

PETs in practice: using federated learning to train machine-learning models without centralising data

Privacy Enhancing Technologies

ODI Research
ADVANCING TRUST IN DATA

June 2024

Contents

PETs in practice: using federated learning to train machine-learning models without centralising data **2**
- Background 3
- What is federated learning? 3
- Why use federated learning for research using patient data? 4
- Deployment of full-stack FL 6
- Challenges in deploying federated learning in hospital settings 8
  - Research approval 8
  - Data collection and de-identification 9
- Resourcing, implementation and maintenance costs 10
- Security measures implemented 10
- Limitations 12
- Value flows and tangible outcomes 13
- Hands-on federated learning 13
- Conclusion 14
- Get in touch 14
- Acknowledgements 14

About

This report has been researched and produced by the Open Data Institute and published in June 2024. Its lead author was Claudine Tinsman, supported by Calum Inverarity, Gefion Thuermer and Neil Majithia. If you want to share feedback by email or would like to get in touch, contact the PETs programme team at pets@theodi.org.

To share feedback in the comments, highlight the relevant piece of text and click the ‘Add a comment’ icon on the right-hand side of the page.

PETs in practice: using federated learning to train machine-learning models without centralising data

Federated learning (FL) promises to train machine-learning (ML) models in a distributed manner, keeping the raw sensitive data safe in its original locations. Like other types of ML, it is used by data scientists to solve specific problems that may involve identifying patterns, correlations or anomalies. Federated learning needs a large amount of training data, from which the model can learn in a systematic, repeatable way.

At the ODI, we are interested in understanding the necessary requirements to make PETs transferable to different use cases, and to increase their adoption by other organisations that may have more limited resources. This case study documents the use of federated learning across hospitals in the UK to develop and evaluate a Covid-19 screening test using distributed data. The application enabled the training of a machine-learning model that demonstrated improved performance and was generalisable. At the same time, due to the use of federated learning, training data was retained at source – rather than shared for centralised training – thus enabling greater data minimisation and privacy preservation.

Increased access to sensitive data can be beneficial for social purposes, as demonstrated by this example. We hope that the documentation of this example will prove useful for other non-profit organisations and researchers working on PETs to establish safe access to data to address societal challenges.

This case study draws on insights from desk research and interviews with the team responsible for developing the machine-learning model, as well as those who developed the federated learning framework used in this particular example.

In creating this explainer, we collaborated with the team at Flower Labs who produced a hands-on federated-learning tutorial that you can run yourself, to see how federated learning works in practice.

Background

What is federated learning?

Federated learning (FL) is an ML technique that allows multiple decentralised devices or servers to collaboratively train a global model without ever transmitting local data from its location.[1] The central server communicates with the participating nodes, which could be devices or organisations that steward data on behalf of others. These nodes receive the model from the central server, train it using their local data, and then transmit the model updates back to the server, where they are aggregated and used to create an updated central model.[2] Critically, no raw data is shared between the participating nodes and the central server. This cycle of updating and retraining can be repeated as often as necessary, with the server continually dispatching the updated model back to the nodes for further training and subsequent updates. Since sensitive data does not need to leave its original location, the decentralised nature of this approach can potentially address concerns related to privacy, security, and data access rights.[3]

This report focuses on the use of FL in the context of patient data. Though there are relatively few cases of FL implemented in the wild, it has a wide range of potential applications across various sectors and use cases. These include:

- **Smartphone applications:** The first significant consumer-oriented application of federated learning was Gboard,[4] the keyboard app on Android. Gboard uses federated learning to offer features like next-word prediction (NWP) without directly sharing user data with Google.
- **Content recommendation:** Most content recommender systems are built by collecting a wide range of user data and centrally training a personalised recommender model. In 2021, Brave News, a privacy-preserving news aggregator, was integrated into browsers. This service allows users to subscribe to RSS feeds of their chosen news sources anonymously, choose from 15 different categories of curated sources, and customise their stream by adding or removing sources.[5]

[1] Rieke et al. (2020), 'The future of digital health with federated learning'.

[2] Ibid.

[3] Shteyn, Kollnig & Inverarity. (2023), 'Federated learning: an Introduction'.

[4] McMahan, B. and Ramage, D. (2017), 'Federated Learning: Collaborative Machine Learning without Centralized Training Data'.

[5] Brave (2024), 'Brave News, the privacy-preserving news aggregator, is revamped for a more personal and easily customizable experience'.

- **Finance and e-commerce:** Fraudulent transactions online pose a significant challenge, costing billions annually. Federated learning can improve fraud detection accuracy while preserving data privacy. Several research organisations are currently developing FL systems to collaboratively train a shared model by leveraging a distributed approach while keeping sensitive data locally.[6]

Why use federated learning for research using patient data?

In traditional ML methods, training is centralised, requiring data to be collected at a singular location. Conversely, FL permits data to remain with local devices or organisations, where model training and evaluation occur directly on-site. This approach enables the collaborative creation of AI models while keeping the data within the confines of the participating organisations, preserving data privacy in the process.[7]

In the context of research with patient data, FL can serve as a valuable alternative to Trusted Research Environments (TREs), which are sometimes referred to as Secure Data Environments (SDEs). TREs are data storage and access platforms that allow researchers to access and analyse sensitive data without that data leaving the environment. The measures built into these TREs are designed to provide greater data privacy and security guarantees and to limit access by potentially malicious actors.[8] TREs have become popular in medical research. Currently, the NHS is developing TREs for research that involves the use of electronic patient records in the UK, in compliance with the recommendations of the Goldacre Review.[9]

[6] IBM (2023), 'Building privacy-preserving federated learning to help fight financial crime'.

[7] Soltan et al. (2022), 'Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening'.

[8] Majithia et al. (2024), 'PETs in Practice: using trusted research environments to enable health data research'.

[9] Goldacre, B & Morley, J. (2022), 'Better, Broader, Safer: Using health data for research and analysis. A review commissioned by the Secretary of State for Health and Social Care'.

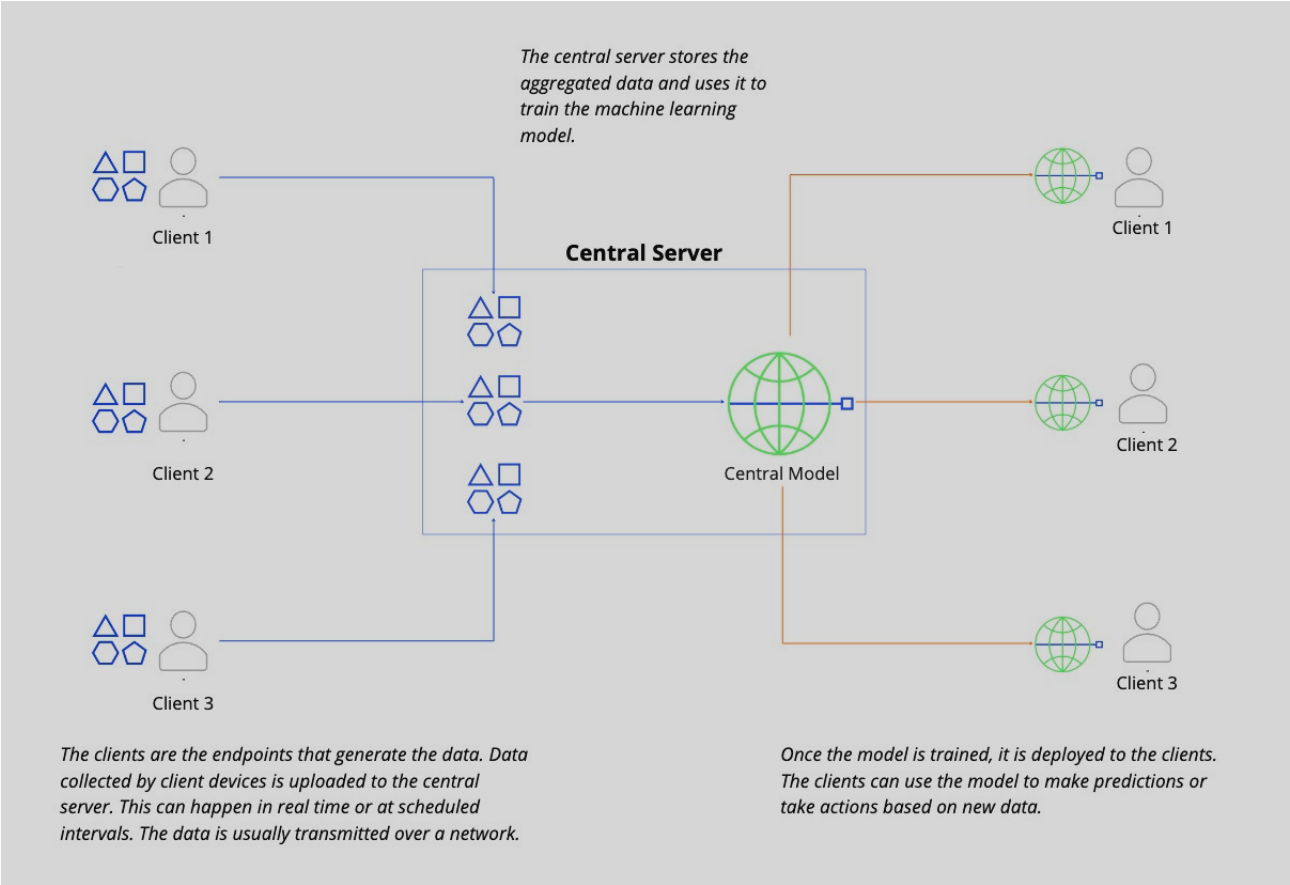


Figure 1: An example of a centralised machine learning model

However, in healthcare research, FL presents two key practical advantages over SDEs. SDEs are a potential single point of failure, while FL's distributed architecture affords a smaller attack surface.

At the height of the Covid-19 pandemic, UK-based researchers developed, validated and piloted the CURIAL-Lab AI screening test for Covid-19 in emergency departments.[10] Using techniques reliant on data centralisation, the CURIAL-Lab test used data routinely collected within one hour of a patient's arrival at a hospital emergency department. Findings indicate that the test yielded a higher negative predictive value for Covid-19 than lateral-flow tests and provided results faster than PCR tests.[11] However, the deployment of the CURIAL-Lab study relied on data centralisation, whereby data was collected from multiple NHS trusts and transferred to a trusted server at the University of Oxford. The data was de-identified for privacy reasons, and secure transfer protocols were implemented.

[10] Soltan et al., Lancet Digital Health (2020), 'Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test'.

[11] Soltan et al (2022), 'Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening'.

The research team developed the CURIAL-Fed-Lab test to address the privacy concerns of patient data transfer. This full-stack FL design aimed to federate the Covid-19 screening test by developing an easy-to-use embedded system – also called full-stack FL—to train and evaluate ML models across four UK hospital groups without centralising patient data.[12]

Because of its utility in training models on sensitive data, FL can motivate healthcare providers to engage in AI research, which could shorten the development timeline, enhance the diversity of data representation, and promote international collaboration.[13] Due to the high degree of technical skill and heterogenous data formats, real-world applications in hospital settings have been limited,[14] with most studies focusing on simulated deployments rather than in-situ implementation.[15]

Deployment of full-stack FL

The figure below provides a visual representation of the FL model training process. The fully annotated diagram allows users to zoom in on details and access additional information about the various elements of the framework. It also includes a narrative walkthrough of the full-stack FL model.

Researchers provided a Raspberry Pi 4 Model B preloaded with a full-stack FL pipeline to four NHS hospital groups in the UK: Oxford University Hospitals NHS Foundation Trust (OUH), University Hospitals Birmingham NHS Foundation Trust (UHB), Bedfordshire Hospitals NHS Foundation Trust (BH) and Portsmouth University Hospitals NHS Trust (PUH). During the training phase, OUH, PUH and UHB de-identified data from patients admitted to emergency departments before the pandemic (who were at the time Covid-19-negative) and tested positive for Covid-19 during the first wave of the pandemic to form and calibrate a global model to predict Covid-19 status.

Model weights, thresholds, and evaluation results were transmitted to a central server, where aggregation and averaging were performed to form a global model. Updated weights for the new global model were transmitted back to the federated clients, facilitating the subsequent round of training. A total of 150 rounds were performed.[16]

---

[12] Ibid.

[13] Ganapathi et al. (2022), 'Tackling bias in AI health datasets through the STANDING Together initiative'.

[14] Crowson et al. (2022). 'A systematic review of federated learning applications for biomedical data'.

[15] Soltan et al. (2022), 'Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening'.

[16] Because researchers could not have predicted in advance that performance would plateau after 50 rounds, they opted for a larger number of rounds to make sure that they hit a plateau. The federated evaluation was carried out after every round of training to monitor performance improvement, and the 'final' global model was evaluated after 150 rounds.

The federated evaluation was conducted by applying the global models to temporally prospective cohorts of patients admitted to hospital during the second wave of the Covid-19 pandemic at OUH, PUH and BH. An additional step of site-specific fine-tuning was conducted for sites also contributing to training (OUH and PUH), and the tuned model was evaluated. Evaluation results were transmitted to the coordinating server for reporting.

The global model evaluation for all patients admitted to BH was repeated on the coordinating server using the same threshold and imputation strategy to confirm the federated evaluation's fidelity. The BH client transmitted summary statistical results from the federated evaluation, which were verified to ensure they matched the results obtained from the evaluation repeated centrally. The extracted BH data was transferred to the server to facilitate this verification.

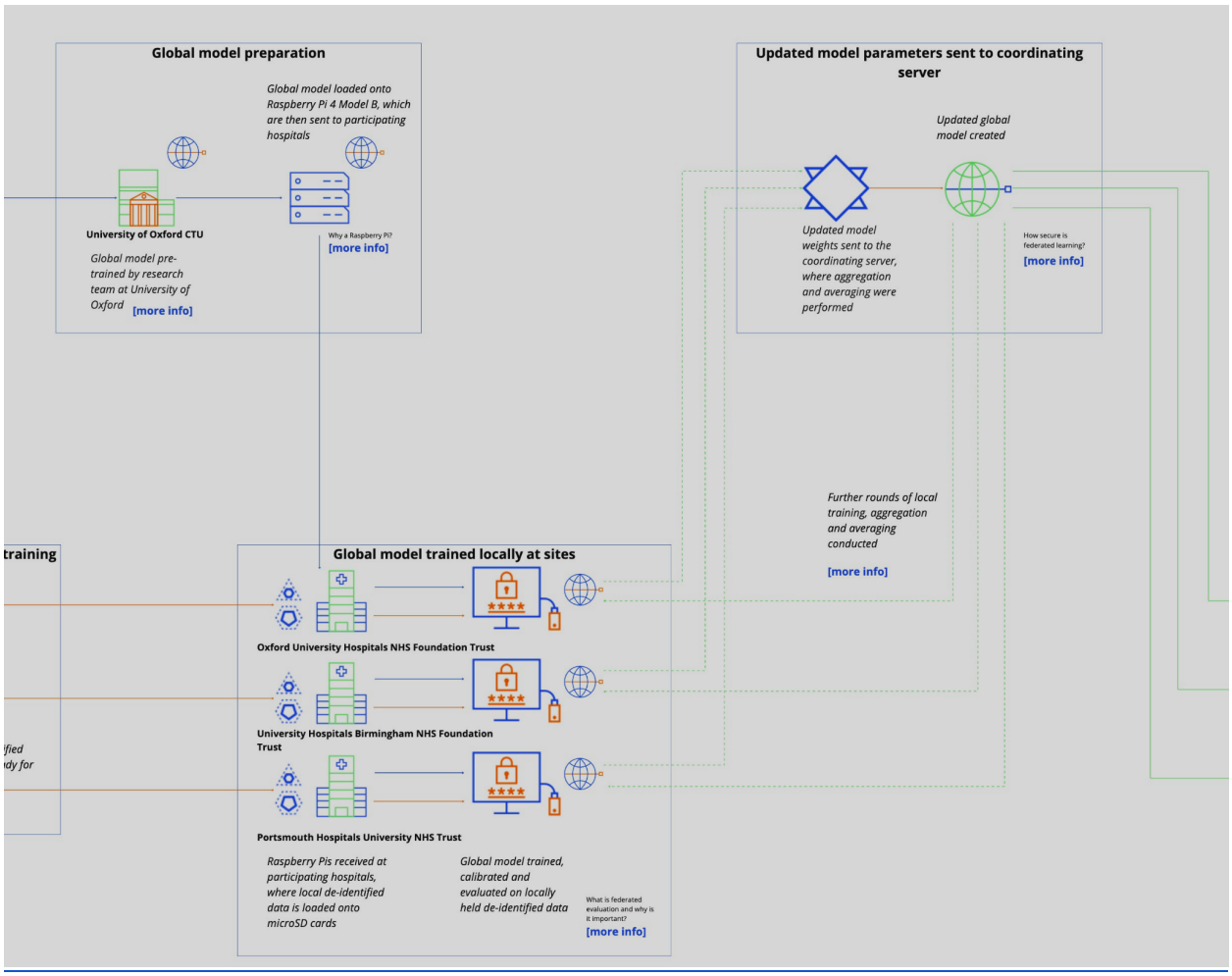


Figure 2: An excerpt of the diagram depicting the deployed CURIAL-Fed-Lab

Challenges in deploying federated learning in hospital settings

Understanding and adapting a technological solution to fit the needs of a specific context and audience is vital to ensure adoption. The full-stack CURIAL-Fed-Lab FL was developed by researchers across multiple universities and NHS Trusts and was tailored to the specific needs of this case study, which was conducted during the Covid-19 pandemic.

Research approval

The study required a lengthy and rigorous approval process before being launched. Researchers developed the full-stack FL and submitted the proposal to the Research Governance and Ethics Assurance Team (RGEA) at the University of Oxford, which evaluated the potential risks and liabilities associated with the study. Upon approval, the university acted as the study's sponsor.[17]

All clinical research conducted within a UK university must have a sponsor. The sponsor's responsibility is to provide insurance for the study and take legal responsibility for initiating and managing the study. Once the university approved the study design, the researchers submitted the proposal to the Health Research Authority (HRA)[18] on 26 March 2020, and received approval within a week. While the administrative process was rapid, the research team faced delays in implementation. Once the team received approval from the HRA, they applied for approval from the research services at each of the study's trust sites. As there were no standardised approval protocols across trusts, the approval processes varied in terms of formality and duration: the longest approval took approximately three months.

Once the approval had been granted at all four trusts, the assigned principal investigator at each site requested the data from the local trust's data team within the study's parameters.

[17] More information on the University of Oxford's process can be found here.

[18] More information about the Health Research Authority can be found here.

Data collection and de-identification

For both training and evaluation, only patient data that met the following criteria was included in the study:[19]

- Over the age of 18
- Unscheduled acute or emergency care admission
- Received a blood draw on arrival or within 24 hours of presenting to hospital
- Had not opted out of electronic health records (EHR) research

At PUH, UHB, and BH, staff members employed by each trust worked onsite to extract and de-identify the data. For data originating from OUH, data extraction and de-identification was performed by Big Data Institute employees at the University of Oxford, within the shared National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre (OH BRC), which spans OUH and the University of Oxford.

The de-identified data was extracted from EHRs, according to the inclusion criteria, as comma-delimited files (.csv files). Local clinicians implementing the study at each participating trust transferred the files to a study folder on the federated clients, using either locally-approved USB flash drives or secure file transfer protocol (sFTP) within the local network.

This study did not require individual informed consent for two reasons. First, the data had been collected within usual medical care. Second, the study investigators did not have access to protected health information or any way to re-identify the data, as it did not include any sensitive or identifying characteristics such as age or ethnicity. During the initial wave of the Covid-19 pandemic, rates of Covid differed amongst age and ethnic groups. At the time this study was conducted, the researchers were concerned that using those variables in the model training would result in a model that predicted the risk of exposure of Covid, rather than the risk of having caught it. Since then, additional research has been conducted to improve fairness and reduce incidences of false negatives/positives when accounting for demographic variables.[20]

[19] Soltan et al. (2022), 'Supplementary Appendix to: A scalable federated learning solution for secondary care using low-cost microcomputing: privacy-preserving development and evaluation of a COVID-19 screening test in UK hospitals'.

[20] For more research on this topic, see Yang et al. (2023), 'Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning'.

Resourcing, implementation and maintenance costs

One key aim of the study was to deploy a system that was inexpensive to produce and easy for local trust staff to use.

NHS hospitals, particularly smaller trusts, run on limited skeleton technical teams with substantial responsibilities.[21] This project aimed to empower local clinical teams so that if one team member was passionate about a project or a cause, they could pursue this relatively autonomously, without needing the hospital's IT team to run the platform, and participate in a federated study. Furthermore, Raspberry Pi devices store data on a removable microSD card. Using the Raspberry Pi also enabled the research team to install the Ubuntu Desktop operating system (22.04.1 LTS), along with the required software and FL pipeline on the devices before sending them to participating organisations. The benefit of this set-up was that local clinicians only needed to follow simple instructions and interact with an easy-to-use graphical user interface to set up the FL and connect to the local network.

The low cost of Raspberry Pi 4's model B (£80-£90 a piece), also allowed the research team to treat them as consumable research equipment that did not need to return. When the study was concluded, local clinical staff were instructed to remove the device's removable microSD card and provide it to their IT team for secure decommissioned equipment disposal.[22] This approach to conducting FL offers a way to potentially reduce the technical and resource burdens associated with this PET, which are often a major deterrent.

Security measures implemented

FL can be considered secure by design because data remains on local devices rather than being centralised, reducing the risk of data breaches. However, FL is not necessarily invulnerable to security risks such as unauthorised on-device access, malicious code injection, or information leakage.[23] To mitigate those risks, researchers included a number of security precautions into the study design:[24]

[21] Ibid.

[22] Ibid.

[23] Zhang (2022), 'Security and Privacy Threats to Federated Learning: Issues, Methods, and Challenges'.

[24] Soltan et al. (2022), 'Supplementary appendix to: A scalable federated learning solution for secondary care using low-cost microcomputing: privacy-preserving development and evaluation of a COVID-19 screening test in UK hospitals.'

- Using single-purpose client and server hardware reduced the risk of inadvertent Trojan attacks or sandbox violations.
- Clients were secured according to local requirements, and participating sites were instructed to physically safeguard devices following local IT processes for hardware containing de-identified data.
- Data on the client was held in de-identified form only for the analysis period, protected by the site network's firewall, and clients were turned off when not in use. Upon completion of participation, sites were instructed to remove and destroy the microSD storage disk.
- Local IT or network security teams established firewall rules, where necessary, to permit two-way communication between the device and the coordinating server via a single, pre-agreed port. The coordinating server adhered to Azure platform security protocols. [25]
- External communication was confined to the pre-agreed port, and the server was deactivated when not used for this study.
- Communications between the client and server consisted solely of weights from the trained model or summary evaluation results, inherently protecting against data leakage if intercepted.
- The model architectures had significantly fewer parameters than the number of training data examples at each site, minimising the risk of raw training data being memorised in model updates.

However, the study did not combine additional privacy-enhancing technologies such as differential privacy in conjunction with FL, but acknowledged the benefit of including such additional measures in future work. [26]

[25] The research team chose Azure because it is an established and widely-used provider within the NHS, and they had easy access to it. The machine on which the platform was used was physically located in the UK.

[26] Soltan et al. (2022), 'Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening'.

Limitations

There were several limitations to this use of full-stack FL.

The FL's full-stack development reduced the need for technical expertise among hospital staff, but the study's design and set-up still required the researchers leading the study to build the models and set up the pipeline from scratch, potentially putting this methodology beyond the skill set of researchers who are non-ML-specialists. The lack of model explainability in FL is a further limitation for this study. Explainability in ML refers to a model's ability to describe, interpret and justify its decisions and behaviour in human terms.[27]

Another limitation is that this FL model was trained using non-independent and non-identically distributed (non-IID) data.[28] In FL, models trained on non-IID tend to perform worse than those trained in the standard centralised learning mode.[29]

Since models are trained on diverse datasets, their learning patterns can vary significantly, making it hard to generate consistent and interpretable explanations across the federated network. In this particular use case, researchers could mitigate that obstacle. Because they had prior knowledge of the data format from the CURIAL-Lab study,[30] they could provide trusts with a data specification and dictionary that local clinicians could follow to carry out preprocessing on clients as part of the study pipeline at each site.[31] This process ensured that the participating sites correctly implemented inclusion/exclusion criteria, feature names were standardised across sites, and data was prepared appropriately for training and evaluation.[32]

---

[27] AWS (2024), 'Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions'.

[28] Zhu., H et al. (2021). 'Federated learning on non-IID data: A survey'.

[29] IID data assumes that all data points are independent of each other and follow the same distribution. Non-independent means that data points are connected or influenced by each other. Non-identically distributed means that data points are not uniformly distributed. For further information on the challenges of using non-IID data in federated learning see Kumarapu (2022), 'Tackling Non-independent and identically distributed data in Federated Learning'.

[30] Soltan, et al. (2022), 'Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening'.

[31] Soltan et al. (2022), 'Supplementary appendix to: 'A scalable federated learning solution for secondary care using low-cost microcomputing: privacy-preserving development and evaluation of a COVID-19 screening test in UK hospitals'.

[32] Soltan et al. (2022). 'Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening'.

Value flows and tangible outcomes

Analysis of the patient data indicated that (1) federation improved the CURIAL-Lab model performance and (2) the resultant global models were generalisable.[33] Prior to this study, the need for specialised technical knowledge at multiple sites to establish and manage a federated client, and the necessity to ensure robust data isolation from active clinical systems, were challenges to the development and deployment of FL systems.

The outcome of the CURIAL-Fed-Lab study indicates that the use of full-stack FL in clinical settings can facilitate hospitals' involvement in AI development at a minimal cost and without needing specialised technical knowledge at each facility.

While no follow-up study was conducted upon the completion of this work, the research was published in the Lancet Digital Health, a high-impact journal publishing innovative work researching the use of digital technology in clinical medicine, public health and global health.[34]

Hands-on federated learning

This report has examined FL and presented a case study that applies it in the context of patient data analysis. However, it can be applied in a variety of other contexts. To illustrate a possible real-world application of FL in a different context, we have partnered with Flower Labs to provide a hands-on low-coding tutorial in which you can build an FL set-up.

Flower[35] is an open-source framework that simplifies and streamlines federated learning tasks. It offers a user-friendly solution for training various models and features ease of use, flexibility, scalability and extensibility.

This tutorial applies the Flower framework to the Adult Census Income dataset to predict income based on attributes including, but not limited to, age, education and occupation. Please access this Google Colab file to complete the tutorial.

[33] Ibid.

[34] The Lancet. (2024). 'About The Lancet Digital Health'

[35] More about Flower Labs and the Flower framework is available here.

Conclusion

This case study presents full-stack FL as a valuable tool in clinical research. It facilitates navigating the privacy challenges associated with handling sensitive patient data.

We feel that breaking down real-world implementations of PETs, as above, into their composite structures and components will increase awareness and understanding of novel PETs. We hope that explainers like this, and the corresponding diagram, will facilitate their wider adoption and contribute to the documentation of case studies.

Get in touch

Please get in touch, if you want to share any feedback, or need help sharing sensitive data. We welcome all forms of input.

Acknowledgements

The writing of this report would not have been possible without the help and insights provided by Andrew Soltan of the University of Oxford, Nic Lane of the University of Cambridge and Flower Labs and Mohammad Naseri at Flower Labs. We are incredibly grateful for their time in answering our questions on their experiences of developing and deploying this use case.

We would also like to thank our colleagues at the ODI and Philpott Design who have contributed their time and assistance to the creation of this written report and our annotated diagram. These include Matt Barber, Emma Thwaites, David Dinnage, Zoe Philpott and Adrian Philpott.