PharmaACE

PHARMAACE WHITE PAPER

UNVEILING THE FUTURE OF MEDICINE WITH FEDERATED LEARNING

Authored by

Monalisha Mallick

PharmaACE

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF PHARMAACE DOCUMENTS

This publication ("Work") is issued by PharmaACE Standards Association ("PHARMAACE") for informational purposes only. It is expressly **not** a consensus standard, nor does it constitute a binding technical or legal standard. The Work is provided **"AS IS"**, **"WITH ALL FAULTS"**, and **without warranty of any kind**, whether express, implied, or statutory.

PHARMAACE **disclaims all warranties**, including but not limited to:

- **Merchantability**
- **Fitness for a particular purpose**
- **Non-infringement**
- **Accuracy, completeness, reliability, or currency of the Work**

PHARMAACE shall **not be liable** for any direct, indirect, incidental, consequential, special, exemplary, or punitive damages, including but not limited to:

- Loss of data, profits, or business interruption
- Procurement of substitute goods or services
- Claims arising under contract, tort (including negligence), strict liability, or any other legal theory

No professional relationship is created by the use of this Work. Users must exercise independent judgment and consult qualified professionals before relying on any information herein. PHARMAACE does not render engineering, legal, or other professional services through this Work.

PHARMAACE makes **no representation** regarding the existence, validity, or enforceability of any third-party intellectual property rights that may be implicated by the use of this Work. Users are solely responsible for:

- Identifying and obtaining necessary licenses
- Assessing risks of infringement
- Ensuring compliance with applicable laws and regulations

PHARMAACE has **not sought nor received** any commitment from rights holders to license patent rights on a reasonable or non-discriminatory basis.

This Work may contain **opinions, interpretations, or guidance** that reflect the views of individual contributors and not necessarily those of PHARMAACE. PHARMAACE does not endorse or guarantee the accuracy of such views.

By accessing or using this Work, the user **acknowledges and agrees** to the foregoing terms and assumes all risks associated with its use.

TABLE OF CONTENTS

UNVEILING THE FUTURE OF MEDICINE WITH FEDERATED LEARNING

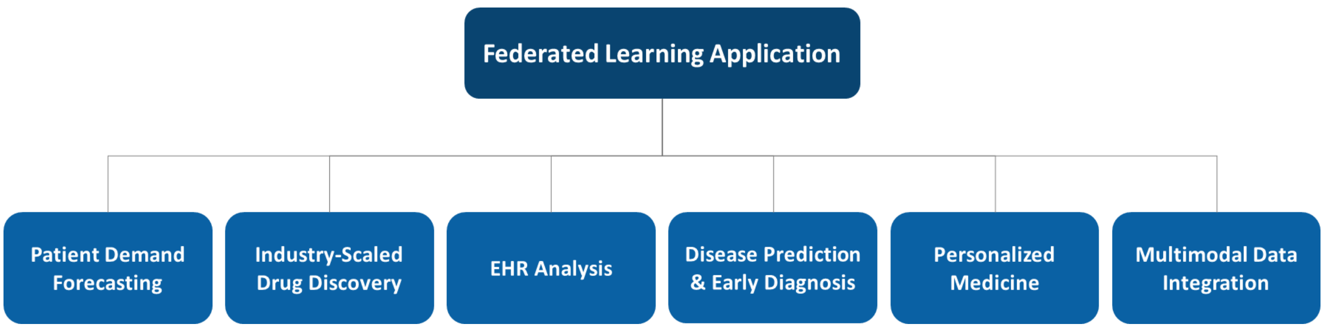
ABSTRACT

The pharmaceutical industry is undergoing a transformative shift driven by the convergence of advanced data management practices and artificial intelligence (AI). With the increasing volume and complexity of pharmaceutical data, efficient, secure, and privacy-preserving data handling is being prioritized. To address these challenges, Federated Learning (FL) has been proposed as a promising decentralized machine learning paradigm. FL is revolutionizing pharmaceutical data management by enabling secure, privacy-preserving collaboration across institutions. Instead of sharing raw data, FL allows models to be trained locally on sensitive datasets such as clinical trials and patient records, and only model updates are aggregated centrally. This approach ensures data privacy, regulatory compliance, and interoperability, making it ideal for the pharmaceuticals industry. The paper clearly outlines the diverse applications of FL in the pharmaceutical sector by first establishing its role in secure data management and then exploring its impact across key domains such as drug discovery, EHR analysis, personalized medicine, and multimodal data integration demonstrating how FL enhances collaboration while preserving data privacy. Despite its advantages, including privacy preservation, scalability, and collaborative potential, FL is not without limitations. Challenges such as system heterogeneity, communication overhead, and model convergence issues continue to be addressed. Nonetheless, as the future of medicine leans increasingly on data-driven innovation, FL stands out as a pivotal technology, fostering collaborative intelligence while safeguarding sensitive medical data.

1. INTRODUCTION: IMPORTANCE OF DATA HANDLING IN PHARMACEUTICALS INDUSTRY

Pharmaceutical data handling is a critical aspect of the healthcare and life sciences industry encompassing the collection, storage, analysis, and protection of sensitive information related to drug development, clinical trials, patient records, and regulatory compliance. With the increasing digitization of pharmaceutical processes, the volume and complexity of data have grown exponentially, necessitating robust data management strategies. Effective data management ensures integrity, accuracy, and confidentiality of information throughout the drug lifecycle. Pharmaceutical data handling is not just a technical necessity but a strategic imperative. It underpins innovation, regulatory compliance, and patient trust, making it essential for companies to invest in modern data management practices and technologies.

Figure 1. Federated Learning in Pharma: Enabling Secure, Collaborative AI Across Key Domains



Artificial Intelligence (AI) is reshaping data management in the pharmaceutical industry by streamlining research workflows and enabling faster, data-driven decisions. One major challenge is inconsistent access to relevant healthcare data, often restricted due to privacy regulations and siloed systems. Federated Learning (FL) is a specialized area within machine learning that adopts a decentralized model training approach[1]. It enables AI systems to analyze data more efficiently while safeguarding the privacy of the original datasets. Google pioneered this concept in 2016, aiming to train models directly on data stored on users' mobile devices without transferring the data itself[2]. In the healthcare sector, data is sourced from various channels including administrative records, digital health platforms, and research databases[3]. FL facilitates the understanding of data structures across these diverse sources without requiring the data to be moved, thereby addressing key challenges related to data accessibility and confidentiality. Thanks to its privacy-preserving capabilities, FL has shown promise in sensitive areas such as transfusion medicine and personalized treatment strategies[4,5]. Additionally, FL can be applied to forecast patient volumes, making it a valuable tool for demand prediction in pharmaceutical operations. Despite its potential, research indicates that only a small fraction of around 5% of studies in healthcare have explored real-world applications of FL, suggesting that technology is still in its early stages[6]. Nevertheless, the growing need for secure and efficient data management in healthcare is gradually driving interest toward integrating FL into pharmaceutical data workflows.

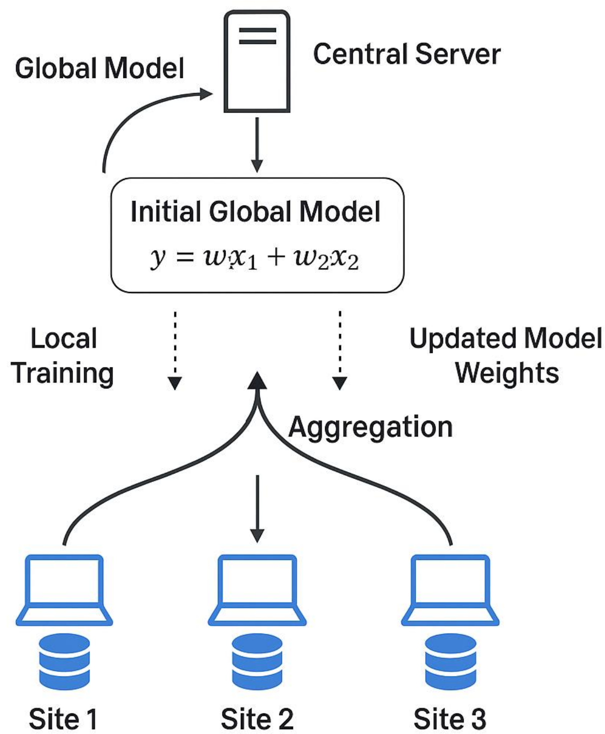
2. HOW FEDERATED LEARNING FUNCTIONS?

In the foundational framework of FL, a central server initiates the process by distributing the initial parameters of a global model to a selected group of participating sites[2]. Each site, using its own local dataset, trains a model that mirrors the architecture of the global one. Instead of sharing raw data, these sites send only the updated model parameters back to the central server. This decentralized approach effectively addresses the limitations of traditional centralized learning, where sensitive data must be transferred to a central repository. Once the server receives updates from all participating sites, it aggregates them to refine the global model. The improved model is then redistributed to a new set of sites for further training. This cycle continues until the global model reaches optimal performance.

FL can be classified based on how data is distributed, primarily into horizontal FL and vertical FL. Horizontal FL, also known as record-level partitioning, involves training models across datasets that share the same features but contain different individual records[7]. For instance, this approach is suitable for building models using patient data from various hospitals, where the variables are consistent, but the patient populations differ. On the other hand, vertical FL referred to as feature-level partitioning combines datasets that contain different types of information about the same individuals[8]. This method is particularly relevant in scenarios where data from multiple sources must be integrated for a comprehensive view of each subject. However, vertical FL is more technically demanding, as it requires careful alignment of data relationships and strategies to prevent issues like overfitting. Due to its simplicity and broader applicability, horizontal FL is more commonly used in health research. It supports the development of models that generalize well across diverse populations, making it especially valuable in studies where data sharing is restricted or fragmented.

Here is a diagrammatic representation of the FL process. It visually illustrates the central server initializing and distributing the model followed by local sites training the model on private data such as age, blood pressure lab results, and the updates (weights/gradients) being sent back to the server where the data is aggregated and refined the global level[9]. Here x_1, x_2, are the input datasets from different sites such as age, blood pressure, lab results and w_1, w_2 are the weight coefficients assigned to each to each data input. w_1x_1, w_2x_2 are the linear combinations used to make predictions.

Figure 2. Federated Learning Process demonstrating a centralized system operated by the central server and gathering inputs from different sites to obtain output on a global basis by performing linear combinations of the input data



3. APPLICATION OF FEDERATED LEARNING IN PHARMACEUTICAL DATA MANAGEMENT

3.1 PATIENT DEMAND FORECASTING

FL has emerged as a transformative approach in patient demand forecasting. Traditional machine learning models often require centralized data aggregation, which poses significant privacy risks and regulatory challenges especially in healthcare, where patient data is highly sensitive. FL addresses these concerns by enabling collaborative model training across multiple institutions without sharing raw data, thereby preserving privacy and complying with data protection regulations. In patient demand forecasting, FL enables pharmacies, hospitals, and clinics to collaboratively contribute to a shared predictive model that estimates future patient volumes, medication needs, and treatment demands[8]. This is accomplished by training local models on each institution's data and then aggregating the model updates centrally. Such a decentralized learning approach ensures that insights are drawn from a wide range of healthcare environments and patient demographics, enhancing the generalizability and accuracy of forecasts. This integrated forecasting helps healthcare providers optimize resource allocation, manage inventory efficiently, and prepare for seasonal or emergent health trends. Importantly, FL also saves time and effort by eliminating the need to access or transfer data from external sources. Each participating entity can contribute to the model using its own data locally, reducing delays and streamlining the forecasting process. This not only accelerates decision-making but also ensures that data remains secure and under the control of its original custodians.

3.2 INDUSTRY-SCALED DRUG DISCOVERY

As the pharmaceutical industry increasingly embraces data-driven innovation, the need for secure, scalable, and collaborative machine learning solutions has become critical. FL offers a promising framework for achieving this by enabling multiple organizations to train shared models without exchanging sensitive data. This approach is particularly valuable in pharmaceuticals industry, where data privacy, intellectual property, and regulatory compliance are paramount. A leading example of this is the MELLODDY project, funded by the European Innovative Medicines Initiative[9]. It developed an industry-scale FL platform involving ten major pharmaceutical companies. Through the use of cryptographic aggregation techniques, MELLODDY enabled secure model training across private datasets, allowing participants to collaboratively improve drug discovery algorithms without compromising data confidentiality. The project demonstrated FL's scalability and effectiveness in real-world scenarios, setting a precedent for future collaborative research in the industry. Another example in drug discovery application is the FLuID framework (Federated Learning Using Information Distillation) which allows companies to share predictive insights using surrogate datasets and knowledge distillation, preserving data privacy and intellectual property[10]. This approach improves predictive performance and fosters collaboration across organizations.

Moreover, FL supports drug-target interaction prediction, especially when data is fragmented across institutions. It enables the creation of robust models for rare diseases or complex conditions, where individual datasets may be insufficient. This collaborative approach enhances model accuracy and broadens the scope of research. FL has been used to benchmark models for predicting drug-target interactions. A study using the GraphDTA model and KIBA dataset showed that FL could improve performance by up to 15% compared to traditional methods, even with non-IID (non-identically distributed) data across clients[11].

3.3 ELECTRONIC HEALTH RECORD ANALYSIS

Pharmaceutical companies increasingly rely on Electronic Health Records (EHRs) to support drug development, clinical trials, and personalized medicine. However, data fragmentation and privacy regulations often limit access to comprehensive datasets. FL offers a scalable solution by enabling secure, decentralized model training across institutions. Instead of transferring sensitive patient data, FL allows encrypted model updates to be shared, preserving privacy and accelerating collaboration. Recent studies have shown FL's effectiveness in handling diverse and complex medical data, improving predictions for rare conditions and enhancing interoperability across systems[12]. Innovations like personalized federated models and differential privacy further strengthen their utility in real-world healthcare applications. For the pharma industry, FL unlocks the potential of multi-source EHR data—streamlining research, improving clinical insights, and supporting regulatory compliance without compromising data ownership.

3.4 DISEASE PREDICTION AND EARLY DIAGNOSIS

Accurate disease prediction and early diagnosis are critical for developing targeted therapies and clinical decision tools. However, accessing diverse patient data across institutions is often restricted by privacy regulations. FL solves this by enabling secure, decentralized model training. Studies have shown FL's effectiveness in predicting conditions like acute pancreatitis, sepsis, and cardiac diseases, achieving high accuracy without centralizing sensitive data[13-15]. Advanced techniques like attention-based models and multimodal frameworks further enhance diagnostic precision. For the pharma industry, FL accelerates biomarker discovery, supports real-time risk stratification, and enables scalable collaboration across healthcare networks, all while maintaining strict data privacy standards.

3.5 PERSONALIZED MEDICINE

Personalized medicine focuses on tailoring treatments to individual patient profiles, including genetics, lifestyle, and clinical history. For pharmaceutical companies, this approach promises more effective therapies and better patient outcomes. However, accessing and integrating diverse, sensitive medical data across institutions remains a major hurdle. FL addresses this by enabling secure, decentralized model training without sharing raw data. Recent innovations like FedMetaMed and GenoMed4All have shown how FL can enhance medication recommendations and accelerate research for rare diseases using distributed datasets[16,17]. Additionally, combining FL with techniques like Differential Privacy has proven effective in real-time diagnostics, such as arrhythmia detection from decentralized ECG data. By supporting privacy-preserving, scalable AI development, FL empowers pharma companies to advance personalized medicine while safeguarding patient data and complying with global regulations.

3.6 MULTIMODAL DATA INTEGRATION

Pharmaceutical companies increasingly rely on diverse data imaging, genomics, EHRs, and wearable sensors to drive precision medicine and clinical research. Integrating these data types securely across institutions is a major challenge. Multimodal Federated Learning (MMFL) offers a decentralized solution, enabling collaborative model training without sharing raw data[18]. Recent frameworks have shown MMFL's effectiveness in combining varied modalities to improve diagnostics, treatment planning, and patient stratification. Innovations like edge computing, secure fusion protocols, and privacy-preserving techniques make MMFL scalable and reliable for real-world applications[19]. For pharmaceutical companies, this translates into quicker insights, enhanced treatment personalization, and adherence to global data protection standards, all while retaining full control over proprietary data.

4. KEY ADVANTAGES AND LLIMITATIONS IN FEDERATED LEARNING

4.1 KEY ADVANTAGES

FL is a decentralized machine learning paradigm that enables multiple clients such as hospitals, mobile devices, or organizations to collaboratively train models without sharing raw data. This approach has gained significant traction in domains like healthcare, finance, and IoT, where data privacy and security are paramount. FL ensures that sensitive data remains local, reducing the risk of privacy breaches and complying with regulations like GDPR and HIPAA. Only model updates such as gradients or weights are shared, not the raw data itself. By training on diverse datasets from multiple sources, FL models often generalize better than those trained on isolated data. This is particularly beneficial in healthcare, where patient demographics and disease patterns vary widely. Since only model parameters are exchanged, FL significantly reduces data transmission requirements, making it suitable for low-bandwidth environments like mobile or remote healthcare settings. FL enables institutions to collaborate on AI development without compromising data ownership. This fosters innovation, especially in fields like drug discovery and medical imaging. FL can be deployed across a wide range of devices, from smartphones to edge computing nodes, allowing for scalable and distributed learning.

4.2 LIMITATIONS

One of the major hurdles in Federated Learning is the presence of non-uniform, non-independent data distributions across participating clients. This can lead to biased model updates and reduced performance if not properly addressed. Although FL reduces raw data transmission, frequent model updates can still incur significant communication costs, especially in large-scale deployments. Aggregating updates from heterogeneous clients requires sophisticated algorithms to ensure fairness and accuracy. Simple averaging may not be effective in all scenarios. FL is susceptible to adversarial attacks such as model poisoning and inference attacks. Ensuring robust security mechanisms like differential privacy and secure multiparty computation is essential. In healthcare, many FL models are still in experimental stages with most FL implementations lacking clinical validation and suffering from methodological flaws, limiting their real-world utility.

5. CONCLUSION

FL is redefining how data-driven innovation unfolds in the pharmaceutical and healthcare sectors. By enabling secure, decentralized model training across institutions, FL addresses one of the most pressing challenges in modern healthcare: the need to collaborate on sensitive data without compromising privacy or regulatory compliance. This paradigm shift allows pharmaceutical companies to unlock the full potential of diverse datasets ranging from EHRs and genomics to medical imaging and wearable sensor data without the need for centralized data aggregation.

FL's applications are vast and impactful. From accelerating drug discovery and improving clinical trial design to enhance disease prediction and enabling personalized medicine, FL empowers organizations to build robust, generalizable models while respecting data ownership. Projects like MELLODDY, GenoMed4All, and frameworks such as FedMetaMed and MMFL demonstrate the scalability and real-world viability of FL across multimodal and multi-institutional settings.

Moreover, the integration of advanced techniques such as differential privacy, edge computing, and personalized federated models further strengthens FL's role in enabling real-time, adaptive, and privacy-preserving analytics. These innovations not only improve diagnostic accuracy and treatment planning but also support strategic decision-making and regulatory alignment.

As the demand for precision medicine and collaborative research grows, FL stands out as a cornerstone technology for the future of healthcare and pharmaceuticals industry. It offers a balanced solution that harmonizes innovation with ethical data stewardship, paving the way for more inclusive, efficient, and secure healthcare systems. Embracing FL is not just a technological upgrade, it is a strategic imperative for organizations aiming to lead in the next era of intelligent, patient-centric care.

6. REFERENCES

The following sources have either been referenced within this paper or may be useful for additional reading:

1. Chaudhary, R.K., Kumar, R. & Saxena, N. A systematic review on federated learning system: a new paradigm to machine learning. *Knowl Inf Syst* **67**, 1811–1914 (2025). https://doi.org/10.1007/s10115-024-02257-6

2. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

3. A. Worku, F.G. Arage, and F.B. Kebede, *Introduction to health care data analytics – an overview*, Discover Health Systems, vol. 4, article 107, 2025.

4. N. Li, A. Lewin, S. Ning, M. Waito, M.P. Zeller, A. Tinmouth, and A.W. Shih, Privacy-preserving federated data access and federated learning: Improved data sharing and AI model development in transfusion medicine, Transfusion, vol. 65, no. 1, pp. 22–28, 2024.

5. A. Chowdhury, H. Kassem, N. Padoy, R. Umeton, and A. Karargyris, A Review of Medical Federated Learning: Applications in Oncology and Cancer Research, in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lecture Notes in Computer Science, vol. 12962, Springer, 2022, pp. 3–24.

6. M. Sandhu, A. Singh, and A. Kaur, A survey on federated learning in medical imaging: From a radiological perspective, Vis. Inf. Commun., vol. 3, no. 2, 2024.

7. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

8. Dayan, I., Roth, H.R., Zhong, A. et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 27, 1735–1743 (2021).

9. Y. Jin, H. Zhu, J. Xu, and Y. Chen, Federated Learning: Fundamentals and Advances, Springer, 2023.

10. Hanser, T., Ahlberg, E., Amberg, A., Anger, L.T., Barber, C., Brennan, R.J., Brigo, A., Delaunois, A., Glowienke, S., Greene, N., Johnston, L., Kuhn, D., Kuhnke, L., Marchaland, J.-F., Muster, W., Plante, J., Rippmann, F., Sabnis, Y., Schmidt, F., van Deursen, R., Werner, S., White, A., Wichard, J. & Yukawa, T. *Data-driven federated learning in drug discovery with knowledge distillation*. **Nat Mach Intell** 7, 451–465 (2025).

11. Mittone, A., et al. A Federated Learning Benchmark for Drug-Target Interaction. arXiv preprint arXiv:2302.07684 (2023).

12. Joshi, H., & Joseph, S. Standardization and Interoperability: Federated Learning's Impact on EHR Systems

and Health Informatics. Advances in Health Information Science and Practice, Vol. 1(1): UBYM3803. 2025.

13. Vieira, P., Maia, E. & Praça, I. Acute Pancreatitis Mortality Prediction with Federated Learning. In: Santos, M.F., Machado, J., Novais, P., Cortez, P. & Moreira, P.M. (eds) Progress in Artificial Intelligence. EPIA 2024.

14. Düsing, C. & Cimiano, P. Improving Early Sepsis Onset Prediction Through Federated Learning. arXiv preprint arXiv:2509.20885 (2025)

15. Ryu, H., Lee, M., Kim, S., Kim, J. H. & Yang, H.-j. Federated Learning for Cardiovascular Disease Prediction: A Comparative Review of Biosignal- and EHR-Based Approaches. Healthcare 13(21), 2811 (2025).

16. Asti, G., Apellaniz, P. A., Carota, L., Casadei, F., Piscia, D., Delleani, M., Isasa, I., Martinez Duarte, D., Rollo, C., Gonzalez Martin, C., Arroyo Galende, B., Saha Cyrille Merleau, N., … Álvarez Garcia, F. (2025). Development, implementation, and validation of an open-source Federated Learning platform to accelerate innovation and boost personalized medicine in rare and ultra-rare haematological diseases: an initiative by GenoMed4All Consortium. medRxiv.

17. Gao, J. & Li, Y. FedMetaMed: Federated Meta-Learning for Personalized Medication in Distributed Healthcare Systems. arXiv preprint arXiv:2412.03851 (2024).

18. Thrasher, J., Devkota, A., Siwakotai, P., Chivukula, R., Poudel, P., Hu, C., Bhattarai, B. & Gyawali, P. Multimodal Federated Learning in Healthcare: a Review. arXiv preprint arXiv:2310.09650 (2023).

19. Aueawatthanaphisut, A. “Secure Multi-Modal Data Fusion in Federated Digital Health Systems via MCP.” *arXiv preprint arXiv:2510.01780* (2025).