicenter studies, and it has become a predominant topic among clinical domains. Unlike the existing multicenter studies, FL does not require centralized data warehouse facilities. Instead, FL allows each institute to have its own data governance policy and controls to access data. Each institute has its own computing and data warehouse facilities and only shares ML models across the institutes to construct global models. Because only models are shared across the institutes, FL can relieve the substantial burden of privacy and security issues. Despite this promise, only 1 preprint[10] applied FL in kidney disease studies with structured medical data.

This review will formally overview FL and its clinical research applications. The rest of the article is organized as follows. In section 2, we introduce prior FL reviews and the scope of our review article, and we briefly describe the FL concept in section 3. Section 4 comprehensively reviews FL applications in clinical studies with structured medical data. We summarize potential challenges and opportunities for further kidney disease applications in section 5. Finally, section 6 concludes our article.

RELATED WORKS AND SCOPE OF THIS REVIEW

Several review studies for FL in clinical studies have been published to date. These efforts include a general overview of FL on clinical studies[16-19] and its application, including

COVID-19-associated AKI[20] and medical imaging.[21-23] Additionally, a few studies have addressed potential privacy and security issues[16,24] and parameter-sharing mechanisms for achieving trust, transparency, and equity in clinical studies.[25] However, to our knowledge, no review has explicitly focused on FL applications in clinical studies that utilize structured medical data, for example, EHR data.

Here, we focus on cutting-edge FL applications utilizing structured medical data in clinical studies. In this regard, we identified the following 3 research questions for this review: (1) What are the health conditions being studied with FL? (2) What are the data used for FL applications in clinical studies? And (3) what are the ML methods developed or used for FL applications in clinical studies utilizing structured medical data?

We initially identified 168 articles that contain FL and 1 of clinical, medical, or health care in either title or text on PubMed. We excluded review articles (n = 37), manuscripts written in languages other than English (n = 1), preprints (n = 4), conference abstracts (n = 1), applications other than structured medical data or clinical/medical studies (99), and applications other than FL (n = 3). Twenty-three articles were selected for our review. Table 1 summarizes recent clinical studies that utilized FL in structured medical data.

FEDERATED LEARNING

FL is a distributed ML framework. FL consists of multiple institutes, each with computing power and data. It is specially designed to build global ML models by aggregating local ML models trained by local computing machines with data without sharing data outside the institutes. Figure 1 illustrates the difference between local ML, centralized ML, distributed ML, and FL.

The naïve way to achieve this FL is the federated averaging (FedAvg).[15] FL involves the following steps: (1) Participant institutes initially construct local models with their computing power using data; (2) participant institutes transfer local models to a predefined centralized server; (3) the centralized server builds a global model by aggregating parameters of local models; (4) the centralized server distributes the global model to participant institutes; (5) the participant institutes update and train local models based on the global model; and (6) FL iterates 2-6 until the global model reaches convergent or maximum iteration rounds. However, FedAvg is not without limitation and, thus, is prone to data issues (eg, skewed distribution)[48] and privacy and security issues (eg, deidentification)[16,24]; more details can be found in the articles by Aouedi and colleagues, Ali and colleagues, Zhu and Jin, and Eichner and colleagues.[16,24,48,49]

APPLICATIONS

Health Conditions Among Federated Learning Applications

COVID-19.—COVID-19 is still an ongoing global public health crisis. COVID-19 was first discovered in December 2019 in Wuhan, China,[50] and soon after, the outbreak spreads to the entire world. Over 97 million patients have been infected with COVID-19, and 1 million have died due to COVID-19 complications by 2022 in the United States.[51] Notably,

COVID-19 is the most studied disease in a short period of time in history. Nation-wide and world-wide collaborative studies have been performed, including the National COVID Cohort Collaborative[52] and Consortium for Clinical Characterization of COVID-19 by EHR,[53] and FL researchers pay great attention on it.

Vaid and colleagues[31] were the first researchers who applied FL to structured medical data. The authors constructed predictive models for 7-day in-hospital mortality using lasso regression[54] and multilayer perceptron[55,56] on 4029 patients admitted to 5 hospitals within the Mount Sinai Health System in New York City. The authors conducted rigorous evaluations of 2 models in local, centralized, and FL settings. They demonstrated that both lasso regression and multilayer perceptron improve the model's performance substantially on FL compared to locally trained models. Moreover, they proved that FL models show a minimal loss in performance than centralized models.

Dayan and colleagues[38] aimed to construct predictive models for 24-hour and 72-hour oxygen treatment using the Deep & Cross Network[57] on 16,148 patients' medical history and chest x-ray imaging data from 20 health systems. The authors evaluated their method on 3 independent cohorts, showing that FL models can achieve substantially higher performance than ML models trained explicitly on data available in 3 independent cohorts.

Other Health Conditions.—In-hospital mortality risk models[27,28,30,33,37,40] are widely studied among the rest of the conditions, yet we want to point out that 5 of 6 studies were carried out based on Medical Information Mart for Intensive Care (MIMIC), the de facto data set for ML studies. We will discuss details in Section 4.2. The rest of the conditions are as follows: lung cancer or chronic obstructive pulmonary disease risk model[32]; 5-year diabetes risk model[33]; heart disease diagnostic model[33]; fetal acidosis at birth risk model; Parkinson's disease stage prognostic model[41]; weight gain trajectory among pregnant patients[39]; 30-day readmission risk among patients undergoing a colectomy surgery[34]; liver, kidney, breast, stomach, and uterine diagnostic model[43]; liver disease diagnostic model[58]; hepatocellular carcinoma diagnostic model[58]; diabetic complications (retinopathy, nephropathy, and neuropathy) risk model[44]; 5-year mortality among patients with larynx cancer[47]; diabetes risk model[35,45]; 1-year hospitalization risk among patients with heart disease[26]; decompensation risk among critically ill patients[40]; length-of-stay prediction model[40]; psychological disorders risk model[45]; and ischemic heart disease risk model.[45]

Data Source for Federated Learning

FL has applied structured medical data from various sources, including EHRs,[3,26-28,30-33,36-38,40,43,44,46,47] insurance claims,[36,45] and clinical study registries.[29,34,35,37,39,41,42,58] Here, we specifically focus on EHR data.

Electronic Health Records.—The recent widespread adaptation of EHRs allows researchers to access rich and longitudinal clinical data inexpensively. Most FL studies utilize EHR data from multiple institutes.[3,30-33,38,43,44,46,47] In contrast, the data used by the rest of the studies from a single institute[26-28,36,37,40] are mostly from MIMIC, which we will explain in detail in the following subsection.

Medical Information Mart for Intensive Care.—The MIMIC[59] comprises data of over 40,000 critically ill patients admitted to the Beth Israel Deaconess Medical Center. MIMIC has served as the de facto data set for ML-based studies. MIMIC provides refined and reusable preprocessing codes for analysis. Therefore, ML researchers can focus on evaluating methods using real-world data for proof-of-concept purposes. However, MIMIC shows less suitable characteristics for FL studies. MIMIC comes from a single institute. Accordingly, FL studies rely on partitioning data into multiple subsets to replicate multiple institutes, yet they may not have sufficient variance on subsets within the study data. We can confirm this problem in the literature where, among 3 studies utilizing in-hospital mortality as a study outcome, 2 studies[28,37] have failed to prove the performance improvement of FL compared to local models. The remaining 1 study[40] does not evaluate local and FL models.

ML Methods for Federated Learning

FL is a framework for facilitating ML methods in a distributed manner without sharing data. We explored which methods had been applied in clinical studies with structured medical data.

Artificial Neural Network.—Artificial neural network (ANN)[55,60] is a computational model that consists of input, hidden, and output layers with connected nodes. FedAvg[15] is the first and most commonly used ANN method for FL-based studies. The objective of FedAvg is to minimize the global loss function by reducing a set of local loss functions. FedAvg has a simple structure but has demonstrated acceptable efficient and robust performance. Accordingly, 5[28,31,37,38,44] out of 12 ANN FL studies applied FedAvg.

Cox Model.—Cox model[61] is a semiparametric regression model exploring the relationship between features and the time a specified event takes place. Developing Cox FL models is a challenging problem because Cox models rely on an underlying baseline hazard function that comes from data and is not parameterized.

Wang and colleagues[46] proposed an innovative method for dealing with this problem. The proposed method, called SurvMaximin, applies a transfer learning approach to build Cox models on target data by robustly combining multiple Cox regression models trained in participant institutes. The authors applied the proposed method to 3-day, 7-day, and 14-day mortality in COVID-19 patients. The models quantify the coefficients robustly and show high concordance compared with local models. However, SurvMaximin does not give coefficients for global models.

Hansen and colleagues[47] tackled the baseline hazard issue. The proposed method iteratively trains stratified Cox models across the participant institutes until the models reach convergence or maximum iteration. Because stratified Cox regression utilizes strata-specific baseline hazards, only parameters for local Cox models are required to share across the participant institutes. The authors evaluated mortality in the Larynx cancer study and demonstrated that the proposed method gives coefficients for global models and shows performance comparable to that of the centralized model. However, the authors did not evaluate local and FL models.

Tensor Factorization.—Tensor factorization (TF),[62] a class of methods for decomposing a tensor into a subset of smaller tensors, is a widely used feature-extraction method that could potentially be used for subphenotype discovery. Recently, Ma and colleagues[36] generalized TF for the FL setting and demonstrated outperformed results on MIMIC and Centers for Medicare & Medicaid Services data than existing methods, including FlexiFact,[63] TRIP,[64] and DPFact.[65] However, TF has less attention in clinical research as TF itself is not reproducible and needs to be addressed in further studies.

Methods for Overcoming Duplicated Patient Data.—Some FL methods address the issue of duplicated patient data across participant institutes. Patient data can spread across multiple institutes. For example, the same patient information can be found with both health care providers and health insurance provider. Another example might be patients referring or moving to different health care providers for various reasons. However, merging or excluding those patients is a challenging problem without sharing patient data, yet these patients are likely to be overrepresenting (due to the duplications) and underrepresenting (due to sparse data) in FL models. Bey and colleagues[27] proposed a novel fold-stratified cross-validation approach by considering similar patients in the same fold. This ensures that duplicated patient data appear on the same test or validation set. However, this approach is prone to an issue of patients transferring to different health care providers for advanced care, which may involve extensive lab tests. Liu and colleagues[45] address underrepresenting issues in the FL setting. The authors construct generative adversarial network models for diagnosis, labs, and medications and use generative adversarial network to infer data. The conventional FL models can be applied on top of the inferred data. The authors applied the proposed method to risk models for diabetes, psychological disorders, and ischemic heart disease and proved superior performance compared to FL-only models.

CHALLENGES AND OPEN QUESTION

Because FL enables building clinical models from multiple institutes without sharing patients' data, it can play a significant role in developing novel, data-driven personalized kidney disease models. Although FL can provide promising clinical studies, many new challenges (including a new perspective on old problems) emerge. This section discusses some important challenges in applying FL in clinical studies with structured medical data.

Common Data Model

Different health systems have different EHR systems with different database schemas and coding systems. In reality, this can even be observed frequently in hospitals within the same health systems. While most ML methods and statistical models require clean study data, the heterogeneous EHR systems can impose a significant huddle on data extraction and preprocessing for facilitating FL.

Recently, common data models (CDMs)[66] have been actively studied for clinical research purposes. CDM aims to organize data into a standard structure to facilitate sharing data and information with different applications and data sources so that it supports collaborative research nation-wise and globally. The Observational Medical Outcomes Partnership (OMOP) CDM[67] is a great example of the use of CDM in clinical studies. Many academic

health systems currently implement an OMOP CDM data warehouse for research. This OMOP CDM data warehouse facilitates collaborative works, including the National COVID Cohort Collaborative.[52] Further studies are desirable for CDM on various medical domains to enable multiomics research.

Missing Data

Missing data are a common but substantial issue in data analysis. Missing data can be broadly categorized as missing completely at random, missing at random, and missing not at random (MNAR).[68] Many studies have addressed this issue by study design and missing data imputation methods, including multivariate imputation by chained equations[69] and block-wise data imputation.[70]

MNAR can be induced as we integrate data from multiple institutes even though we apply the same study design for addressing MNAR. Specifically, MNAR can occur due to the varying health care policies and infrastructures across the institutes. For instance, HbA1c and fasting plasma glucose are the 2 measurements for diabetes. While HbA1c is widely used for diabetes diagnosis and treatment, some health systems use fasting plasma glucose for treatment administration precisely. Therefore, we may be unable to avoid MNAR when we conduct multicenter studies. The first-hand approach to this is discretization based on prior clinical knowledge. However, discretization can cause a loss of substantial information. On the other hand, missing data imputation might be an alternative option. However, using missing data imputation on data having MNAR may cause bias.[71] Developing novel methods that can address the MNAR issue systemically is worth exploring.

Phenotypes

The diagnostic code is essential data for understanding patients' characteristics. However, diagnostic codes vary over time[72] and around the world[73] and may be incomplete[74] due to the nature of the clinical data. Several approaches are used to overcome these challenges. First, current clinical guidelines are used to attribute the missing diagnostic code. This approach protects us from incomplete diagnostic codes for patients' medical records due to changes in diagnostic code definitions over time. Second, phenotypic algorithms[75,76] (eg, eMerge algorithms[77]) are used to identify patients with specific health conditions based on different conditions, lab test results, or drugs.

Interpretability

Many FL studies have demonstrated that FL models can show equal or reasonably comparable performance to ML models from centralized data. However, only some studies have pointed out the interpretability issue called explainable artificial intelligence.[78] In fact, explainable artificial intelligence has become a key topic in clinical ML. Clinicians are liable for their clinical practices, which makes clinicians reluctant to adopt black box models. Many ML studies address this interpretability issue through visualization approaches, for example, SHapley Additive exPlanations (SHAP diagram),[79] yet most of those require data access.

CONCLUSION

As ML becomes a predominant method for clinical studies, accessing more extensive and diverse data becomes a significant bottleneck. Multicenter studies are a promising solution to address these issues, yet sharing patient data across the institute is a challenging problem due to ethical, regulatory, business, and technical challenges. FL is a recently developed ML approach that allows the construction of global models without sharing data across participant institutes. Here, this article reviewed 23 recently published cutting-edge articles for FL in clinical studies with structured medical data. In addition, we have addressed various challenges and open questions that arise in FL in clinical studies with structured medical data. Future FL systems for clinical studies need to consider CDMs and issues arising in multicenter studies to facilitate data-driven personalized medicine studies.

Support:

This work was supported by the National Institutes of Health (NIH) grants R01DK108803, U01HG007278, U01HG009610, and U01DK116100 awarded to G.N.N. and T32DK007757 awarded to W.O. The content is solely the responsibility of the authors and does not necessarily represent the views of the NIH.

REFERENCES

1. Yadav P, Steinbach MS, Kumar V, Simon GJ. Mining electronic health records (EHRs): a survey. ACM Comput Surv. 2018;50(6):1–40.
2. Ben-Israel D, Jacobs WB, Casha S, et al. The impact of machine learning on patient care: a systematic review. Artif Intell Med. 2020;103:101785. 10.1016/j.artmed.2019.101785 [PubMed: 32143792]
3. Xu J, Zhang Y, Yu H, et al. High performance of privacy-preserving acute myocardial infarction auxiliary diagnosis based on federated learning: a multicenter retrospective study. Ann Transl Med. 2022;10(18):1006. [PubMed: 36267731]
4. Mercado MG, Smith DK, Guard EL. Acute kidney injury: diagnosis and management. Am Fam Physician. 2019;100(11):687–694. [PubMed: 31790176]
5. Chan L, Nadkarni GN, Fleming F, et al. Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease. Diabetologia. 2021;64(7):1504–1515. [PubMed: 33797560]
6. Bai Q, Su C, Tang W, Li Y. Machine learning to predict end stage kidney disease in chronic kidney disease. Sci Rep. 2022;12(1):8377. [PubMed: 35589908]
7. Zhang K, Liu X, Xu J, et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. Nat Biomed Eng. 2021;5(6):533–545. [PubMed: 34131321]
8. Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. NPJ Digit Med. 2020;3(1):139. [PubMed: 33134556]
9. Bhatraju PK, Mukherjee P, Robinson-Cohen C, et al. Acute kidney injury subphenotypes based on creatinine trajectory identifies patients at increased risk of death. Crit Care. 2016;20(1):372. [PubMed: 27852290]
10. Jaladanki SK, Vaid A, Sawant AS, et al. Development of a federated learning approach to predict acute kidney injury in adult hospitalized patients with COVID-19 in New York City. medRxiv Prepr Serv Heal Sci. 2021.
11. Chaudhary K, Vaid A, Duffy Á, et al. Utilization of Deep learning for subphenotype identification in sepsis-associated acute kidney injury. Clin J Am Soc Nephrol. 2020;15(11):1557–1565. [PubMed: 33033164]

12. Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K. Efficient machine learning for big data: a review. Big Data Res. 2015;2(3):87–93.
13. Centers for Disease Control and Prevention. Health insurance portability and accountability Act of 1996 (HIPAA). https://www.cdc.gov/phlp/publications/topic/hipaa.html (Accessed 15 October 2022).
14. European Commission. What is GDPR, the EU's new data protection law? Accessed October 28, 2022. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations_en
15. 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. 2017;54:1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html. (Accessed 28 October 2022).
16. Aouedi O, Sacco A, Piamrat K, Marchetto G. Handling privacy-sensitive medical data with federated learning: challenges and future directions. IEEE J Biomed Heal Inform. 2022:1–14. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9804708. (Accessed 28 October 2022).
17. Rahman A, Hossain MS, Muhammad G, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. Cluster Comput. 2022. https://link.springer.com/article/10.1007/s10586-022-03658-4. (Accessed 28 October 2022).
18. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res. 2021;5(1):1–19. [PubMed: 33204939]
19. 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. [PubMed: 33847779]
20. Gulamali FF, Nadkarni GN. Federated learning in risk prediction: a primer and application to COVID-19-associated acute kidney injury. Nephron. 2022:1–5. 10.1159/000525645.
21. Darzidehkalani E, Ghasemi-rad M, van Ooijen PMA. Federated learning in medical imaging: Part I: toward multicentral health care ecosystems. J Am Coll Radiol. 2022;19(8):969–974. [PubMed: 35483439]
22. Darzidehkalani E, Ghasemi-rad M, van Ooijen PMA. Federated learning in medical imaging: Part II: methods, challenges, and considerations. J Am Coll Radiol. 2022;19(8):975–982. [PubMed: 35483437]
23. Ng D, Lan X, Yao MM-S, Chan WP, Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quant Imaging Med Surg. 2021;11(2):852–857. [PubMed: 33532283]
24. Ali M, Naeem F, Tariq M, Kaddoum G. Federated learning for privacy preservation in smart healthcare systems: a comprehensive survey. IEEE J Biomed Heal Inform. 2022:1–14. 10.1109/JBHI.2022.3181823.
25. Li D, Luo Z, Cao B. Blockchain-based federated learning methodologies in smart environments. Cluster Comput. 2022;25(4):2585–2599. [PubMed: 34744493]
26. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform. 2018;112:59–67. 10.1016/j.ijmedinf.2018.01.007. [PubMed: 29500022]
27. Bey R, Goussault R, Grolleau F, Benchoufi M, Porcher R. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. J Am Med Inform Assoc. 2020;27(8):1244–1251. [PubMed: 32620945]
28. Lee GH, Shin S-Y. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res. 2020;22(10):e20891. [PubMed: 33104011]
29. Mangold P, Filiot A, Moussa M, et al. A Decentralized framework for Biostatistics and privacy Concerns. 2020.
30. Shao R, He H, Chen Z, Liu H, Liu D. Stochastic Channel-based federated learning with neural network pruning for medical data privacy preservation: model development and experimental validation. JMIR Form Res. 2020;4(12):e17265. [PubMed: 33350391]
31. 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. [PubMed: 33400679]

32. Rajendran S, Obeid JS, Binol H, et al. Cloud-based federated learning implementation across medical Centers. JCO Clin Cancer Inform. 2021;(5):1–11. [PubMed: 33411624]
33. Cui J, Zhu H, Deng H, Chen Z, Liu D. FeARH: Federated machine learning with anonymous random hybridization on electronic medical records. J Biomed Inform. 2021;117:103735. https://www.jmir.org/2020/10/e20891/. (Accessed 15 October 2022). [PubMed: 33711540]
34. Huang Y, Jiang X, Gabriel RA, Ohno-Machado L. Calibrating predictive model estimates in a distributed network of patient data. J Biomed Inform. 2021;117:103758. [PubMed: 33811986]
35. Chen H, Mohapatra S, Michalopoulos G, He X, McKillop I. Federated Deep learning architecture for personalized healthcare. Stud Health Technol Inform. 2021;281:193–197. 10.3233/SHTI210147. [PubMed: 34042732]
36. Ma J, Zhang Q, Lou J, Xiong L, Ho JC. Communication efficient federated generalized tensor factorization for collaborative health data analytics. In: Proceedings of the Web Conference 2021. ACM; 2021:171–182. 10.1145/3442381.3449832
37. Sadilek A, Liu L, Nguyen D, et al. Privacy-first health research with federated learning. NPJ Digit Med. 2021;4(1):132. [PubMed: 34493770]
38. 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. [PubMed: 34526699]
39. Puri C, Dolui K, Kooijman G, et al. Gestational weight gain prediction using privacy preserving federated learning. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE; 2021:2170–2174. 10.1109/EMBC46164.2021.9630505
40. Thakur A, Sharma P, Clifton DA. Dynamic neural graphs based federated reptile for semi-supervised multi-tasking in healthcare applications. IEEE J Biomed Heal Inform. 2022;26(4):1761–1772.
41. Melissourgos D, Gao H, Ma C, Chen S, Wu SS. On outsourcing artificial neural network learning of privacy-sensitive medical data to the cloud. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). IEEE; 2021:381–385. 10.1109/ictai52525.2021.00062
42. Hauschild A-C, Lemanczyk M, Matschinske J, et al. Federated Random Forests can improve local performance of predictive models for various health care applications. Bioinformatics. 2022. 10.1093/bioinformatics/btac065.
43. Ma Z, Zhang M, Liu J, et al. An assisted diagnosis model for cancer patients based on federated learning. Front Oncol. 2022;12. 10.3389/fonc.2022.860532.
44. Islam H, Mosa A, Famia. A federated mining approach on predicting diabetes-related complications: demonstration using real-world clinical data. AMIA Annu Symp Proc. 2022;2021:556–564. [PubMed: 35308968]
45. Liu D, Fox K, Weber G, Miller T. Confederated learning in healthcare: training machine learning models using disconnected data separated by individual, data type and identity for Large-Scale health system Intelligence. J Biomed Inform. 2022;134:104151. 10.1016/j.jbi.2022.104151. [PubMed: 35872264]
46. Wang X, Zhang HG, Xiong X, et al. SurvMaximin: robust federated approach to transporting survival risk prediction models. J Biomed Inform. 2022;134:104176. 10.1016/j.jbi.2022.104176. [PubMed: 36007785]
47. Rønn Hansen C, Price G, Field M, et al. Larynx cancer survival model developed through open-source federated learning. Radiother Oncol. 2022;173:319–326. 10.1016/j.radonc.2022.06.009. [PubMed: 35738481]
48. Zhu H, Jin Y. Multi-objective evolutionary federated learning. IEEE Trans Neural Networks Learn Syst. 2020;31(4):1310–1322.
49. Eichner H, Koren T, Mcmahan B, Srebro N, Talwar K. Semi-cyclic stochastic gradient descent. Proc 36th Int Conf Mach Learn. 2019;97:1764–1773. https://proceedings.mlr.press/v97/eichner19a.html. (Accessed 15 October 2022).
50. Lake MA. What we know so far: COVID-19 current clinical knowledge and research. Clin Med. 2020;20(2):124–127.
51. Centers for Disease Control and Prevention. COVID data tracker. Accessed October 28, 2022. https://covid.cdc.gov/covid-data-tracker/

52. Haendel MA, Chute CG, Bennett TD, et al. The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28(3):427–443. [PubMed: 32805036]
53. 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(1):109. 10.1038/s41746-020-00308-0. [PubMed: 32864472]
54. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58(1):267–288.
55. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–507. [PubMed: 16873662]
56. LeCun Y, Bengio Y, Hinton GE. Deep learning. Nature. 2015;521(7553):436–444. [PubMed: 26017442]
57. Wang R, Fu B, Fu G, Wang M. Deep & cross network for ad click predictions. In: Proceedings of the ADKDD'17. ACM; 2017:1–7. https://arxiv.org/abs/1708.05123. (Accessed 15 October 2022).
58. Hauschild A-C, Lemanczyk M, Matschinske J, et al. Federated Random Forests can improve local performance of predictive models for various healthcare applications. In: Wren J, ed. Bioinformatics. 2022;38(8):2278–2286. [PubMed: 35139148]
59. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, Physio-Toolkit, and PhysioNet. Circulation. 2000;101(23):E215–E220. [PubMed: 10851218]
60. Hinton GE. Learning multiple layers of representation. Trends Cogn Sci. 2007;11(10):428–434. [PubMed: 17921042]
61. Cox DR. Regression models and Life-tables. J R Stat Soc Ser B. 1972;34(2):187–220.
62. Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev. 2009;51(3):455–500.
63. Beutel A, Talukdar PP, Kumar A, Faloutsos C, Papalexakis EE, Xing EP. Scalable flexible factorization of coupled tensors on Hadoop. In: Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2014:109–117. https://epubs.siam.org/doi/abs/10.1137/1.9781611973440.13. (Accessed 28 October 2022).
64. Kim Y, Sun J, Yu H, Jiang X. Federated tensor factorization for computational phenotyping. KDD. 2017;2017:887–895. 10.1145/3097983.3098118. [PubMed: 29071165]
65. Ma J, Zhang Q, Lou J, Ho JC, Xiong L, Jiang X. Privacy-Preserving tensor factorization for collaborative health data analysis. Proc ACM Int Conf Inf Knowl Manag. 2019;2019:1291–1300. 10.1145/3357384.3357878. [PubMed: 31897355]
66. Weeks J, Pardee R. Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in U.S. Health care research. EGEMS (Wash DC). 2019;7(1):4. [PubMed: 30937326]
67. Hripcsak G, Duke JD, Shah NH, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574–578. [PubMed: 26262116]
68. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20(1):199. [PubMed: 32711455]
69. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014;179(6):764–774. [PubMed: 24589914]
70. Xiang S, Yuan L, Fan W, Wang Y, Thompson PM, Ye J. Bi-level multisource learning for heterogeneous block-wise missing data. Neuroimage. 2014;102:192–206. 10.1016/j.neuroimage.2013.08.015. [PubMed: 23988272]
71. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40–49. [PubMed: 21499542]
72. Schwartz LM, Woloshin S. Changing disease definitions: implications for disease prevalence. Analysis of the third national health and Nutrition examination survey, 1988-1994. Eff Clin Pract. 1999;2(2):76–85. [PubMed: 10538480]
73. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. World Health Organ Tech Rep Ser. 2000;894:i–xii.1-253. [PubMed: 11234459]

74. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for Handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013;1(3):1035. [PubMed: 25848578]

75. Chiu PH, Hripcsak G. EHR-based phenotyping: bulk learning and evaluation. J Biomed Inform. 2017;70:35–51. 10.1016/j.jbi.2017.04.009. [PubMed: 28410982]

76. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc. 2013;20(e2):e206–e211. [PubMed: 24302669]

77. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20(e1):e147–e154. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence-draft. (Accessed 28, November 2022). [PubMed: 23531748]

78. Phillips PJ, Hahn CA, Fontana PC, et al. Four Principles of Explainable Artificial Intelligence. 2021. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence-draft. Accessed November 28, 2022.

79. Lundberg SM, Lee S-I. A Unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30 (NIPS 2017). https://papers.nips.cc/paper/2017/hash/8a20a8621978632d7a6c43dfd28b67767-Abstract.html; 2017.

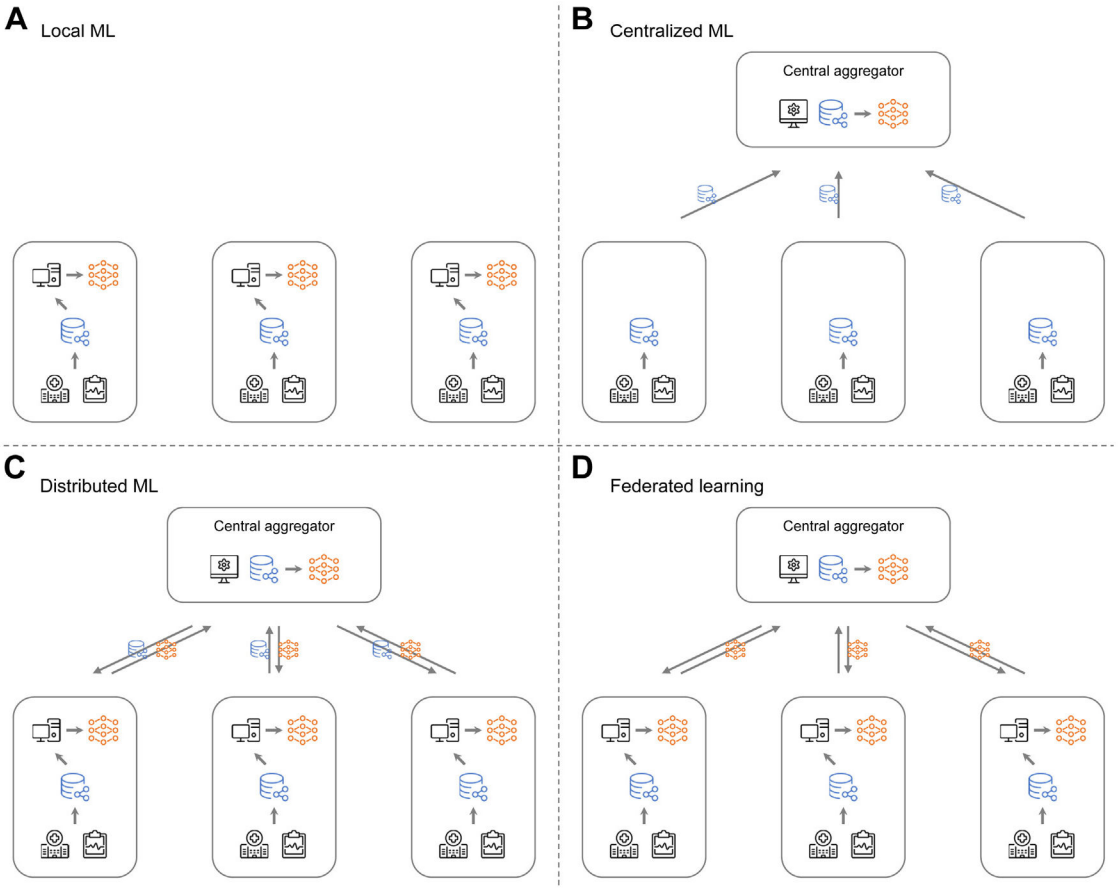

Figure 1.

(A) Local ML: Each institute constructs local models with its own computing power and data. (B) Centralized ML: Central aggregator polls patient data from participant institutes. A global model is constructed on the central aggregator's computing power. (C) Distributed ML: The central aggregator distributes computing tasks and data to the participant institutes to construct a global model. (D) Federated learning: Only local and global models can be moved across the participant institutes and central aggregator. A global model is constructed by aggregating local models from participant institutes.

Table 1.

Studies Using Federated Learning With Structured Medical Data

Title	Study Type	Topic	Outcome	Model	Evaluation Metrics	# Of Institutes	# Of Patients	Comment
Federated learning of predictive models from federated Electronic Health Records[26]	Prognostic	Heart diseases	Hospitalizations	SVM	AUROC	1	45,579	Pros: The proposed SVM shows higher performance with lower iteration, requiring low computing power and communication.
Fold-stratified cross-validation for unbiased and privacy-preserving federated learning[27]	Prognostic	General	Mortality	XGBoost	Accuracy; AUROC	1	Not specified	Pros: This might be the first study addressing the issue of the same patients appearing at multiple institutes. Others: The same patient may appear in multiple institutes, but his/her condition may differ across the institutes. In this case, is the question of what the authors want to solve still valid?
Federated Learning on Clinical Benchmark Data: Performance Assessment[28]	Prognostic	ICU	48-h in-hospital mortality	LSTM	F1 score; precision, recall; AUROC	1	21,139	Others: No difference in raw and imbalanced settings; Evaluation on data from same institute.
A Decentralized Framework for Biostatistics and Privacy Concerns[29]	Prognostic	Obstetric; HF	Fetal acidosis at birth; HF	LR	OR	1; 4	775; 740	Cons: Not sufficient evaluation to show the value of FL; some coefficients may show huge variances, but these could be due to the data itself; it would be great to use AUROC as well.
Stochastic Channel-Based Federated Learning With Neural Network Pruning for Medical Data Privacy Preservation: Model Development and Experimental Validation[30]	Prognostic	Not specified	Mortality	NN	AUROC; AUPRC	Not specified	30,760	Pros: Reduce the number of channels so that it can minimize the data and increase robustness. Others: Not sure the higher performance is an effect of channel selection as a normalized term
Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach[31]	Prognostic	COVID-19	7-d in-hospital mortality	LR; NN	AUROC; binary cross-entropy loss	5	4029	Pros: In depth comparison of 2 models on FL setting. Cons: interestingly, some evaluation even show that NN FL shows higher performance, but not sufficient information about the underlying reason. Others: Both LR and ANN improve performance substantially on FL.
Cloud-Based Federated Learning Implementation Across Medical Centers[32]	Prognostic	Not specified	Either lung cancer or COPD	LR; NN	F1 score; precision, recall; accuracy; AUROC	2	Not specified	Pros: Comparison across various ML models. Others: ANN improved on FL, while LR improved minimally.
FeARH: Federated machine learning with anonymous	Prognostic	ICU	Mortality	NN	AUROC; AUPRC	208	30,760	Pros: FL without centralized server. Cons: Apparently, pooled data and conducted

Title	Study Type	Topic	Outcome	Model	Evaluation Metrics	# Of Institutes	# Of Patients	Comment
random hybridization on electronic medical records[33]								traditional train/test/validation data (does not fully take advantage of data from 208 institutions.); No comparison between FL and local models.
Calibrating predictive model estimates in a distributed network of patient data[34]	Prognostic	Colectomy surgery	30 d readmission	SIR	AUROC; H-L C statistics; H-L H statistics; ECE; MCE	Not specified; 1	17,184	Pros: The distributed calibration measurement is the most significant part. Cons: The source of incorrect calibration needs to be clearly specified.
Federated Deep Learning Architecture for Personalized Healthcare[35]	Prognostic	Diabetes	Diabetes	GLMM	Accuracy	Not specified	768	Pros: One of a few papers dealing with the performance of vertical and horizontal FL. Cons: Results consist of many missing and incomplete evaluations.
Communication Efficient Federated Generalized Tensor Factorization for Collaborative Health Data Analytics[36]	Feature extraction	ICU		TF	Logit loss	Not specified; 1	91,999; 34,272	Cons: Tensor factorization itself is not reproducible and, thus, may get less attention from domain experts.
Privacy-first health research with federated learning[37]	Diagnostic; prognostic	HF; diabetes; in-hospital mortality; COVID-19; avian influenza; bacteremia; azithromycin; tuberculosis	Survival; 5-y diabetes risk; inpatient mortality; COVID-19 infection; fatality; relapse; adverse events; extrapulmonary TB	LR; NN	AUROC; OR; coefficient	Not specified	299; 768; 53,423; 9275; 294; 159; 1712; 3342	Cons: An approach for splitting data is not specified. Others: Evaluation on 8 different benchmark data sets.
Federated learning for predicting clinical outcomes in patients with COVID-19 Federated learning for predicting clinical outcomes in patients with COVID-19[38]	Prognostic	COVID-19	24-h oxygen treatment; 72-h oxygen treatment	NN	AUROC	20 + 3	16,148	Pros: 3 independent external validation cohorts; use of standard codes. Cons: No comparison between ML models from centralized data; only ANN (Deep & Cross network) was applied in this study. Others: FL models show better results than local models.
Gestational weight gain prediction using privacy preserving federated learning[39]	Prognostic	Obstetric	Weight gain	PR	MAE	Not specified	80	Pros: The authors generalized polynomial regression for FL. Cons: Potentially prone at biased data.
Dynamic Neural Graphs Based Federated Reptile for Semi-Supervised Multi-Tasking in Healthcare Applications[40]	Diagnostic; prognostic	ICU	In-hospital mortality; decompensation; phenotype classification; length-of-stay prediction	NN	AUROC	1	21,139	
On Outsourcing Artificial Neural Network Learning	Diagnostic	Parkinson's outcome	HYstage	NN	Accuracy	Not specified	4900	Pros: Training models outside of the institute with minimum privacy leakage

Title	Study Type	Topic	Outcome	Model	Evaluation Metrics	# Of Institutes	# Of Patients	Comment
of Privacy-Sensitive Medical Data to the Cloud[41]								through matrix mask. Cons: Potentially prone at demask approach.
Federated Random Forests can improve local performance of predictive models for various health care applications[42]	Diagnostic	Liver disease	Liver disease; Hepatocellular carcinoma	RF	AUROC	Not specified; 1	583; 685	Pros: Various experiments, including smaller patients per site, different numbers of patients per site, and imbalanced patients. Others: 3/5 evaluations were done with nonclinical data.
An Assisted Diagnosis Model for Cancer Patients Based on Federated Learning[43]	Diagnostic	Cancer	Liver; kidney; breast; stomach; uterine	NN	Not specified	3	Not specified	Cons: Insufficient information regarding the study design.
A Federated Mining Approach on Predicting Diabetes-Related Complications: Demonstration Using Real-World Clinical Data[44]	Prognostic	Diabetes	Diabetic retinopathy; nephropathy; neuropathy	LR; NN	F1 Score; Precision, Recall	22; 31; 31	10,599; 17,455; 23,682	Cons: Does not fully specify how to construct study data with over or undersamples; Need confidence interval to better model assessment.
Confederated learning in healthcare: Training machine learning models using disconnected data separated by individual, data type and identity for Large-Scale health system Intelligence[45]	Prognostic	General	Diabetes; psychological disorders; ischemic heart disease	NN, GAN	AUROC; AUPRC	Not specified	82,143	Pros: Another study addressing the issue of the same patients appearing at multiple institutes. Cons: The authors disclose they only show the numerical performance of FL models from diagnosis-only data, but still, it would be better to show FL models from full data.
	Prognostic	COVID-19		Cox		17	83,178	
SurvMaximin: Robust federated approach to transporting survival risk prediction models[46]			3-d, 7-d, 14-d mortality		AR (1); compound symmetry; AUROC			Pros: SurvMaximin can give coefficients of the Cox model for the target population without accessing data. Cons: The method itself does not give a global model from the entire data. Others: Cox is the first-hand tool for clinical studies, yet, it is hard to build over FL since the baseline hazard is semiparametric.
Larynx cancer survival model developed through open-source federated learning[47]	Prognostic	Larynx cancer	Mortality	Cox	C-statistic	3	786	Pros: Interesting to see the cancer application from clinician friendly methods; Provide coefficients for the global model. Cons: Limited comparison between FL and local models.
High performance of privacy-preserving acute myocardial infarction auxiliary diagnosis based on federated learning: a multicenter retrospective study[3]	Diagnostic	Not specified	Acute myocardial infarction	NN	Precision; sensitivity; accuracy	3	3614	Cons: Confidence interval is not provided. Others: Cohorts are appeared to be from RCT, but no detailed information about it. Otherwise, there is no way the cohort has a perfect balance regarding outcomes.

Abbreviations: SVM, Support Vector Machine; AUROC, Area Under the Receiver Operating Curve; ICU, Intensive Care Unit; LSTM, Long Short Term Memory Network; HF, Heart Failure; LR, Logistic Regression; OR, Odds Ratio; NN, Neural Network; AUPRC, Area Under the Precision Recall Curve; ANN, Artificial Neural Network; COPD, Chronic Obstructive Pulmonary Disease; ML, Machine

Adv Kidney Dis Health. Author manuscript; available in PMC 2023 May 24.

Learning; SIR, susceptible-infected-recovered; ECE, expected calibration error; MCE, maximum calibration error; GLMM, generalized linear model; TF, Tensor Flow; TB, Tuberculosis; FL, Federated Learning; PR, Prediction; MAE, Mean Absolute Error; RF, Random forest; GAN, Generative Adversarial Networks; RCT, Randomized Controlled Trial.