International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

The role of federated learning in improving predictive analytics in public health data systems without compromising privacy

David Kajovo *

Independent Researcher.

International Journal of Science and Research Archive, 2025, 16(02), 046-051

Publication history: Received on 23 June 2025; revised on 29 July 2025; accepted on 01 August 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.2.2278

Abstract

The growing use of data-driven methods in the public health sector has augmented the need for quality and heterogeneous data sets to drive predictive analytics. However, there are significant privacy implications and regulatory and technological risks related to using centralized data systems, especially when handling sensitive data such as health records. This conceptual review addresses how the Federated Learning (FL) concept can revolutionize and make privacy-preserving predictive analytics realizable in the context of public health data systems. FL uses a decentralized model training method with the possibility of many institutions coming together and creating strong analytical models without exchanging raw information. The paper synthesises the main theoretical premises, e.g., privacy-by-design principles, regulatory frameworks, e.g., GDPR and HIPAA, and the main mechanics of FL, e.g., cross-silo and cross-device architecture. It considers the nature of how its privacy-utility trade-off is addressed. It reviews a conceptual framework that could depict the concept of FL integration into government public health systems. It is also possible to mention among the possibilities of FL that it might be used with other promising technologies, such as blockchain and AI, to enhance both outbreak prediction and responsiveness of the health system. Upon declaring numerous opportunities mushrooming, the challenges raised include data heterogeneity, communication overheads, and infrastructural constraints, particularly in low- and middle-income countries that are subjected to serious analysis. The paper ends with recommendations on the policy and system thresholds and directions of future empirical research and conceptual development on implementing FL in the public health sector. Presenting FL as a technology and ethical innovation, this review outlines what changes it can bring to how the public health systems use data and why, resulting in trust, transparency, and regulatory compliance.

Keywords: Federated Learning; Public Health; Predictive Analytics; Data Privacy; Decentralised Systems; Blockchain; Artificial Intelligence

1. Introduction

The field of predictive analytics has been revolutionising public health with the ability to perform data-driven actions (awareness of an upcoming outbreak, dedicated resource allocation, personal care, etc.) (Raghupathi & Raghupathi, 2020; Wiens et al., 2019). These functions are dependent on what is known as big, heterogenous data comprised of information collected in platforms such as electronic health records (EHRs), laboratory systems, and mobile health applications (Dash et al., 2019; Sim, 2019). Nevertheless, the typical centralised data system, although suitable to provide powerful analytics, is dangerous to privacy (Price & Cohen, 2019). Concentration of sensitive health information in central stores raises the chances of a potential breach and unauthorized access, more so in low-governance data environments (Zhang et al., 2019). It discredits trust in the authorities and discourages data sharing, which results in reducing the efficiency of public health campaigns (Vayena et al., 2018). A considerable option is Federated Learning (FL) (Yang et al., 2019). It is possible to train predictive models in FL in a decentralised manner using data holders

* Corresponding author: David Kajovo

without transfer (Li et al., 2020). The data at each node are locally stored, and each node just broadcasts updates to their model, which means that each node is kept anonymous and at the same time allows them to form collective intelligence (Kairouz et al., 2021). This strategy corresponds with ethical information control and new governing guidelines. This paper presents a conceptual review of FL's role in enhancing predictive analytics in public health without compromising privacy. It explores theoretical foundations, examines challenges and opportunities, and proposes a model for integrating FL into health data systems. Guided by the question, *How can Federated Learning improve public health predictive analytics without compromising individual privacy?,* the review synthesises literature from 2015 to 2025, with a focus on foundational and emerging frameworks in decentralised health analytics.

2. Theoretical Foundations / Key Concepts

To integrate Federated Learning (FL) in effective ways into the systems used in predictive analytics in the field of public health, one must be familiar with the essential concepts, theoretical basics, mainly concerning the decentralised machine learning, data privacy, and predictive health modelling notions (Li et al., 2020; Rieke et al., 2020). Collectively, they give the framework to the design of ethically correct, data-driven interventions without trading privacy (Kaissis et al., 2021). Federated Learning (FL) refers to a distributed machine learning framework whereby several institutions or devices learn jointly to train models in a manner that they do not exchange raw data (Kone250c relativny et al., 2016). Firstly, FL has become a rather important approach in sensitive fields like healthcare, which was first introduced by Google in 2016 (Yang et al., 2019). It enables a distributed global model to be trained over multiple data sources, each typically in a different location (e.g., at different hospitals or mobile health apps), and each using locally generated data as input, with only the model itself being distributed (e.g., model updates, e.g., weights are propagated) (McMahan et al., 2017). These updates are collected on at central server, enhancing the worldwide model without ever looking at the initial data (Bonawitz et al., 2019).

A FL architecture usually contains three essential parts: local data holders (i.e., clients), the central coordinating server, and an iterative communication loop through which the model can be refined (Lim et al., 2020). In every round, the model is sent out to clients, and they update it locally, whereupon the updates are sent back to a central position to be aggregated, sometimes with Federated Averaging (FedAvg) (McMahan et al., 2017). There are two prominent flavors of FL; cross-silo FL (a limited number of trusted entities hold abundant data records - suited to hospitals and health departments, Xu et al., 2021) and cross-device FL (many small-data clients, individuals and wearables, or community surveillance - Kairouz et al., 2021). Predictive analytics in population health employs tools and methods of statistics and machine learning to predict disease trends and risks and make decisions (Wiens et al., 2019). The uses are epidemic, stratification of patients according to risk, and allocation of resources optimally (Rajkomar et al., 2019). Nonetheless, the privacy and compliance concerns associated with centralising the data used to develop such models are highly raised, especially when electronic health records (EHRs), laboratory findings, demographic and social determinants of health (SDOH) are used (Price & Cohen, 2019). With increased regulation regarding privacy, FL promises to be an interesting decentralised solution that does not have to sacrifice analytic ability at the cost of protection (Sheller et al., 2020).

Privacy-by-Design (PbD) frameworks incorporate into FL implementation, considering that the FL deployment is based on privacy-distinctive frameworks, which encourage the implementation of privacy capabilities in the system during initial stages of its development (Cavoukian, 2018). FL directly matches the PbD ideas by reducing information exposure and mitigating secure and efficient methods such as differential privacy, secure multiparty computation (SMPC), and authorised access (Geyer et al., 2017). Moreover, legal standards such as the General Data Protection Regulation (GDPR) or the Health Insurance Portability and Accountability Act (HIPAA) contribute even more to the validity of FL (Mantelero, 2018). GDPR is focused on data minimisation and consent, which promotes the decentralised and transparent character of FL (Voigt & Von dem Bussche, 2017). HIPAA requires the protection of the health information, and FL is of interest to U.S. institutions that are wading in predictive analytics under regulatory guidelines (Cohen & Mello, 2018). Data governance is also subject to biomedical ethics, including autonomy, beneficence, non-maleficence and justice (Beauchamp & Childress, 2019). FL also enables justice because it can lead to the representative training of models in a diverse range of people, which potentially leads to the decrease of the algorithmic bias and enhances health equity (Holzinger et al., 2022). To conclude, Federated Learning should be viewed as a paradigm shift with regard to using predictive analytics in the context of public health (Li et al., 2022). It reduces the distance between innovation and regulation since it maintains privacy, increases security, and ethical use of data (Kaissis et al., 2021). With health systems growingly dependent on predictive modelling, conceptual solutions to integrating FL present a solution to an increasingly responsible course of action (Rieke et al., 2020).

3. Conceptual Integration / Analysis

The overriding problem when working with public health data systems has been a privacy-utility trade-off that can now be solved by implementing Federated Learning (FL) into data systems (Li et al., 2022). Such traditional centralised methods also necessitate collecting sensitive health data that leads to a significant privacy concern and regulatory issues (Kaissis et al., 2021). Conversely, FL decentralises the model training, which allows institutions to have a complete control of their data and also contributes to shared, high-performance models (Rieke et al., 2020). This maintains confidentiality and facilitates the adherence to the regulations on privacy, like GDPR and HIPAA (Voigt & Von dem Bussche, 2017; Cohen & Mello, 2018). The use of FL can be depicted as a conceptual three-layer structure: (1) the data proprietorship layer, where information stays in the possession of the rightful parties (Sheller et al., 2020); (2) the collaborative model training layer, where learning would be organized with safe protocols (Bonawitz et al., 2019); and (3) the decision-support layer, where the insights could be used to trigger public health initiatives, such as an outbreak alert or vaccine distribution (Wiens et al., 2019). The model corresponds with privacy-by-design and ethics of data governance (Cavoukian, 2018; Beauchamp & Childress, 2019).

Adoption of FL however has its challenges. Heterogeneity in data can compromise model performance, namely, because of the differing formats, standards, and collecting practices (Xu et al., 2021). Bias can be generated by algorithmic differences produced by a difference in data quality at the sites (Holzinger et al., 2022). Besides, the communication overhead, where nodes and servers commonly update each other, may strain the systems, particularly in low-resource situations, like in Nigeria or India, where a poor infrastructure and lack of technical skills are still a major problem (Lim et al., 2020). Whilst these seem to be hurdles, there are great opportunities in FL. Transparency and accountability in a multi-stakeholder area may be enhanced by incorporating FL into blockchain (Kairouz et al., 2021). In the same way, FL + AI/ML help to better detect localized health trends and make real-time forecasts even in such delicate areas as mental health or infectious disease monitoring (Yang et al., 2019). More specifically, FL provides essential support to real-time outbreak forecasting based on distributed data analysis, which is not stored centrally, and thus, it will be highly valuable in responding to global health threats in an ethically and time-sensitive manner (Rajkomar et al., 2019). To conclude, FL has transformative prospects within the premise of building inclusive, safe, and smart systems in the field of public health (Rieke et al., 2020).

4. Discussion

Based on these conceptual connections that are discussed in this review, Federated Learning (FL) is evident as a promising and transformational way to address the conflict between data utility and privacy in public health analytics (Li et al., 2022; Rieke et al., 2020). Decentralised model training incorporation into the predictive model infrastructures enables the use of large-scale heterogeneous datasets by health systems without compromising the freedom and privacy of data custodians (Sheller et al., 2020). The overlap between FL and critical areas of public health policy, i.e., epidemic prediction, equitatable healthcare access, and resource allocation demonstrates the possibility of FL to serve beyond the technological novelty as a structural facilitator of ethical and non-discriminatory health policy (Holzinger et al., 2022; Wiens et al., 2019). Among the main findings of this review, it is possible to highlight the fact that the issue of the privacy-utility trade-off is no longer regarded as an unsolvable dilemma (Kaissis et al., 2021). Federated Learning enables to build powerful predictive models with the help of distributing the analytical efforts on a variety of nodes thus preserving complete control of data to each node (Yang et al., 2019). This setup ensures high institutional confidence and reduces dangers of data hack in centralised infrastructure, rule infringements, and social distrust of the use of data (Cohen & Mello, 2018). Moreover, this architecture matches the ethical principle of autonomy quite, which gives people and institutions more control over their data contributions (Beauchamp & Childress, 2019).

As a policy point of view, the implementation of analytics based on FL should be led by the public health authorities as it is their chance to be responsible leaders (Vayena et al., 2018). To begin with, FL should be duly accepted as a legitimate and safe approach to collaborative data analysis by policy frameworks (Mantelero, 2018). The legal authorities like those who protect data in the country should work out particular instructions related to decentralised learning, e.g. the governance of models, the methods of consent, and audit of update (Voigt & Von dem Bussche, 2017). Funding, piloting and strategic partnerships with academia or private sector stakeholders should also be used as incentives to incentivise the adoption of FL platforms by public health institutions (Xu et al., 2021). In addition, it is necessary to incorporate FL into the practice of public health through changes at several levels (Lim et al., 2020). In terms of their technological infrastructure, the governments should invest in safe cloud platforms, standardised health information exchange protocols, and stable connectivity networks, particularly, beyond the underserved areas (Bonawitz et al., 2019). It will be critical to construct a robust fabric of decentralised computing, which will allow the institutions in charge of

maintaining the health of the population in various geographic locations to seamlessly engage in federated networks (Kairouz et al., 2021).

Besides hardware and network investments, it is essential to enlarge capacity and train the workers (Rajkomar et al., 2019). The technical competencies that data scientists, epidemiologists and health informaticians need to tackle are to design, deploy, and assess the FL systems (Geyer et al., 2017). This necessitates the inclusion of decentralised analytics in academic syllabuses, professional training institutions as well as talent development programmes in the public sector (Price & Cohen, 2019). Open-source systems and knowledge-sharing communities will be able to serve as a useful mean of rapid adoption and dissemination of skills through collaborative working groups (McMahan et al., 2017). Another factor is the aspect of interoperability (Xu et al., 2021). To operate in scale, the terms of communication between FL participating institutions should be lacking in barriers and there should be consensus over a set of data standards and model formats. It requires reconsidering legacy health information systems which were not created with distributed analytics (Dash et al., 2019). Standard-setting bodies and governments will have to collaborate to facilitate the smooth cross-system work by harmonising the data schema, ontology, and software interface (Konečný et al., 2016).

Ethically, FL represents some principles that are important values in the development of fair and responsible data science (Floridi et al., 2018). Being decentralised, it facilitates the involvement of more institutions as smaller health institutions that would otherwise not be able to provide data centrally are also included in collective intelligence (Rieke et al., 2020). In such a fashion, FL can minimise the concentration of power in data-rich organisation and guarantee that health innovations incorporate a variety of perspectives and experiences (Holzinger et al., 2022). Nevertheless, ethical deployment is also a matter that requires attention (Mehrabi et al., 2021). FL systems should be built in such a way so as to prevent the reproduction of local biases represented in datasets. This would require equity-sensitive learning algorithms, model validation, and input in model interpretation by the stakeholders (Zhao et al., 2017). What is more, transparency and explainability should not be sacrificed at the alter of complexity; the application of FL should be open with clear documentation and communication regarding the process of training, updating of models, and its applications with regards to the making of decisions regarding public health (Wachter et al., 2017).

Recommendations

Future research should focus on practical substantiation of Federated Learning by simulating it on everyday public health conditions, primarily through pilot-based studies in low- and middle-income countries. The implementations can estimate the feasibility in practice, infrastructural requirements, and model performance in various health settings. More conceptual advancement is also required to combine FL with new themes like Social Determinants of Health (SDOH) and his knowledge of machine learning and blockchain-based auditability. The participatory design approach should also be researched to ensure that health workers, the custodians of data, and affected communities are not left out when it comes to the implementation of FL. It would be interesting to have comparative studies on the effectiveness of various FL architectures (e.g., cross-silo vs cross-device) in different epidemiological settings.

5. Conclusion

The conceptual overlap of Federated Learning (FL), data privacy, and predictive analytics and their application in public health systems has been reviewed in this paper. FL can resolve the old privacy-utility trade-off by decentralizing the model's training and keeping data ownership local. It supports the establishment of valid, diverse, and morally valid models that have been able to guide essential intervention of Public health infrastructures in the populace- without divulging confidential information. The analysis revealed that FL is technically feasible and meets ethical, legal, and institutional priorities in health governance. As revolutionary, FL will boost outbreak forecasting, distributing resources, and health equity by relying on collaborative intelligence. The ideas presented below will form the foundation of more harmonized, decentralized health data systems based on privacy and trust.

References

[1] Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). Oxford University Press.

[2] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Van Overveldt, T. (2019). Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 1, 374–388.

[3] Cavoukian, A. (2018). Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario. https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

[4] Cohen, I. G., & Mello, M. M. (2018). HIPAA and protecting health information in the 21st century. JAMA, 320(3), 231–232.

[5] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis, and future prospects. Journal of Big Data, 6(1), 54.

[6] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

[7] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client-level perspective. arXiv preprint arXiv:1712.07557.

[8] Holzinger, A., Dehmer, M., & Jurisica, I. (2022). Knowledge discovery and interactive data mining in bioinformatics—State-of-the-art, future challenges, and research directions. BMC Bioinformatics, 23(Suppl 6), 1–18.

[9] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1–210.

[10] Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., ... & Rueckert, D. (2021). End-to-end privacy-preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence, 3(6), 473–484.

[11] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

[12] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

[13] Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., ... & Feng, A. (2022). Privacy-preserving federated brain tumour segmentation. Machine Learning in Medical Imaging, 12966, 133–143.

[14] Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y. C., Yang, Q., ... & Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(3), 2031–2063.

[15] Mantelero, A. (2018). AI and big data: A blueprint for a human rights, social and ethical impact assessment. Computer Law & Security Review, 34(4), 754–772.

[16] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 1273–1282.

[17] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

[18] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

[19] Raghupathi, W., & Raghupathi, V. (2020). Big data analytics in healthcare: Promise and potential. Health Information Science and Systems, 2(1), 3.

[20] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

[21] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

[22] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 12598.

[23] Sim, I. (2019). Mobile devices and health. New England Journal of Medicine, 381(10), 956–968.

[24] Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689.

[25] Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. Springer.

[26] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99.

[27] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Horvitz, E. (2019). Do no harm: A roadmap for responsible machine learning for healthcare. Nature Medicine, 25(9), 1337–1340.

[28] Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., & Wang, F. (2021). Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 5(1), 1–19.

[29] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19.

[30] Zhang, X., Hailu, B., Tabor, D. C., Gold, R., & Jiang, X. (2019). Role of health information technology in addressing health disparities: Patient, clinician, and system perspectives. Medical Care, 59(Suppl 1), S8–S16.

[31] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.