Federated Learning: A Solution to Breaking Silos in Healthcare Data

Abdelkader Bouregag, Yuan Lian, Lydia Tanque, Zachary Webel

Abstract:

Demand for organizations to implement advanced artificial intelligence models is increasing at a rapid pace. However, as the architecture of these models advances, the needs for significantly larger datasets also grows. The main problem with healthcare, biotech, etc. That collecting data is costly, and sharing data requires significant efforts to maintain privacy. Therefore we propose the implementation of a federated machine learning platform. This platform allows each organization to train one shared model with its own data. Each data set will be normalized to the agreed-upon data standard. Once the training is done, the organization saves, encrypts, and shares the model parameters (weights, biases) with the cloud platform. This then allows other organizations on the platform to train the shared model without exposing their data. The resulting model is more accurate and generalizable than one only trained on a single dataset. The federated machine learning approach will allow organizations to leverage their data and contribute to advances in artificial intelligence without violating privacy standards.

Problem Statement:

Biomedical-healthcare applications in artificial intelligence aimed at improving clinical outcomes have struggled to keep pace with advancements in other fields of artificial intelligence such as financial, generative and the entertainment markets. The question of why this is happening is fairly straightforward and that answer is data. More specifically, “the rigorous regulation of patient data and the requirements for its protection” (Kaissis). The process of creating an artificially intelligent model that not only is accurate, but also generalizes to unseen data points is more so a problem of data collection and manipulation, rather than complex vector calculus and linear algebra.

Consider the standard data interface in health informatics: the Electronic Health Record. It stores labs, radiology, prescriptions, demographic, and other historical health data, most, if not all of which is considered HIPAA protected. A common practice in the EHR is to implement clinical decision support tools, resources assisting the caregiver in their workflow. Machine learning or deep learning models can become a great resource for such a use case. However, as mentioned above, the need for a large corpus of diverse data is mandatory to generate a model that will aid the caregiver and not harm the patient. This brings along two problems. First, it is very unlikely that an organization will have enough diverse data in their own EHR to create a generalized model sophisticated enough to improve clinical outcomes. Next, if this is the case, the organization cannot just borrow some protected health data from another organization’s EHR implementation. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) “mandates strict rules regarding the storage and exchange of personally identifiable data and data concerning health, requiring authentication, authorization, accountability” (Kaissis). This accounts for all domains in health care data such as imaging, genomics, labs, medications, ect.

This paper discusses the identification and acquisition of a federated machine learning system. The need for an implementation of this system is because of the problems highlighted above.

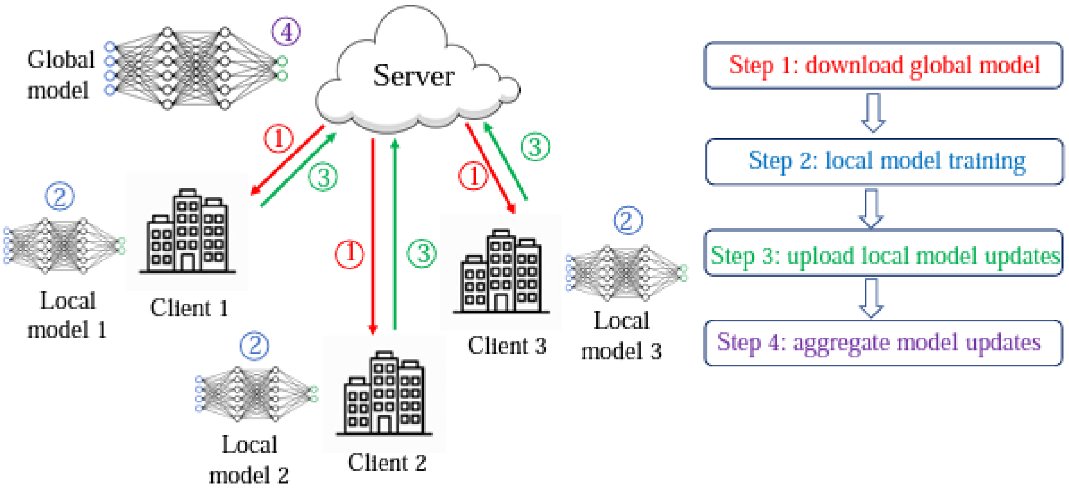
Our goal is to create an artificially intelligent model without violating HIPAA data sharing policies. As a preview, we propose the use of a federated machine learning system in order to collaborate with other healthcare organizations in training one model, without ever exposing our collection of identifiable health data, therefore keeping compliance with data privacy standards.

Solutions development:

In this section, we'll present the two main approaches for federated learning: Cross-device and cross-silos methods

Federated learning (FL) is a distributed machine learning schema where a group of clients trains a global model on their private local data under the coordination of a central server. Based on the clients' infrastructures and training scale, FL can be divided into cross-device and cross-silos. In cross-device FL, clients are usually small distributed entities with limited computational resources and data (e.g. wearables, smartphones, connected devices). However, in cross-silos FL, clients are usually a relatively small number of companies or organizations where each client is expected to have a certain amount of computational power and data and to be ready to participate in the entire training process. (Huang, 2022)

Since we are focusing on the application of federated learning at the institution level, we will only consider cross-silos architecture, the figure below shows the different steps involved in this architecture:


Global model
④
Server
①
③
①
③
①
③
②
Local model 1
Client 1
②
Local model 2
Client 2
Client 3
②
Local model 3
Step 1: download global model
Step 2: local model training
Step 3: upload local model updates
Step 4: aggregate model updates

Cross-silo Federated learning architecture (Huang, 2022)

Moreover, sharing data on a federated learning platform can be done through the following infrastructures types:

Horizontal Federated Learning (HFL):

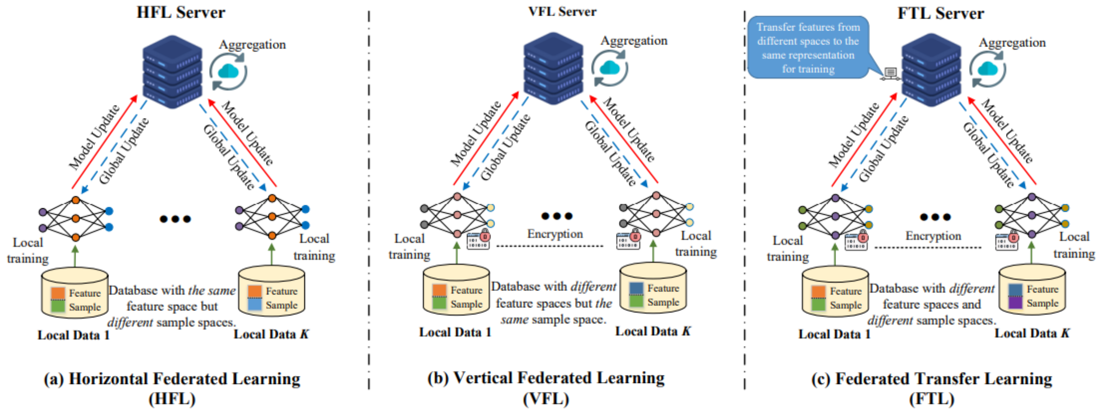
HFL involves training a shared global model using local datasets that have the same feature space but different sample spaces. Local participants can adopt the same AI model for training their datasets, and the server combines the local updates to create a global model without direct access to the data. An example of HFL in smart healthcare is the detection of speech disorders where users speak the same sentence (feature space) with different voices (sample space) on their smartphones

Vertical Federated Learning (VFL):

VFL involves federated training of health datasets that have the same sample space but different feature spaces. This is achieved by integrating solutions based on entity alignment and encryption techniques during local training. An example of VFL can involve the joint training of an AI model by a hospital and an insurance company (different data features) that serve the same patient (same sample space) that incorporate historical medical records at the hospital and healthcare costs at the insurance company to enable intelligent healthcare decision-making.

Federated Transfer Learning (FTL)

FTL can handle datasets with different sample spaces and different feature spaces. Unlike VFL, FTL uses transfer learning to calculate feature values from different feature spaces to the same representation, which is then used to train local datasets. In smart healthcare, FTL can be used to improve disease diagnosis by allowing multiple hospitals in different countries with different patients (sample space) and therapeutic programs (feature space) to collaborate. By enriching the shared AI model output, FTL can improve the accuracy of diagnosis.
The following figure illustrates the comparison between different federated learning types:



(a) Horizontal Federated Learning (HFL)

(b) Vertical Federated Learning (VFL)

(c) Federated Transfer Learning (FTL)

Federated learning types (Nguyen, 2021)

Federated learning platforms:

Federated AI Technology Enabler (FATE):

FATE (Federated AI Technology Enabler) is an open-source federated learning platform that allows multiple organizations to collaborate on data while preserving data security and privacy. It offers a range of federated learning algorithms, including logistic regression, tree-based algorithms, deep learning, and transfer learning, and implements secure computation protocols based on homomorphic encryption and multi-party computation (MPC). FATE is hosted by Linux Foundation, and the Technical Charter outlines the responsibilities and procedures for technical contribution to, and oversight of, the FATE Project. FATE claims to be the world's first industrial-grade federated learning open-source framework.

FedML:

FedML proposes open-source and enterprise-level solutions for federated learning that provides a library of algorithms and system infrastructure for easy experimentation and benchmarking. It supports various types of federated learning, including horizontal, vertical, and federated transfer learning, and provides a modular architecture for researchers to develop new algorithms and protocols. The company is Venture Capital backed and recently raised $6 million to further develop its platform.

PaddleFL:

PaddleFL is a Federated Learning (FL) platform developed by Baidu, which is based on the deep learning platform PaddlePaddle (PArallel Distributed Deep LEarning: Machine Learning Framework) and implemented using C++ and Python programming languages. PaddleFL supports both differential privacy and secure multi-party computation and can work on honest-but-curious parties. It consists of four components at compile time, including FL strategies, user-defined models, and algorithms.

Fedlearner:

Fedlearner is a collaborative machine learning framework that enables joint modeling of data distributed between institutions. It defines the platform as a multi-party collaborative machine learning framework

TensorFlow Federated:

TensorFlow Federated (TFF) is an open-source framework for decentralized machine learning, specifically Federated Learning (FL), which allows a shared global model to be trained across many participating clients keeping their data locally. TFF enables developers to simulate and experiment with FL algorithms and provides interfaces for high-level and low-level implementations of FL and other computations on decentralized data. TFF's interfaces are organized in two main layers: Federated Learning (FL) API and Federated Core (FC) API, allowing developers to apply the included implementations of federated training and evaluation to existing TensorFlow models and to concisely express novel federated algorithms.

Flower:

Flower is a customizable and extendable open-source framework for building federated learning systems. It is designed to be framework-agnostic, which means it can work with any machine-learning framework. Flower is also written with maintainability in mind, making it understandable and easy to contribute to.

FLUTE (Microsoft):

FLUTE (Federated Learning Utilities for Testing and Experimentation) is a high-performance open-source platform for federated learning research and offline simulations at scale. The vision for FLUTE is to support progress in the state-of-the-art in Federated Learning by providing task-agnostic support for a wide variety of scenarios and cutting-edge algorithms with strong experimental results in a user-friendly environment while lowering the FL-related entry barriers to data scientists and researchers. The key differentiator behind FLUTE is the ease of implementing new scenarios for experimentation in core areas of active research—such as optimization, quantization, privacy, and scalability in a robust simulator. (*Project FLUTE*,)

CrypTen:

CrypTen is a PyTorch-based framework for privacy-preserving machine learning, which aims to make secure computing techniques more accessible to machine learning practitioners. It uses Secure Multiparty Computation and presents the protocols via a CrypTensor object that works like a PyTorch Tensor. CrypTen implements a tensor library, allowing practitioners to debug, experiment on and explore ML models. It addresses real-world challenges by not oversimplifying the implementation of secure protocols.

FedTree:

FedTree is a federated learning system that allows for the training of tree-based models in a highly efficient, effective, and secure manner. It supports federated training of gradient boosting decision trees, parallel computing on multi-core CPUs and GPUs, and implements features such as homomorphic encryption, secure aggregation, and differential privacy. It also supports both classification and regression tasks.

System requirements:

After defining the different types and architectures of Federated Learning systems, as well as some of the platforms proposing FL solutions, we will proceed to develop a list of system requirements that will allow us to assess the different Federated learning solutions and evaluate the ones answering to the system's needs based on UniFed framework: a benchmark for federated learning (Lui, n.d.,)

Functionality support:

There are different federated learning frameworks in the market that support different types of functionalities. In our use case, we focus on the implementation of most used machine learning methods in healthcare including shallow learning, deep learning, and imaging data support. We also look into the support of the different types of federated learning, Horizontal only / All-in frameworks

FL framework support:

The system requirements for federated learning framework support includes the ability to securely transfer model weight updates. In addition, the ability to at any point in time change the number of training providers is crucial. Provided the same common data model for inputs and outputs, the scalability of the federated learning platform is defined by the ability to add or remove any training providers.

Model support:

The model support requirements are defined by the federated machine learning platform to have the ability to utilize multiple coding languages and multiple artificial intelligence platforms. First, different types of languages can be used in machine learning such as python, R, c++, ect. The federated learning platform needs the ability to exchange model updates that are interpretable by these different languages. Next, artificial intelligence platforms such as keras, tensorflow, pytorch are all commonly used. Once again, the chosen platform must have the ability to exchange model updates that are returned by the load_weights and save_weights functions across all of the platforms.

Privacy engagement and security concerns:

In federated learning, raw data is kept in the local organization, minimizing the risk of data leakage or unauthorized access. Model updates are shared instead of actual data, which helps protect private patient information.

Aggregation Security: While sharing model parameter updates, secure aggregation methods can be used to ensure that individual client updates remain private. Techniques such as Homomorphic Encryption (Monique,2013) can protect the privacy of the exchanged model parameters during the aggregation process.

Access control: Federated learning systems can be designed to require authentication or access control, ensuring that only authorized organizations can participate in the collaborative model training. This contributes to maintaining trust among data-sharing organizations and prevents non-authorized access leak model parameters and patient information.

Encrypted communication: All data updating and between organizations and the central server can be encrypted to ensure the privacy of the exchanged model updates.

Benchmark comparison:

In the following table, we present the comparison results between the different FL platforms that the benchmark has evaluated using the UniFed framework. We align each platform with the requirements mentioned above.

		FATE	**Fed ML**	**PaddleFL**	**FedLearner**	**TFF**	**Flower**	**FLUTE**	**CrypTen**	**FedTree**
Model support - Horizontal	Regression	Y	Y	Y	Y	Y	Y	Y	N/A	N
	Neural network	Y	Y	Y	Y	Y	Y	Y	N/A	N
	Tree-based models	Y	N	N	N	N	N	N	N/A	Y
Model support - Vertical	Regression	Y	Y	Y	N	N	N	N	Y	N
	Neural network	Y	N	Y	Y	N	N	N	Y	N
	Tree-based models	Y	N	N	Y	N	N	N	N	Y
Deployment Support	Single-host simulation	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Multi-hostdeployment(<16hosts)	Y	Y	Y	Y	N	Y	Y	Y	Y
	Cross-devicedeployment(>100host)	N	Y	Y	Y	N	Y	Y	N/A	Y
	Networking protocols	Customized	MPI	gRPC	gRPC	/	gRPC	MPI	Torch Distributed	gRPC
Privacy protection against the semi-honest server	Does not require a 3rd party aggregator (vertical)	Y	N	Y	Y	N	N	N	Y	Y
	Aggregator does not learn model param(arbiter scenario)	Y	N	Y	N	N	N	N	N/A	Y
	Aggregator does not learn individual model gradient (secagg)	Y	Y	Y	N	Y	N	N	N/A	Y
Privacy protection against the semi-honest peer client	Clients do not learn anything about the model param(vertical)	Y	N	Y	N	N/A	N/A	N/A	Y	Y
	Clients do not learn gradiants about the model param(vertical)	Y	Y	Y	N	N/A	N/A	N/A	Y	Y
Privacy protection in the final model	Support training with central DP (Differentially-Private Stochastic Gradient Descent)	N	N	Y	N	Y	Y	Y	N	Y

Overall concept:

The project revolves around the concept of federated machine learning as a solution to address the challenges of data silos in healthcare. The report highlights the need for larger and more diverse datasets to develop accurate and generalizable AI models in healthcare, including horizontal, vertical, and federated transfer learning, and provides an overview of various federated learning platforms such as FATE, FedML, PaddleFL, TensorFlow Federated. The strict regulations surrounding patient data protection, such as HIPAA, make it difficult to share data across organizations.

Proposed solution is the implementation of a federated machine learning platform that allows organizations to train a shared model using their own data without exposing sensitive information. By encrypting and only disclosing the model parameters, such as weights and biases, this collaborative approach enables organizations to leverage their data collectively and contribute to advancements in artificial intelligence without compromising privacy standards.

Technical Infrastructure:

A successful implementation of federated learning requires an environment with the ability to securely and rapidly transfer a h5 file or equivalent depending on the type of algorithm used. We reference the most crucial aspect of federated machine learning as a “Federated Learning Server”. The purpose of this server is to host the model, send and receive model updates and store historical model logs. The server will never access any data, only model parameters. The server must employ encryption methods to securely transfer the model parameters to and from participating organizations. The server will need to employ communication protocols such as HTTP to transfer data. This server will be hosted in the cloud using a contractual agreed upond cloud provider.

On an individual-organizational level, the technical infrastructure required in terms of federated machine learning is the ability to send and receive model parameters to the federated learning server. This technical infrastructure is defined by the personal machines and coding environments.

Integration and interoperability challenges:

One of the significant challenges of Federated Learning in the healthcare system is the lack of interoperability among different healthcare providers and systems. To address this challenge, standards and protocols must be established to ensure that data can be shared securely and efficiently. According to an article on Medium, “The use of standardized data formats and interfaces can help facilitate interoperability between different systems and devices, making it easier to share data and collaborate on Federated Learning projects” (Kumar, A. 2021)

Another critical aspect of implementing Federated Learning in the healthcare system is integration. Healthcare systems comprise a wide array of different workflows, processes, and data. Healthcare providers often use various systems and sources to provide quality care. This can make it hard to combine resources into a single standardized source. Lack of interoperable architectures in these systems can inhibit and cause barriers to care. One important point to note is that while federated Learning has the potential to improve healthcare outcomes, it is crucial to address the ethical and privacy concerns that come with using sensitive data.

Furthermore, implementing federated learning in the healthcare system is integration. Quality healthcare workflow uses numerous specialized resources to drive precision health for patients. Federated Learning research has drastically increased with efforts to overcome these challenges. The platforms are designed to enable secure optimal data sharing across health systems and data sources. This integration can enable data to be collected from multiple sources and consolidated into a single source model across multiple organizations and improve the knowledge and accuracy of Federated Machine Models by allowing it to learn from multiple sources of diverse data. Other implantation barriers comprise additional workload for healthcare providers.

As avenues to safe data sharing grow, providers must ensure that patient data is anonymized and protected throughout the Federated Learning process. Furthermore, patients should be educated on how their data is being used and given the option to opt-out if they wish. By prioritizing patient privacy and ethical considerations, the healthcare system can ensure that Federated Learning is a responsible and effective tool for improving healthcare outcomes.

Overall, the interoperability and integration changes that are required for Federated Learning are significant and these models in healthcare still face many risks. With that being said, the gain outweighs the risk. In addition, as the development of standards and protocols for data sharing improves and demand for a scaled design and architecture grow, Federated Learning has proved that it can be a valuable tool for the research, development, and innovation of digital health.

Meaningful use:

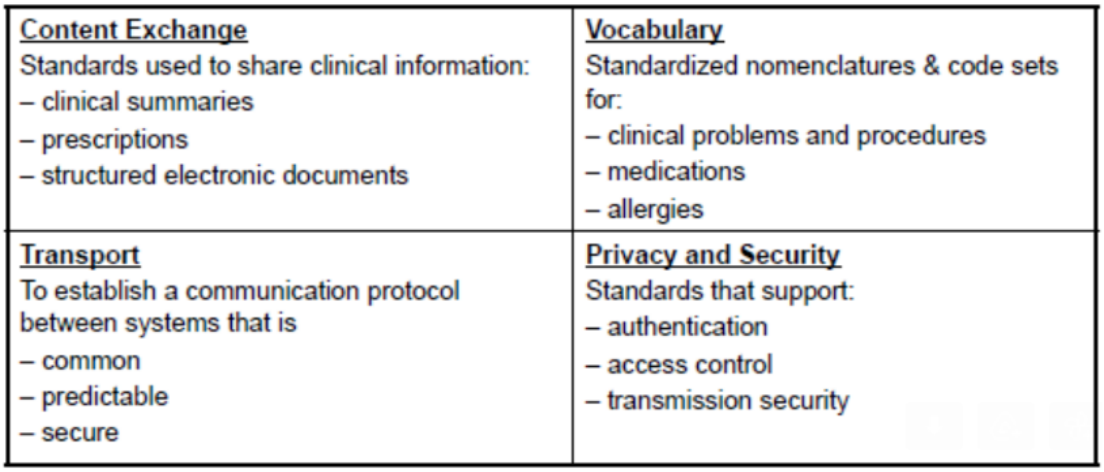
Electronic health records (EHR) adoption and resource utilization focus on Meaningful Use (MU). In addition, the advent of new healthcare applications, technology, tools, and platforms has improved patient care, safety, and delivery outcomes. Thus EHR's MU spans technology adoption, federal & industry standards, delivery, and resource utilization used to deploy technical solutions to hospitals, private clinics, or agencies such as the Centers for Medicare and Medicaid (CMS). This process also involves how emerging technology solutions, such as federated machine learning, improve patient care, data security & privacy, and device and server decentralization on multiple endpoints. MU can be used in healthcare, e.g., EHR, to capture and share patient health data. It is also used to track information on patient care data. In

MU, patient data confidentiality and security must always be maintained for optimal clinical decision-making and healthcare solution delivery.

As a result, healthcare organizations can use EHRs responsibly or meaningfully to improve patient safety and efficiency, deliver quality patient care, and reduce administrative costs, health disparities, and burdens that patients often incur when seeking treatment. This concept affects not only EHR but also the population-public health, clinical care coordination, and other relevant healthcare delivery areas. MU and EHR are integral components for ensuring that healthcare organizations are committed to maintaining patient privacy and data security efficiently. Some of the MU components include but are not limited to, Meaningful Manner, Electronic Exchange, and Clinical Quality Measures. Meaningful Manner uses certified electronic health record technology.

In contrast, Electronic Exchange Information focuses on electronic health exchange. Healthcare organizations rely on electronic health exchange information to improve healthcare quality (Clark et al., 2020). The U.S. Congress defines MU as using certified EHR technology, that is, e-prescribing, to exchange health information to improve patient quality of care. For instance, MU can be used to report identified clinical quality measures to CMS. MU's key components are critical to adopting, utilizing, and delivering optimal healthcare solutions. First, its key objective is to track necessary patient-level clinical information. Second, MU ensures that health providers have visibility into patient populations' health statuses.

Third, when deployed as an optimized/scalable solution, MU renders clinical decision support that health providers design and need to assist them in adhering to and improving evidence-driven best practices. Some of the critical components of MU stem from the electronic healthcare execution and transaction. This process involves formulary drug information, lab results, summarized primary patient data, patient eligibility checking, and how data among crucial stakeholders can be exchanged. The need for reporting evidence-based process metrics, that is, patients with hypertension but blood pressure under control, is vital to continued patient care, safety, and solution delivery. Some MU standards include the listed quadrants/categories – content exchange, vocabulary, transport, and privacy & security.



Content Exchange Standards used to share clinical information: – clinical summaries – prescriptions – structured electronic documents	**Vocabulary** Standardized nomenclatures & code sets for: – clinical problems and procedures – medications – allergies
Transport To establish a communication protocol between systems that is – common – predictable – secure	**Privacy and Security** Standards that support: – authentication – access control – transmission security

Personalized Treatment: Federated learning enables the development of more precise and tailored treatment strategies by fostering cooperation among research institutions and hospitals in training AI models with genomics and clinical data without disclosing the original patient information.

The healthcare and health insurance industries can leverage federated learning to protect sensitive patient health information(PHI) in its original source. Utilization of federated learning in health care could improve diversity in health data through incorporation of diverse data extracted from multitudes of health resources and organizations. As research surrounding Federated Learning increases and improves, there will be a lateral need for new learning methodologies, processes, and model architectures to support larger computing and sample populations being trained on shared models. (Dilmegani, 2023)

Telemedicine: Additionally, federated learning can be applied in telemedicine and remote monitoring, allowing healthcare providers to more effectively analyze and interpret data gathered from wearable devices like smartwatches and other sensors such as blood pressure monitors, enhancing patient care and outcomes.

Drug Discovery: Facilitate collaboration among pharmaceutical companies and research labs to update and leverage distributed data sources and test processes to train more accurate and comprehensive models, leading to more informed decision-making and potentially faster drug development.

Cloud Solutions:

Cloud solutions play a vital role in implementing federated machine learning in healthcare. The uses are hosting the federated learning server in the cloud and utilizing the prebuilt security standards of large cloud providers. Through extensive encryption, access controls, and compliance certifications, it places a strong emphasis on data privacy and security. Strict security procedures are used by AWS, Azure, GCP, and IBM Cloud to guarantee the confidentiality and integrity of sensitive model updates. Healthcare organizations can cooperate on training shared models by integrating federated learning with cloud platforms while maintaining data privacy, abiding by legal obligations, and taking advantage of the scalable computing power and secure infrastructure offered by these cloud solutions.

Conclusion:

In conclusion, the implementation of a federated machine learning platform is crucial in optimizing the return on investment of machine learning clinical decision support tools. Leveraging not only our organizational database, but also industry wide data samples will allow for these advancements to take place. The individual organizations control which data scientists can generate machine learning models and each individual data sample never leaves the organization's IT infrastructure. The federated learning server is hosted in the cloud with a remote backup. The common data model and feature store is agreed upon before the implementation of the learning system. The security standards are also decided at this point. This allows for maximum interoperability and security.

While federated learning offers promising solutions, there are several challenges that need to be addressed for its successful implementation in healthcare data. Such as data heterogeneity and bias, regulatory frameworks and computational constraints. To support efficient collaboration and model training, federated learning platforms must ensure data quality, address the heterogeneity of data structures, and standards, and mitigate biases, which might require crucial adjustment of the system. Along with compliance with informed consent requirements and regulatory frameworks, ensuring alignment while collaboration increases the level of difficulty. In addition, since Participating organizations may have limited processing capabilities, it will constrain their contribution to the training process. Balancing the computational load and optimizing resource usage among different organizations are significant challenges in federated learning.

Addressing these challenges requires a multidisciplinary approach. Collaborative efforts and ongoing research are essential to realize the full potential of federated learning to Breaking Silos in Healthcare Data.

References

1. Huang, C. (2022). Cross-Silo Federated Learning: Challenges and Opportunities. https://arxiv.org/abs/2206.12949
2. Nguyen, D. C. (2021). Federated Learning for Smart Healthcare: A Survey. https://www.semanticscholar.org/paper/Federated-Learning-for-Smart-Healthcare%3A-A-Survey-Nguyen-Pham/d457f7760237df1f147ad0075b34f38711bc74d3
3. Kaissis, G.A., Makowski, M.R., Rückert, D. et al. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2, 305–311 (2020). https://doi.org/10.1038/s42256-020-0186-1
4. Monique Ogburn, Claude Turner, Pushkar Dahal,Homomorphic Encryption,Procedia Computer Science,Volume 20,2013,Pages 502-509,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2013.09.310.
5. Dilmegani, C. (2023, 15 03). What is Federated Learning? Use Cases & Benefits in 2023. Retrieved from AI Multiple: https://research.aimultiple.com/federated-learning/#:~:text=What%20are%20potential%20use%20cases%20and%20examples%20of,Vehicles%20...%204%20Manufacturing%20%E2%80%93%20predictive%20maintenance%20

6. Kumar, A. (2021). Federated Learning in Healthcare: Challenges and Opportunities. Medium. Retrieved from https://towardsdatascience.com/federated-learning-in-healthcare-challenges-and-opportunities-59b1e61b1f8e
7. Clark, P., Myers, L., and Stauffer, C. (August 2020). The program Year 2020 Stage 3 Objectives for Meaningful Use. https://healthcurrent.org/wp-content/uploads/Final-AHCCCS-Program-Year-2020-Stage-3.pdf
8. Tian Li ,Anit Kumar Sahu(2020) Federated Learning: Challenges, Methods, and Future Directions, Institute of Electrical and Electronics Engineers , Signal Processing Magazine,vol 37, 50--60 https://doi.org/10.1109%2Fmsp.2020.2975749