Federated Learning and Analysis for Collaborative Research in Healthcare at a National and International Scale

White Paper by
RLS-Digital Health

RLS-SCIENCES

Fonds de recherche
Nature et
technologies
Québec

CONSORTIUM
SANTÉ
NUMÉRIQUE

Université
de Montréal

CONSORTIUM SANTÉ NUMÉRIQUE

RISC Software GmbH

SDFMU

Uniklinikum Erlangen

USP Universidade de São Paulo

UNIVERSITY OF GEORGIA

Introduction

In the last decade, Big Data and Artificial Intelligence (AI) have generated a wave of transformations that we still have difficulty to comprehensively assess and envision. In healthcare, as in several other domains, the exponential increase of data and information, coupled with an enhanced capacity to process and store them, have radically changed the way we conduct research, development, and decision making (Mehta et al., 2019). This rapid overhaul, however, has been raising questions and concerns among researchers, policy makers, and civil society, as shown by present debate on AI regulation to maximize benefits and mitigate risk in the public interest.

Healthcare is often presented as both a promising field for Big Data applications and a domain that is lagging behind in AI implementation (Shaw et al., 2019). Hence, AI algorithms applied to diagnostics and to treatment prediction and precision, especially in fields such as radiology and cardiology, have undergone considerable developments in the last years (Fornell, 2023). Yet, experts have emphasized the presence of a gap between AI development in medicine and concrete applications at the bedside (McCradden et al., 2019).

The AI chasm may have lessened due to the advancement generated during the COVID-19 pandemic. Nonetheless, there are still many challenges today to the collection, use, and sharing of quality health data to develop useful predictive applications for patients, professionals, and decision makers.

The connection between science and policy is especially important in the field of health and for responsible applications of Big Data and AI. In this frame, the multiregional, international science network, RLS-Sciences, includes a dedicated group of researchers who work on digital health. RLS-Sciences is "bringing the benefits of diverse cultural and scientific perspectives through its multilateral collaboration among scientists, policy makers, and science managers"[1]. RLS-Sciences operates under the umbrella of a political forum, the Regional Leaders Summit (RLS). RLS convenes biennially to exchange amongst seven regional governments: Bavaria (Germany), Georgia (USA), Québec (Canada), São Paulo (Brazil), Shandong (China), Upper Austria (Austria), and Western Cape (South Africa) on the theme "Policy for Generations"[2].

The RLS partner regions first agreed to support scientific cooperation between their regional research institutions in 2012, and chose the field of renewable energy[3]. They invited regional research actors to establish the RLS-Energy Network, where researchers exchange best practices, data, and privileged research results in a trusted scientific network. In 2016 at the 8th RLS in Munich, based on an analysis of the scientific strengths in the regions, the Bavarian RLS members encouraged the establishment of three further groups: RLS-Small Satellites, RLS-Global Aerospace Campus, and the RLS-Expert Dialogue on Digitalization. To support the RLS-Sciences network as a whole, as well as the four groups, regional research administrators created governance and research support infrastructures, including the appointment of dedicated regional coordinators from the governmental, scientific, and science management sectors, as well as lead scientists from each thematic group for each region.

[1] https://oecd-opsi.org/innovations/rls-sciences/
[2] https://www.mdpi.com/2071-1050/13/1/76
[3] https://www.rls-sciences.org/about.html

Together, these representatives actively work to create the conditions for effective international collaboration at the regional level in science. RLS-Sciences members participated at the biennial political summits in 2016, 2018, and 2021, strengthening the science-policy nexus. In 2020, at the invitation of the Government of Upper Austria, representatives of the partner regions and their heads of government, along with the RLS-Sciences network, met for a "Virtual Roundtable on COVID-19". At the roundtable the partner regions exchanged on experiences, best practices, challenges, and exit strategies for combating the COVID-19 pandemic. Ahead of the round table, Bavaria proposed the creation of a fifth group focusing on the field of digital health: RLS-Digital Health. The group undertook a preparatory phase beginning in 2020 and launched as the fifth RLS-Sciences group in 2022.

As the newest RLS-Sciences group, RLS-Digital Health works according to a roadmap which was co-created to a matrix of methods and medical use cases most relevant for the partner regions.

The group is committed to advancing knowledge and promoting translational research in a wide spectrum of topics in the digital health field by bridging the gap between research and application. The group has identified Federated Learning and Federated Analysis as particularly relevant for their future collaborative work, particularly as enabling mechanisms for taking best advantage of the diversity of data within their regions.

At the time of the conception and writing of this White Paper, the following researchers were designated as Lead Scientists in RLS-Digital Health:

> PD Dr. Sebastian Bickelhaupt, Universitätsklinikum Erlangen, Bavaria

> Prof. Dr. Jessica Kissinger, University of Georgia, Georgia

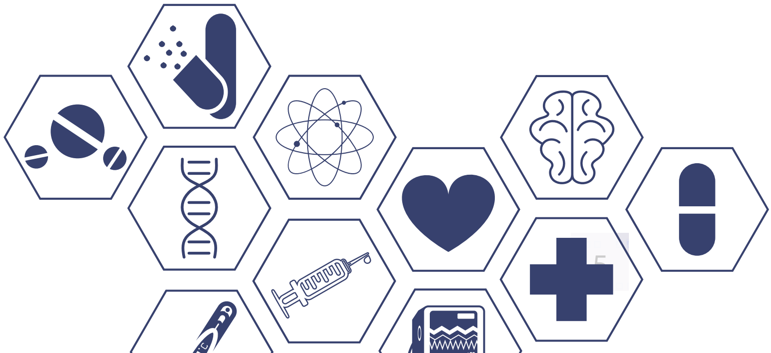
> Prof. Dr. Yves Joanette, Université de Montréal, Québec

> Prof. Dr. Agma Traina, Universidade de São Paulo, São Paulo

> Prof. Dr. Yu Changbin, Shandong First Medical University, Shandong

> Dr. Michael Giretzlehner, RISC Software GmbH, Upper Austria

The present White Paper chooses to approach the advancement of Big Data and AI in healthcare by focusing on a major trend in digital health: Federated Learning (FL) and Federated Analysis (FA). This approach proposes to harness the full potential of health data by enabling the secure exploitation of multiple data sources without having to pool data in a single site (AbdulRhaman, 2020). FL/FA can be presented as a response to present legal, ethical, and technical challenges that limit data sharing across institutions and jurisdictions and thereby reduce the capacity to conduct collaborative data-driven research at a national and international scale (Kairouz et al., 2021). While FL/FA presents genuine opportunities for the enhancement of Big Data and AI for research and innovation, this approach also raises several questions regarding privacy protection, data reliability, and resource utilization, among others. These are the specific issues investigated in this document.



While exploring the potential and challenges of FL/FA for collaborative research in digital health, this white paper also describes robust platforms and technologies showing how FL/FA can be made possible in today's healthcare systems.

The collaboration between the RLS-Digital Health members has shed light on projects that stand out as powerful examples of the promises of FL/FA for data-driven research and innovation in healthcare. Based on these inspiring initiatives, this white paper presents key conditions that could drive the establishment of successful infrastructures able to connect and analyze high quality and real-time data sources, while ensuring that the best standards for data protection and normalization are in place. These conditions could help us define and build a model for data-driven collaborative research at the national and international scales that could benefit researchers, innovators, decision-makers, and patients.

Data-driven research in healthcare: potentials and challenges

Advances in data-driven research across institutions and jurisdictions

Despite the crises and tragedies that it produced, the COVID-19 pandemic opened pathways to genuine progress in data-driven collaborative research and innovation across the globe (Bragazzi et al., 2020). When SARS-CoV-2 was first identified in China in January 2020, researchers knew very little about this new coronavirus or how to respond to this threat. However, after only a few days, the virus' entire genome sequence was identified and shared with the entire research community[4]. By comparison, during the SARS outbreak in 2003, this same effort took almost three months, and before that, the disease was originally thought to be caused by chlamydia.

The COVID-19 pandemic led to several initiatives aimed to accelerate national and international data sharing for research, treatment development, and policy evaluation. Shared information included: molecular data (from sequences to drug targets), epidemiological data, intervention data, as well as and public policies and strategies that were key to facilitate international collaboration and evidence-based decision-making to combat the virus (Dagliatti, 2021). New data commons and repositories were put in place to accelerate the pooling and use of COVID-19 data for research, such as the European Data Portal and Coronanet. Above all, current infrastructures for data storage, access, and processing were used at full potential to enable research collaboration at national and international scales. This is illustrated by the COVID-19 initiatives launched by recognized data infrastructures such as the UK Biobank and International Severe Acute Respiratory and emerging Infection Consortium (ISARIC).

Such developments emphasized the importance of data governance alignment between research teams

[4] https://ec.europa.eu/research-and-innovation/en/horizon-magazine/covid-19-how-unprecedented-data-sharing-has-led-faster-ever-outbreak-research

and institutions, as well as the significance of common standards and terminologies for data harmonization (Ros et al., 2021).

Regarding data formats in particular, common data models such as OMOP and communication standards such as FHIR played a key role in the sharing and use of COVID-19 data for research and innovation.

In the end, these collaborative efforts led to a rapid increase in knowledge production and dissemination, as shown by the thousands of publications produced during the pandemic. Vaccines were developed in record time and new treatments were offered to patients within a year. However, this rapid leapfrog in data-driven research and collaboration also revealed important limitations in the ability to use and share health data to improve population health and address the challenges of healthcare systems.

Challenges in data-driven research at a national and international scale

The COVID-19 pandemic showed how in a short period of time and a moment of emergency an international community of various actors can agree on common standards and terminologies to name a disease, variants, genes, tests, etc. This was an international effort that brought tremendous benefits to research and to the development of effective solutions to contain and combat the pandemic.

However, there are still many challenges to data-driven research within and across borders which limit the capacity to conduct collaborative projects based on large diverse high-quality health data (Abdulrahman et al. 2021; Nguyen et al. 2022; Ros et al., 2021):

> **Privacy issues:** health data are sensitive personal information which require an additional level of protection and attention to ensure privacy and confidentiality;

> **Restrictive regulations:** As health data is considered sensitive personal information, existing regulations (i.e. GDPR, HIPPA, etc.) tend to restrict the transfer of data from one country to another, and most often from one institution to another.

> **Ethical concerns:** Citizens in most countries share similar concerns about the use of their health data and require guarantees for reuse: individual informed consent, information about data use, ability to withdraw consent, etc.

> **Lack of data standardization and interoperability:** Health data are collected and stored in diverse formats that do not always follow international standards and terminologies, which make data linkage and reuse more difficult.

> **Poor data quality and availability:** Health data is known to be of highly varying quality and difficult to exploit due to missing standardization, especially across healthcare systems. Some data can be found in a structured format (imaging data, sociodemographic data, tabular lab findings), yet much data is unstructured (free-text notes, scanned information, etc.)

> **Lack of practice alignment in research evaluation:** On top of all challenges, Institutional Review Boards within and across countries adopt different evaluation standards and practices to assess research projects, which can create barriers to international collaboration.

These challenges are not new and have posed serious obstacles to health data reuse for research and innovation for decades (Price et al., 2019). Yet, with the progress of data analytics and the rapid increase in data generation through information systems, sensors, and applications, these challenges are now seen as hazards that could put an halt to AI development and implementation in healthcare (Morley et al., 2020).

Hence, AI algorithms, especially machine learning and deep learning applications, require a large amount of structured and curated data for training and validation. High quality, diverse, and representative data are responsible for the precision, robustness, and transferability of AI algorithms across healthcare environments (Peifer et al., 2020).

Without good data, algorithms may bring very poor results in practice; this is precisely what the common concept “garbage in, garbage out” implies.

As pointed out before, the availability and accessibility of health data sets is greatly restricted at present, and current efforts to pool massive data sets for AI research and innovation may face additional limitations. First, although the number of data repositories has increased in the last years, these mainly contain research data that are not representative of patient populations and genuine care pathways (ie. clinical studies data, non-observational data, etc.). In other words, data from research repositories are valuable but contain biases, which can limit the real-life performance of algorithms that are trained and calibrated using this information.

Second, efforts to structure real-life data (i.e. EHR, IoTs, etc.) for secondary use have been remarkable in the last years, notably through the development of data lakes and warehouses for research and analytics at an organizational or national scale (Rieke et al., 2020). Still, these infrastructures are mostly based on a centralized model where real-life data is stored, curated, and exploited in a unique site.

Sharing real-life data beyond the site where they are collected and generated remains a major challenge in several jurisdictions. It poses not only legal and ethical challenges, related to privacy and data protection, but also technical ones.

Anonymizing data, managing safe and efficient access, and transferring data is a non-trivial activity that requires important resources and multi-disciplinary expertise, especially in a context where data protection regulation moves fast (Daglieti, 2021).

Therefore, as an alternative to data centralization and releasing, alternative models have been envisioned and experimented. Federated Learning and Analysis may be an efficient way to bypass current hurdles related to data sharing and accelerate data-driven research and innovation across sites and frontiers. In the following, this paper will explore the benefits of this new paradigm for data exploitation and present some of the new challenges that it raises for data-driven research in healthcare.

Federated Learning and Analysis in health care: potential and challenges

What is Federated Learning and Federated Analysis?

Guided by privacy concerns and legal restrictions to data sharing, a paradigm has recently emerged in data analytics and machine learning (ML) which has been called Federated Learning (FL). According to AdbulRahman and colleagues (2021), FL "is a privacy-preserving decentralized approach, which keeps raw data on devices and involves local ML training while eliminating data communication overhead". FL thus presents an alternative to centralized systems for processing data and training AI algorithms. In FL, data is kept on the original sites where it is collected or generated (these sites are sometimes called the "nodes"). Sites agree to collaborate on jointly processing data or training a model, under the coordination of a central server. The server receives analysis information from the sites (statistics, parameters, gradients, weights, etc.), but never the raw data itself. The information is then aggregated on the central server for an enhanced and more performant algorithm trained on a much larger data base. The model or statistics can then be shared with the different sites in order to contribute to collaborative knowledge. At no point in time do sites have access to raw data from collaborators, only to analytics results.

Through this decentralized privacy-preserving process, FL offers the potential to bypass legal, ethical, and technical issues related to the pooling and sharing of sensitive and potentially identifiable information in healthcare (Rieke et al., 2020).

Research has shown that ML models trained through FL can attain levels of performance that are comparable to those trained on centrally hosted data sets and are superior to models that are trained on isolated single-site data (Li et al., 2019; Roy et al., 2019).

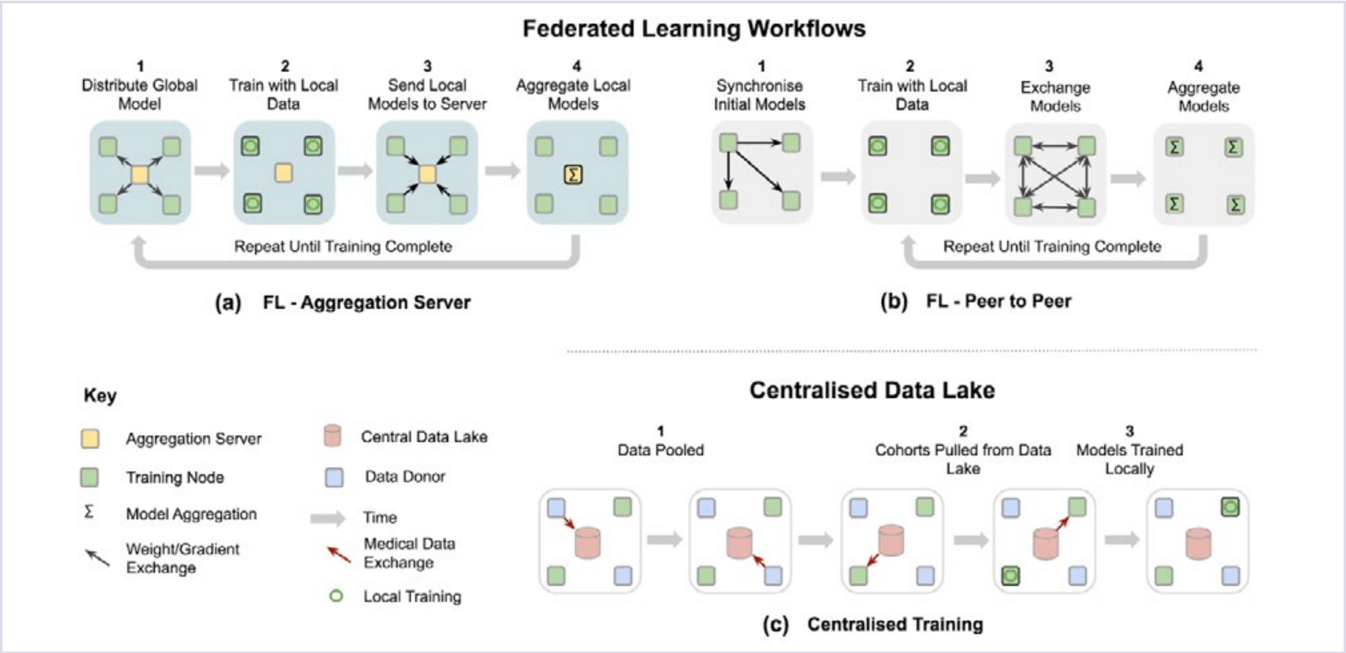


Figure 1: Example federated learning (FL) workflows and comparison to learning on a Centralised Data Lake (extracted from Rieke et al., 2020)

FL, however, is only one approach to Federated Analysis (FA), which covers a greater scope of methodologies and processes for decentralized data analysis.

Hence, FA involves applying basic data science methods (statistics, linear regression, etc.) for data analysis, while FL focuses on training machine-learning models remotely and obtaining aggregated prediction results.

In this sense, FL can be considered as a subset of FA, as it involves a specific type of machine learning-based analysis within the broader context of decentralized data analysis.

In this larger context, one disadvantage of FL is the dependence on a central server, which requires all participating sites to agree on one trusted central body, whose failure or inefficiency could jeopardize the training and analysis process. One alternative to FL is a decentralized approach that does not require a central coordinator. This has already been envisioned and put in place through an infrastructure such as the Decentralized Zero-Trust IoT Data Fabric, developed at the University of Georgia (USA)[5].

This research endeavor aims to establish a decentralized data architecture founded on the principles of web 3.0 and blockchain. The proposed system would enable each data owner to exercise complete authority over their data and allow them to grant or revoke access and use to any user without the involvement of an intermediary.

Additionally, all alterations made to the data and records of access would be traceable and subject to audit, thus ensuring transparency and accountability. This model is called “zero-trust” because it does not require data owners or fiduciaries to have trust in a central coordinator so as to decide on which data access and usage are acceptable and should be authorized.

[5] To see more: https://sensorweb.engr.uga.edu/index.php/wns/

Types of Federated Learning Frameworks

There are several types of Federated Learning frameworks that can be implemented in healthcare to facilitate decentralized data analysis and AI model training (Joshi et al., 2022; Mammen, 2021):

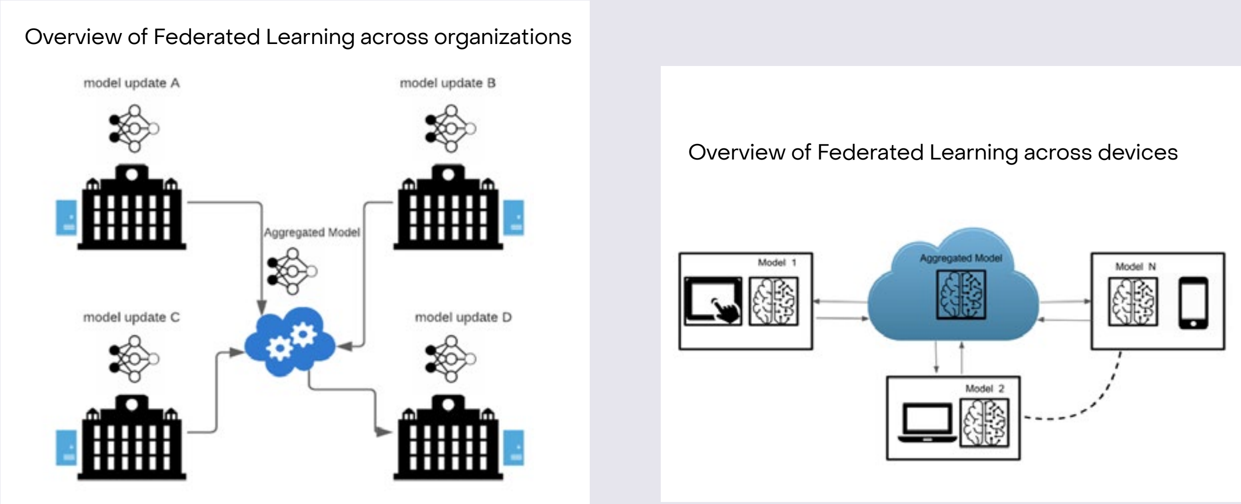
Vertical Federated Learning - can be used for instance when different organizations have data about the same group of patients but with different features. FL allows for the building of an AI model based on a more complete dataset.

Horizontal Federated Learning - can be used when different organizations have data with the same features but about different groups of patients. FL can be applied to train a model on a larger dataset containing a higher number of patients and more variability.

Federated Transfer Learning - involves adding a new feature to a pre-trained model, similar to traditional machine learning. An example of this is extending vertical FL to include additional sample instances that may not be present in all collaborating organizations.

Cross-Silo Federated Learning - is used when the participating nodes/centers are fewer in number and available for all rounds of model training. This framework is normally applied to organizations which have a large amount of data but cannot share them. It can rely on vertical or horizontal FL.

Cross-device Federated Learning - this framework is appropriate in the case when the number of participating nodes is high and when each node has small amounts of data. As a result, the cross-device framework develops models for large-scale dispersed data within the same application. This can be the case when training a model on devices such as mobile and IoTs.


Overview of Federated Learning across organizations

Overview of Federated Learning across devices

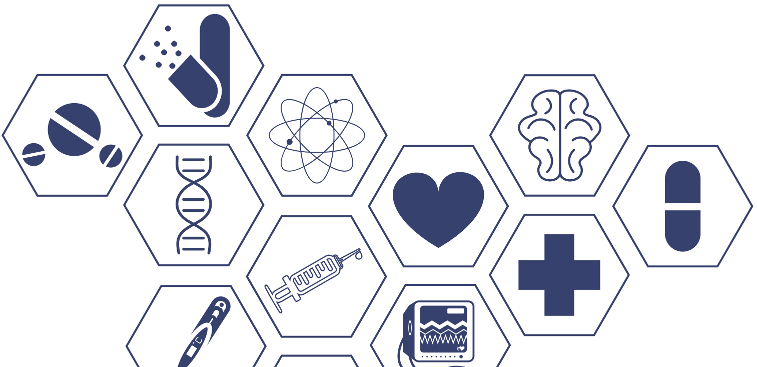
Figure 2: Federated Learning across devices and across organizations (extracted from Mammen., 2021)

Potential benefits of Federated Learning and Analysis for healthcare research

In comparison with centralized data analytics frameworks, FL/FA may offer several benefits for data-driven research and collaboration in healthcare. Below are those most reported by the experts consulted and the literature (Mammen, 2022; Rieke et al., 2020; Sheller et al., 2020; Xu et al., 2020):

- **Data Privacy:** FL/FA could enhance the confidentiality of local data by training machine learning models on devices and nodes without sharing raw data (only aggregated information or weights and gradients) with a central server. This safeguards sensitive patient data and ensures compliance with privacy regulations such as HIPAA, PIPEDA and GDPR, reducing the risk of data breaches.
- **Local data control:** FL/FA could empower healthcare institutions to maintain control over their local data while still generating insights and performing statistical analysis in a collaborative manner. This provides autonomy and control over data, which may be necessary for regulatory or institutional requirements.

- **Data diversity and scalability:** FL/FA allows for the aggregation of diverse data from multiple healthcare institutions, enabling the development of robust and generalizable AI models. This can enhance model performance and provide better insights, leveraging the potential of more complete and diversified datasets while eliminating the need for data transfer or centralization.
- **Real-time updates:** FL/FA allows for continuous model updates as local models on edge servers and/or devices are refined, enabling real-time adaptation and improvement of the global model. This is particularly valuable in dynamic healthcare settings where data distribution and characteristics may change over time.
- **Cost-effectiveness:** FL/FA bears potential to be cost-effective as it can eliminate the need for data transfer or central data storage, reducing communication overhead and associated costs. It also allows healthcare institutions to leverage their existing infrastructure and resources for local analysis.
- **Collaboration and knowledge sharing:** FL/FA can promote collaboration among healthcare institutions by allowing them to work together in developing a global model or project, while preserving local data security and control. This may foster knowledge sharing, expertise exchange, and joint research efforts, which could lead to advancements in healthcare practices and patient outcomes.



Challenges raised by Federated Learning and Analysis and associated requirements

Despite the potential benefits of FL/FA for healthcare research and innovation highlighted, there are still several concerns and challenges raised by experts and the literature that limit its implementation within and across institutions (Kairouz et al., 2021; Liu et al., 2022). This is why it is necessary to implement specific techniques and processes to address these challenges and facilitate the deployment of FL/FA in healthcare settings. Some of these challenges and associated requirements are as follows :

Data heterogeneity: Health data is known to exhibit variability in aspects such as data formats, data quality, data distribution, and data representation. Therefore, datasets across nodes and institutions can present different characteristics that can limit the ability to analyze them in a decentralized manner (Kairouz et al., 2021; Liu et al., 2022)

Requirements: Data standardization and interoperability are key to ensure the success of a FL/FA infrastructure. Harmonization across nodes and sites can be facilitated through the application of formats such as the Fast Healthcare Interoperability Resources (FHIR) format for EHR data and the Digital Imaging and Communications in Medicine (DICOM) format for imaging data. Moreover, corrections need to be applied to ensure that non-identically distributed datasets can be used together to train a model.

Data bias: This issue refers to the presence of systematic errors or prejudices in the data stored in nodes participating in the FL/FA process. This can occur for instance due to the choices made by each participating institution during data collection and processing (Kairouz et al., 2021). Different devices can be used to collect, analyze, and store data, and data can be collected from different populations across sites, which can lead to biased results in the aggregated model.

Requirements: Within and across sites, bias checks and corrections need to be constant during the FL/FA process. Moreover, some techniques can be used to control for this risk and ensure more fairness in the process; this is the case in agnostic federated learning (Mohri et al., 2019). In this scenario, the central server can apply weighting or fairness mechanisms when aggregating the contributions from different nodes. This allows for equal representation and consideration of nodes with varying data characteristics.

Data security and confidentiality: Recent studies have shown that FL/FA fails to provide sufficient privacy guarantees, as sensitive information could be revealed during the analysis and training process (Mothukuri et al., 2021 ; Yin et al., 2021). Hence, during the process, nodes send information to a central server, which renders the infrastructure vulnerable to several types of attacks such as property inference attacks, reconstruction attacks, and membership inference attacks (Hu et al., 2021). These attacks aim to identify whether or not an individual was present in the training datasets, and this presents privacy risks for individuals.

Requirements: Several privacy-preserving techniques must be implemented in order to limit the risk of data reidentification during the training process (see box below). Still, risk can never be reduced to zero, as there is a balance to find between the accuracy of information sent to the central server and the level of privacy that participants wish to maintain within the FA/FL infrastructure.

Data poisoning: This is a case where one malicious institution aims to poison the global model by sending model updates derived from mislabeled data. Such data poisoning attacks can cause substantial drops in information accuracy and model precision, even with a small percentage of malicious participants (Tolpegin et al. 2020).

Requirements: The risk of data poisoning can be reduced then there are controls on which institution can contribute to the FA/FL process. Within an academic network, for instance, risk of malicious attacks on models coming from participating institutions is greatly reduced as requirements to participate can be very strict and follow stringent laws and ethical guidelines (see part 3 for instances).

Communication and computational challenges: FL/FA entails information analysis and exchanges between local models or servers, which can result in increased communication and computational overhead. This can pose challenges in terms of bandwidth, latency, and computational resources, particularly in large-scale healthcare settings with distributed data sources.

Requirements: FL/FA require significant resources, including computational power, storage, and expertise, at local sites. Ensuring that all participants have the necessary resources to actively participate in the analysis and training process can be challenging, especially for smaller healthcare institutions or resource-constrained settings.This can be difficult as human expertise in data engineering and data science tend to be scarce in publicly-funded healthcare institutions.

Governance and legal/ethical considerations: FL/FA relies heavily on trust, collaboration, and alignment between multiple stakeholders, including healthcare institutions, data owners, ethics review boards, and other parties involved. Developing effective governance models and addressing legal and ethical concerns, such as patient consent, intellectual property, and liability, can be intricate and demanding and necessitates careful consideration and implementation of appropriate measures.

Requirements: Robust governance mechanisms and legal frameworks to ensure proper coordination, data sharing agreements, and compliance with regulations and ethical guidelines need to be implemented and accepted by all participants in the federated infrastructure. These mechanisms and frameworks can take the form of a shared documentation based on a unified consent form, data sharing agreement, and data access pathway; for instance, all of those developed in partnership and implemented through the establishment of a committee representing all parties involved in the FL/FA infrastructure.

Privacy-preserving techniques in Federated Learning - simple definitions

Differential Privacy: A technique used to protect the privacy of individuals by adding noise or randomness to data in a way that preserves statistical accuracy while preventing re-identification of individuals. It provides a mathematical framework for quantifying and controlling the privacy risks associated with sharing or analyzing sensitive data, such as personal or confidential information, while enabling data analysis.

Homomorphic Encryption: Homomorphic encryption is a cryptographic technique that allows data to be encrypted in a way that can be processed by users without decrypting it. In other words, it allows computations to be performed directly on encrypted data, without the need to decrypt it first. This provides a high level of data privacy and security, as the original data remains encrypted throughout the computation process, including when it is stored and transmitted.

Secure multi-party computation (SMPC): Also known as secure computation or secure function evaluation, this cryptographic technique enables multiple parties to jointly compute a function on their data, without revealing them to each other. It allows parties to collaborate on computations while keeping their individual data confidential. SMPC ensures that the inputs and outputs of the computation remain private and secure, even when the computation is performed on untrusted or potentially malicious systems or networks.

Focus on national and international FL/FA platforms across RLS-Sciences' regions

The GA4GH Beacon Project

Summary

One of the main challenges facing human genomics research is data scarcity. To overcome this obstacle, the Global Alliance for Genomics and Health (GA4H) launched the Beacon Project in 2014. This initiative aims to facilitate the sharing of genomic and clinical data among federated networks (Fiume et al., 2019). As genomics data are particularly sensitive, the project aims to provide regulatory, ethical, and security guidelines to ensure appropriate measures are in place for data analysis and sharing, following the GA4GH's "Framework for Responsible Sharing of Genomic and Health-Related Data" (Knoppers, 2014). The Beacon Project enables the integration of genomics data from various sources worldwide through a shared query protocol.

By "beaconizing" their omics dataset, hospitals or research institutions can contribute to joint scientific efforts for accelerating genomics research and precision medicine, without compromising data privacy or ownership.

Hence, the Beacon API was designed for researchers and specialists to allow the query of genomic variants and associated information. Thanks to a robust data infrastructure and responsible health practices, genomics data sharing through the Beacon Project enables to derive valuable insights into disease, prognosis, and lifestyle-related genomic variations.

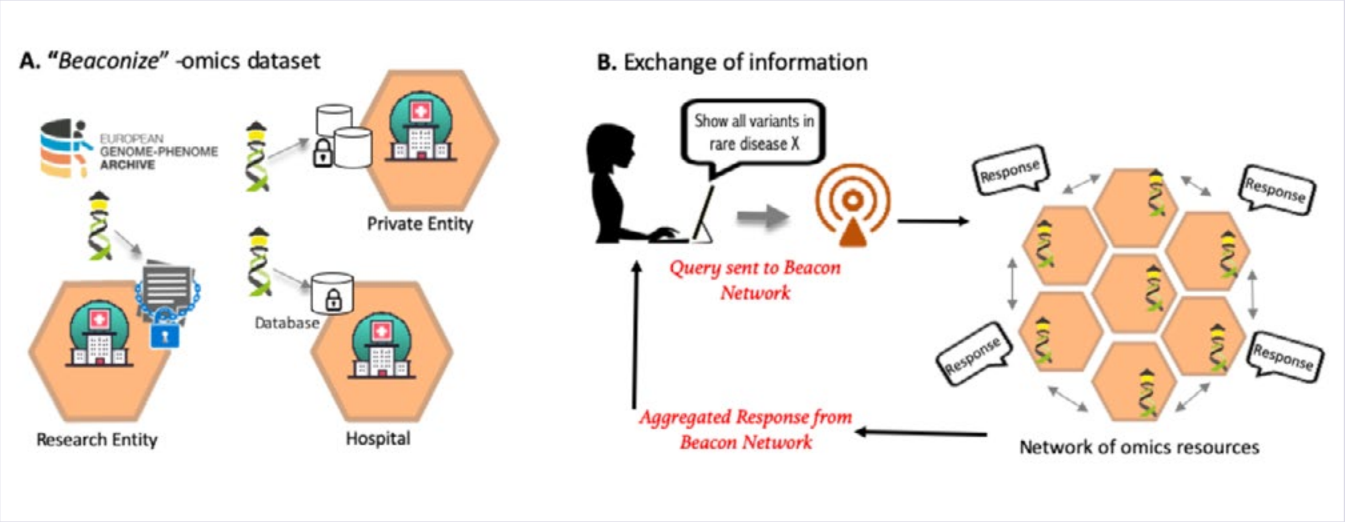

A. "Beaconize" -omics dataset
EUROPEAN GENOME-PHENOME ARCHIVE
Private Entity
Database
Research Entity
Hospital
B. Exchange of information
Show all variants in rare disease X
Query sent to Beacon Network
Response
Response
Response
Response
Aggregated Response from Beacon Network
Network of omics resources

Figure 3: Illustration of the Beacon's infrastructure and functioning (extracted from: https://beacon-project.io/)

Further details

The original versions of the Beacon Project (versions 0 and 1) were limited to indicating the presence or absence of a specific genomic mutation in a dataset from patients with a particular disease or from the general population. The API could enable researchers to query information about a specific allele. For example, a query may involve asking whether a nucleotide, such as C, has been observed at a specific genomic location, such as position 32,936,732 on chromosome 13. The Beacon's response is either "yes" or "no". Therefore, the Beacon API enables remote searches for allelic information of interest without requiring the identification of a particular patient or sample, thereby reducing privacy risks.

The latest version of the Beacon Project (version 2) offers researchers greater flexibility in searching for genomic variants and allows for the inclusion of additional questions about the dataset and participant attributes.

In secure settings, authorized users can link Beacon results to privacy-protected data, such as a patient's electronic health record, and connect it to expert variant annotation. Alternatively, researchers may request access to a dataset returned in their query results, and Beacon Version 2 can provide contact information and data use restrictions to help with the process. This aims to enable researchers to investigate and share genomic variant data alongside critical metadata such as clinical and phenotypic information. This functionality could allow researchers to investigate more questions related to rare and complex diseases[6].

[6] https://www.ga4gh.org/news/new-release-of-ga4gh-beacon-expands-genomic-and-clinical-data-access/

The CODA platform

Summary

The COVID-19 pandemic brought to light the limitations of pooling data at a single place for analysis and research. This was particularly evident in the challenge of providing timely insights into a rapidly emerging public health crisis. To respond to this, the University of Montreal Hospital in Quebec, Canada developed the Collaborative Data Analysis (CODA) platform. From January 2020 onwards, it was designed with input from stakeholders from various fields, including research, clinical, administrative, governmental, ethical, and legal. The key requirements included the ability to perform federated analytics and machine learning, provide support for common medical standards and terminologies, implement measures to minimize disclosure of individual patient data, and ensure deployment using noncommercial software libraries.

The feasibility of the CODA platform was tested at eight hospitals in Canada by enrolling patients with suspected or confirmed COVID-19 over three years. The FL capabilities were tested on reference clinical and imaging data sets from critically ill patients.

The CODA platform was successfully deployed and results from the deployment feasibility study will be published soon in the Lancet Journal. The software code, documentation, and technical documents of the CODA platform were released under an open-source license. From now on, the platform will be used to develop and prospectively validate models for risk assessment, proactive monitoring, and resource usage forecasting among hospitalized and outpatients, as part of ongoing validation efforts.

Further details

The CODA platform consists of a collection of microservices that work in conjunction to enable the decentralized computation of healthcare data (as depicted in Figure 4).

The system includes various services that carry out data ingestion and computation at individual hospital sites (site nodes), a coordinating mechanism for local computations to execute distributed tasks (orchestration hub), and frontend components (dashboard and notebook applications) that facilitate customized analytical queries, data visualizations, and machine learning model training. Prior to data ingestion into the CODA platform, data is de-identified. All communication channels between platform components are secured using Secure Sockets Layer/Transport Layers Security (SSL/TLS).

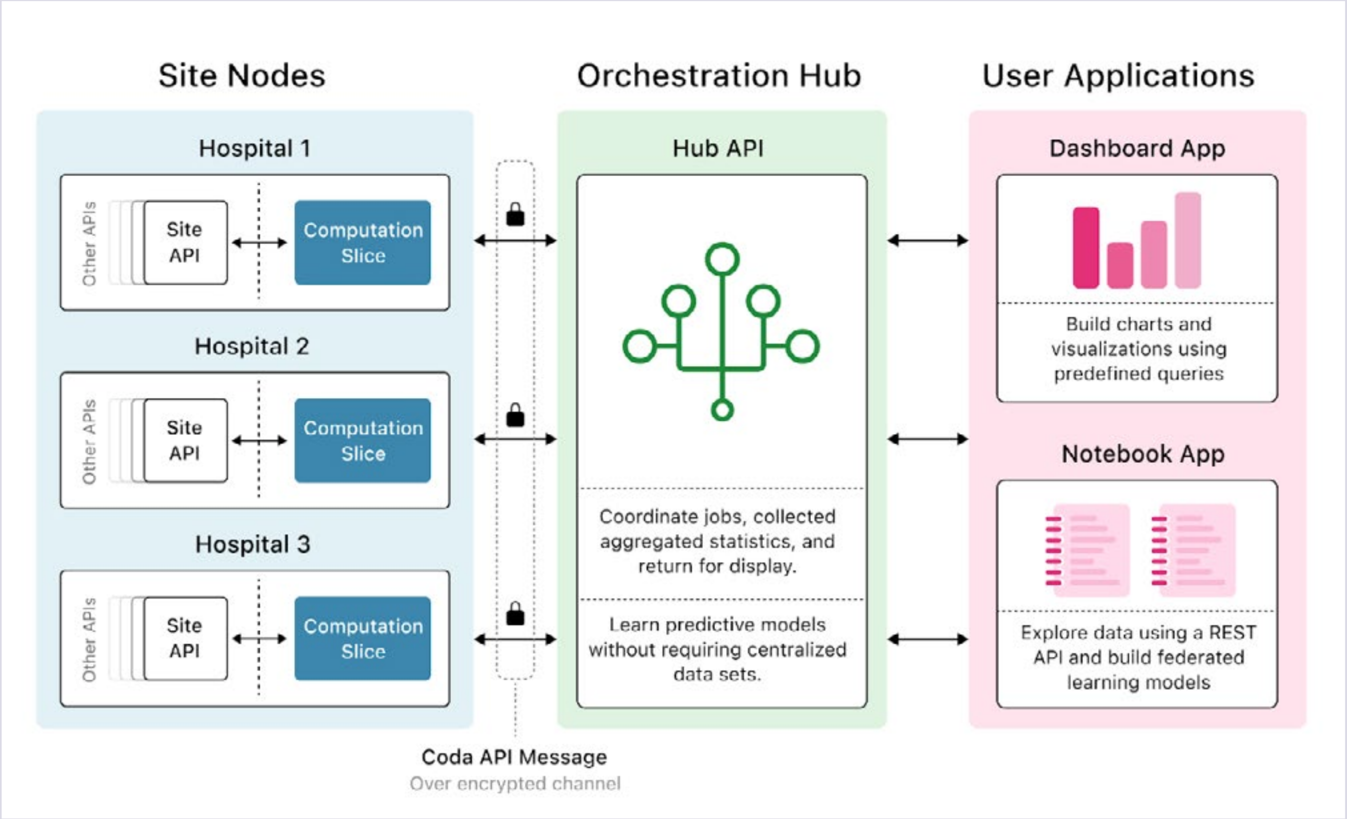

Site Nodes
Orchestration Hub
User Applications
Hospital 1
Other APIs
Site API
Computation Slice
Hospital 2
Other APIs
Site API
Computation Slice
Hospital 3
Other APIs
Site API
Computation Slice
Hub API
Coordinate jobs, collected aggregated statistics, and return for display.
Learn predictive models without requiring centralized data sets.
Dashboard App
Build charts and visualizations using predefined queries
Notebook App
Explore data using a REST API and build federated learning models
Coda API Message
Over encrypted channel

Figure 4: A high-level overview of the CODA network infrastructure (source : CITADEL)

The site nodes, located within institutional firewalls at participating healthcare institutions, include components for storing and retrieving de-identified electronic health record (EHR) data, imaging, and waveform data. The EHR data is stored in the Fast Healthcare Interoperability Resources (FHIR) format, while imaging and waveform data are stored in the Digital Imaging and Communications in Medicine (DICOM) format.

The CODA feasibility study aimed to deploy the platform across nine public hospitals in Québec, Canada, as illustrated in Figure 5.

A Governance Framework was established to formalize the legal and ethical terms of collaboration between participating institutions[7].

Eight out of the nine enrolled sites successfully deployed the platform locally and are connected to the CODA network, while one site dropped out due to IT resource limitations, and two sites have not yet provided patient data. As of publication, the CODA feasibility study cohort comprises 1,091,540 patients, with a total of 46,181,904 FHIR objects and 3,777,716 imaging studies.

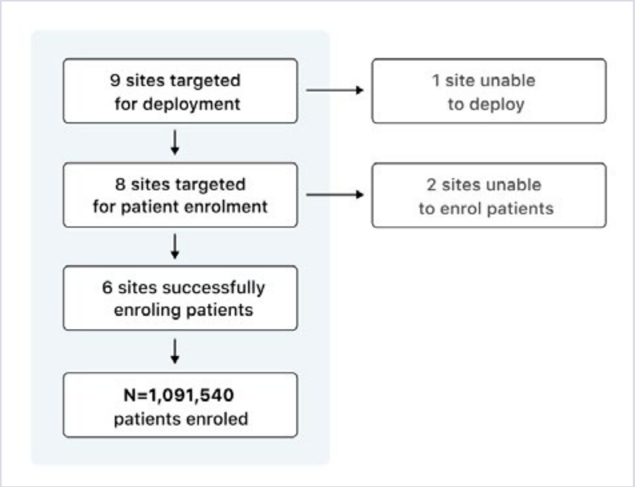

9 sites targeted for deployment
1 site unable to deploy
8 sites targeted for patient enrolment
2 sites unable to enrol patients
6 sites successfully enroling patients
N=1,091,540 patients enroled

Figure 5: Flowchart of Site And Patient Enrolment In Feasibility Study (source : CITADEL)

Software code, documentation, and technical documents regarding the CODA platform were released under the GPL v3 license (www.coda-platform.com). A set of standard FHIR templates were developed to assist users in migrating from legacy storage formats. An API Reference Specification was developed to guide the implementation of the various platform components. A Deployment Guide was created to facilitate the creation of sandbox/testing environments. A Data Security Framework was created to govern implementation practices relating to the authentication and authorisation of users, as well as data protection.

GAIA-X in Europe with a focus on Germany

Summary

Gaia-X is a European initiative launched in 2019 by the former German Minister of Economic Affairs, Peter Altmaier, and his French counterpart, Bruno Le Maire. Alongside the efforts to create a European Health Data Space, the Franco-German collaboration aimed to further cooperation in the field of data sharing and AI development through an open-source and secure data infrastructure that could safeguard and enhance Europe's digital sovereignty. In early 2022, the initial implementation of Gaia-X began with the development of data spaces and associated services, such as the Mobility Data Space. The German Federal Ministry for Economic Affairs and Climate Action facilitated this progress through its funding competition titled "Innovative and Practical Applications and Data Spaces in the Gaia-X Digital Ecosystem."

[7] https://github.com/coda-platform/guides-and-policies/tree/main/policies/governance

Eleven projects were selected with the aim to implement Gaia-X and develop innovative digital solutions with significant market potential.

Gaia-X establishes the groundwork for an autonomous, united, and transparent data infrastructure that adheres to European principles. It represents a strategic approach that fosters collaboration among diverse stakeholders in constructing a data environment that complies with European regulations and promotes trust (Federal Ministry for Economic Affairs and Climate Action, 2021; Otto and Burnmann, 2021).

Specifically, the health data space encompasses shared capabilities and federated health data spaces, where data can be accessed in a granular and selective manner. In the future, these federated data spaces will be implemented at regional, national, and European levels. The aim is to ensure that the data space contributes to patient care processes while enabling the secondary utilization of data at a cohort or population scale. This will establish a data value chain connecting data holders and users within the extensive and intricate European health ecosystem.

According to Core Gaia-X infrastructural components include:

> Identity & Trust: federated identity management for individuals and organizations

> Federated Catalogue: to publish the registration, consent, and query services

> Sovereign Data Exchange: to manage registration, consent, cloud/edge services, and data query and access services

> Compliance: rights management, onboarding, and certification

The organizational structure of Gaia-X rests upon three fundamental pillars: 1) Gaia-X Association for Cloud and Infrastructure (AISBL), 2) National Gaia-X Hub, and 3) Gaia-X Community. The German Gaia-X Hub serves as the primary contact point for companies, organizations, and individuals in Germany seeking to acquire more information about the project or engage with the open-source community.

To see more:

https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

https://gaia-x.eu/wp-content/uploads/2022/05/Gaia-X-Event-Report_Health-Data-Space-Event-4_4_2022.pdf

https://healthmanagement.org/uploads/article_attachment/gaia-x-federated-data-infrastructure-the-future-of-data-management.pdf

Use Case in Germany: Health-X dataLOFT

Health-X, supported by the German Federal Ministry of Economy and Climate Protection, aims to establish a validated, transparent, and interconnected platform for health data called dataLOFT. Gathering several public and private actors in the field of health data in Germany, **Health-X dataLOFT intends to ensure compliance with Gaia-X standards while enhancing health data accessibility.** The primary focus of the project is utilizing data from two key healthcare sectors: primary care institutions like hospitals and doctors' offices, as well as personal health data from the secondary health market, which includes data from apps and sensors.The project also aims to place citizens at the core, emphasizing their authority in determining the collection, usage, and control of their own health data. They possess the autonomy to decide which data is made available for personal health concerns, medical care, and research, based on their own values and preferences.

Through four use cases, the HEALTH-X dataLOFT project is constructing a data space specific to the health domain. The data space aims to serve as the foundation for these use cases, addressing important issues related to empowering citizens, promoting health prevention, facilitating healthy aging, and enhancing clinical care.

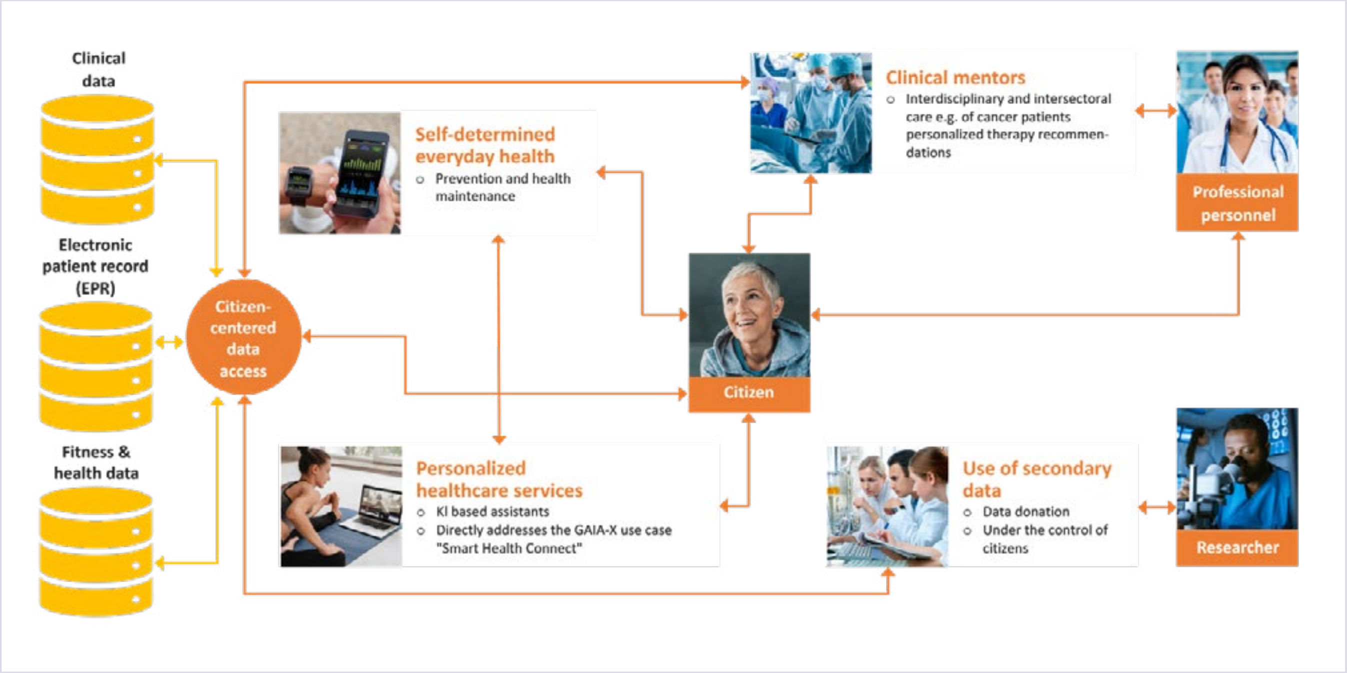

Clinical data
Electronic patient record (EPR)
Fitness & health data
Citizen-centered data access
Self-determined everyday health
Prevention and health maintenance
Clinical mentors
Interdisciplinary and intersectoral care e.g. of cancer patients personalized therapy recommen-dations
Professional personnel
Citizen
Personalized healthcare services
KI based assistants
Directly addresses the GAIA-X use case "Smart Health Connect"
Use of secondary data
Data donation
Under the control of citizens
Researcher

Figure 6: Use Cases for which the dataLOFT platform will be tested (source: Health-X 2023)

To see more:

https://www.computer.org/csdl/magazine/mu/2022/01/09770011/1D830WmTsDS
https://www.isst.fraunhofer.de/en/business-units/healthcare/projects/HEALTH-X-dataLOFT.html

The BORN Project in Bavaria, Germany

Launched in 2022, the BORN Project is a collaboration between the Bavarian Center for Cancer Research (BZKF) and the radiological institutes of all six university hospitals in Bavaria. It was launched with the participation of Markus Blume, the Bavarian State Minister for Science and the Arts, and Klaus Holetschek, the Bavarian State Minister for Health and Care. Mint Medical and Brainlab will closely collaborate with the university hospitals and the BZKF to establish a consistent and structured reporting system for oncological imaging. They will also develop a secure IT infrastructure to facilitate the capture and exchange of data.

Initially, standardized reading templates will be developed for six different entities, ensuring uniform reporting of cancer cases across all Bavarian university hospitals. Following clinical evaluation at these hospitals, the standardized collection and assessment of imaging data can be expanded to other healthcare facilities and practices, benefiting patients throughout Bavaria. This will also create an unprecedented data repository, covering a population of 13 million residents, that can be utilized for prospective and retrospective studies.

The ambition is to create a unique dataset for the development of image-based biomarkers and machine learning techniques. This will help collect and reuse structured and comprehensive health data while safeguarding data privacy, as well as enhance the usability of healthcare data for patient treatment, research, development, and policymaking.

In the BORN Project, data collection is decentralized to individual clinics, ensuring the confidentiality, protection, and management of the data. The respective centers retain ownership of the data they contribute, allowing for differential release ranging from general use in medical research to specific collaborative projects.

Conditions for success of FL/FA for data-driven international research

In light of the projects presented in the previous section and the analysis of the potentials and challenges of FL/FA in health care, several conditions can be drawn to ensure that FL/FA comes to consolidate and facilitate data-driven international research:

> **Foundational principles and values:** These must be shared by all participants of a FL/FA platform in order to ensure that governance and technical frameworks and processes will reflect the main objectives and interests of the participants, be they institutions or individuals. Examples of these principles and values can be: Open Science, FAIR principles, and prioritizing public-interest research.

> **Regulatory compliance and alignment:** FL/FA platforms need to adhere to relevant legal requirements, such as data protection laws (e.g., GDPR, PIPEDA, HIPAA) and AI regulations (e.g. EU AI Act, Bill C-27 in Canada), as well as broader ethical guidelines for research involving human subjects. Nonetheless, these regulations and guidelines can vary from one jurisdiction to another; this is why a minimum level of legal and ethical alignment across countries and regions is necessary to ensure that FL/FA platforms can reach an international scale.

- **Interoperability and standardization prioritization:** Data quality, accessibility and usability across an FL/FA platform is highly dependent on the existence of data formats, standards, terminologies, and protocols shared by all participants. Systems interoperability and data standardization are key to ensure compatibility and harmonization across the data sources. Moreover, ethics documentation needs to be standardized through shared consent ontologies describing informed consent, opt-in and opt-out procedures, terms-of-use, specific research usage, and more.
- **Robust security frameworks and processes:** A FL/FA platform should integrate robust privacy-preserving mechanisms to protect patient confidentiality, comply with data protection regulations, and prevent unauthorized access or data breaches. These mechanisms need to be adequate with privacy policies and based on top industry standards to verify data users' identities and roles, and monitor data usages.
- **Standardized application programming interfaces (APIs):** APIs can streamline access to data stored in different systems participating in a FL/FA platform. These APIs empower researchers and other users to query and analyze distributed datasets, and enable data holders to effectively manage the data entrusted to them.
- **Enhanced patient accessibility and control:** Trust in FL/FA platform largely relies on the confidence patients will have in its security and benefits for care enhancement and research advances. Along with professionals and researchers, patients need to be given the opportunity to access data and results from analysis and research, and be able to preserve control over data usages. This can be accomplished through digital consent processes, support for multiple languages and cultures, and dynamic consent practices that allow patients to have control over how their data is utilized, and to discover which benefits are produced out of them.

Acknowledgements

This work was conducted by the RLS-Digital Health with the contributions of multiple experts coming from the participating regions. Their expertise was called upon for the writing and revision of this white paper, and interviews were conducted with members of RLS-Digital Health as well as further experts from their regions to obtain relevant information, examples, and references regarding the development and implementation of FL/FA in healthcare.

Below are the names of the experts consulted during qualitative semi-structured interviews that took place from February to April 2023:

- Michael Chassé, Centre de recherche du Centre hospitalier de l'Université de Montréal, Québec, Canada
- Philippe Després, Université Laval, Québec, Canada
- Vincent Ferretti, Centre hospitalier universitaire Sainte-Justine, Québec, Canada
- Louis Mullie, Centre de recherche du Centre hospitalier de l'Université de Montréal, Québec, Canada
- Jessica Kissinger, University of Georgia, Georgia, USA
- Jaewood Lee, University of Georgia, Georgia, USA
- WenZhan Song, University of Georgia, Georgia, USA
- Peter Zinterhof, LRZ and Salzburg Federal Hospital, Bavaria and Upper Austria, Germany and Austria
- Bjoern Eskofier, Friedrich-Alexander-Universität, Bavaria, Germany
- Christian Wachinger, Technical University of Munich, Bavaria, Germany
- Michael Giretzlehner, RISC Software GmbH, Upper Austria
- Mohit Kumar, Software Competence Center Hagenberg, Upper Austria
- Bernhard Moser, Software Competence Center Hagenberg, Upper Austria
- Agma Traina, University of Sao Paulo, Sao Paulo, Brazil
- Changbin Yu, Shandong First Medical University, Jinan (Shandong), China

The RLS-Digital Health Group would like to acknowledge the financial support of the Fonds de recherche du Québec – Nature et technologies and the Digital Health Consortium of the University of Montreal which enabled the production of the white paper.

The Group thanks Stéphanie Hauschild for graphic design and Cécile Petitgand, CEO of Data Lama, for her remarkable work in the coordination of the production of this White Paper.

References

AbdulRahman, Sawsan, et al. (2020)."A survey on federated learning: The journey from centralized to distributed on-site learning and beyond." IEEE Internet of Things Journal 8.7: 5476-5497.

Bragazzi, Nicola Luigi, et al. (2020)."How big data and artificial intelligence can help better manage the COVID-19 pandemic." International journal of environmental research and public health 17.9: 3176.

Dagliati, Arianna, et al. (2021)."Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview." Briefings in bioinformatics 22.2: 812-822.

Federal Ministry for Economic Affairs and Climate Action. (2021). GAIA-X: Eine vernetzte Datenstruktur für ein europäisches digitales Ökosystem. In 2021. https://www.bmwk.de/Redaktion/DE/Dossier/gaia-x.html

Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S. O., ... & Scollen, S. (2019). Federated discovery and sharing of genomic data using Beacons. Nature biotechnology, 37(3), 220-224.

Fornell. D. (2023) FDA has now cleared more than 500 healthcare AI algorithms. https://healthexec.com/topics/artificial-intelligence/fda-has-now-cleared-more-500-healthcare-ai-algorithms

Hu, H., Salcic, Z., Sun, L., Dobbie, G., & Zhang, X. (2021, December). Source inference attacks in federated learning. In 2021 IEEE International Conference on Data Mining (ICDM) (pp. 1102-1107). IEEE.

Joshi, Madhura, Ankit Pal, and Malaikannan Sankarasubbu. "Federated learning for healthcare domain-Pipeline, applications and challenges." ACM Transactions on Computing for Healthcare 3.4 (2022): 1-36.

Kairouz, Peter, et al. "Advances and open problems in federated learning." Foundations and Trends® in Machine Learning 14.1–2 (2021): 1-210.

Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. The HUGO journal, 8(1), 3.

Li, Wenqi, et al. "Privacy-preserving federated brain tumour segmentation." Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10. Springer International Publishing, 2019.

Mammen, Priyanka Mary. "Federated learning: Opportunities and challenges." arXiv preprint arXiv:2101.05428 (2021).

McCradden, Melissa D., Elizabeth A. Stephenson, and James A. Anderson. "Clinical research underlies ethical integration of healthcare artificial intelligence." Nature Medicine 26.9 (2020): 1325-1326.

Mehta, Nishita, Anil Pandit, and Sharvari Shukla. "Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study." Journal of biomedical informatics 100 (2019): 103311.

Mohri, M., Sivek, G., & Suresh, A. T. (2019, May). Agnostic federated learning. In International Conference on Machine Learning (pp. 4615-4625). PMLR.

Morley, Jessica, et al. “The ethics of AI in health care: a mapping review.” Social Science & Medicine 260 (2020): 113172.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619-640.

Otto, B., & Burmann, A. (2021). European data infrastructures: approaches and tools for using data for the benefit of individuals and communities. Computer Science Spectrum, 44(4), 283-291.

Price, W. Nicholson, and I. Glenn Cohen. “Privacy in the age of medical big data.” Nature medicine 25.1 (2019): 37-43.

Rieke, Nicola, et al. “The future of digital health with federated learning.” NPJ digital medicine 3.1 (2020): 119.

Ros, Francisco, et al. Addressing the Covid19 pandemic and future public health challenges through global collaboration and a datadriven systems approach. Vol. 5. No. 1. 2021.

Roy, Abhijit Guha, et al. “Braintorrent: A peer-to-peer environment for decentralized federated learning.” arXiv preprint arXiv:1905.06731 (2019).

Shaw, James, et al. “Artificial intelligence and the implementation challenge.” Journal of medical Internet research 21.7 (2019): e13659.

Sheller, Micah J., et al. “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data.” Scientific reports 10.1 (2020): 1-12.

Tolpegin, V., Truex, S., Gursoy, M. E., & Liu, L. (2020). Data poisoning attacks against federated learning systems. In Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25 (pp. 480-501). Springer International Publishing.

Xu, Jie, et al. “Federated learning for healthcare informatics.” Journal of Healthcare Informatics Research 5 (2021): 1-19.

Yin, X., Zhu, Y., & Hu, J. (2021). A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys (CSUR), 54(6), 1-36.

Federated Learning and Analysis for Collaborative Research in Healthcare at a National and International Scale

White Paper by
RLS-Digital Health

RLS-SCIENCES

Fonds de recherche
Nature et
technologies
Québec

CONSORTIUM
SANTÉ
NUMÉRIQUE

Université
de Montréal

CONSORTIUM
SANTÉ
NUMÉRIQUE

RISC
Software GmbH

SDFMU

Uniklinikum
Erlangen

USP
Universidade de São Paulo

UNIVERSITY OF
GEORGIA