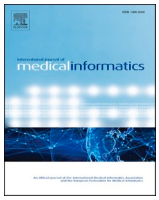
International Journal of Medical Informatics 195 (2025) 105708

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Developing a prototype for federated analysis to enhance privacy and enable trustworthy access to COVID-19 research data

Solmaz Eradat Oskoui [a,d], Matthew Retford [b,*], Eoghan Forde [a], Rodrigo Barnes [a], Karen J Hunter [b], Anne Wozencraft [b], Simon Thompson [c], Chris Orton [c], David Ford [c], Sharon Heys [c], Julie Kennedy [c], Cynthia McNerney [c], Jeffrey Peng [c], Hamed Ghanbariadolat [c], Sarah Rees [c], Rachel H Mulholland [d], Aziz Sheikh [d], David Burgner [e,f], Meredith Brockway [g,h], Meghan B. Azad [i,j], Natalie Rodriguez [i,j], Helga Zoega [k], Sarah J Stock [d], Clara Calvert [d,l], Jessica E Miller [e], Nicole Fiorentino [i], Amy Racine [m], Jonas Haggstrom [m], Neil Postlethwaite [b]

[a] *Aridhia Informatics, UK*
[b] *Health Data Research UK (HDR UK) , UK*
[c] *Swansea University Medical School, UK*
[d] *Usher Institute, University of Edinburgh, Edinburgh, UK*
[e] *Infection, Immunity and Global Health Theme, Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria, Australia*
[f] *Department of Paediatrics, University of Melbourne, Parkville, Victoria, Australia*
[g] *Faculty of Nursing, University of Calgary, Canada*
[h] *Alberta Children's Hospital Research Institute, Calgary, AB, Canada*
[i] *Pediatrics and Child Health, University of Manitoba, Winnipeg, Canada*
[j] *Children's Hospital Research Institute of Manitoba, Winnipeg, Canada*
[k] *Centre of Public Health Sciences, Faculty of Medicine, University of Iceland, Reykjavík, Iceland*
[l] *Department of Population Health, London School of Hygiene and Tropical Medicine, UK*
[m] *Cytel, USA*

ARTICLE INFO

Keywords:
Federated Networks
Federated Analytics
COVID-19
Health Data Research
Privacy-Preserving
Secondary Data
Data Re-use

ABSTRACT

Background: The use of federated networks can reduce the risk of disclosure for sensitive datasets by removing the requirement to physically transfer data. Federated networks support federated analytics, a type of privacy-enhancing technology, enabling trustworthy data analysis without the movement of source data.
Objectives: To set out the methodology used by the International COVID-19 Data Alliance (ICODA) and its partners, the Secure Anonymised Information Linkage (SAIL) Databank and Aridhia Informatics in piloting a federated network infrastructure and consequently testing federated analytics using test data provided from an ICODA project, the International Perinatal Outcome in the Pandemic (iPOP) Study. To share the challenges and benefits of using a federated network infrastructure to enable trustworthy analysis of health-related data from multiple countries and sources.
Results: This project successfully developed a federated network between the SAIL Databank and the ICODA Workbench and piloted the use of federated analysis using aggregate-level model outputs as test data from the iPOP Study, a one-year, multi-country COVID-19 research project. This integration is a first step in implementing the necessary technical, governance and user experiences for future research studies to build upon, including those using individual-level datasets from multiple data nodes.
Conclusions: Creating federated networks requires extensive investment from a data governance, technology, training, resources, timing and funding perspective. For future initiatives, the establishment of a federated network should be built into medium to long term plans to provide researchers with a secure and robust data analysis platform to perform joint multi-site collaboration. Federated networks can unlock the enormous potential of national and international health datasets through enabling collaborative research that addresses critical public health challenges, whilst maintaining privacy and trustworthiness by preventing direct access to the source data.

* Corresponding author.
E-mail address: matthew.retford@hdruk.ac.uk (M. Retford).

https://doi.org/10.1016/j.ijmedinf.2024.105708
Received 25 January 2024; Received in revised form 30 October 2024; Accepted 15 November 2024
Available online 20 November 2024

1. Introduction

The outbreak of COVID-19 highlighted the need for timely access to, and trustworthy sharing of high-quality data to inform and aid decision making in the pandemic [1]. However, challenges such as data silos and privacy constraints can prevent data access for research purposes and preclude the development of more accurate and robust statistical models [2,3]. The use of emerging techniques such as federated analysis can help address these issues. Federated analytics (or analysis) is a form of privacy-enhancing technology that enables trustworthy access to data for analysis queries and machine learning models without the data moving from its source [4]. This technology supports the "Five Safes" framework [5] – Safe People, Safe Projects, Safe Settings, Safe Data and Safe Outputs – by facilitating secure and trusted research access to data. This benefits the research community as it can simplify data governance issues, reduce data silos and provide access to data to enable research that was not previously possible.

The International COVID-19 Data Alliance (ICODA) [6] was convened in July 2020 and, in collaboration with partners and researchers, sought to develop a trustworthy approach to enable researchers in multiple countries to overcome the challenges of accessing and harnessing the power of data to respond effectively to the COVID-19 pandemic [7]. ICODA supported twelve driver projects, each aiming to achieve their own research goals and support development of the ICODA research platform, including the testing of new technology, approaches and services.

Federated analytics and federated learning are supported through the establishment of a federated network, a series of decentralised, interconnected data nodes [8]. Federated networks help address the challenges of data silos and fragmentation by enabling researchers to analyse data from multiple sources without the need to transfer or centralise data. This approach minimises data breach risks and improves the efficiency of research efforts by avoiding redundant data collection and processing. In addition, federated networks for data sharing and analytics in health research can help facilitate rapid and collaborative responses to emerging threats, such as new variants of COVID-19, by enabling real-time international data sharing and analysis [9]. The World Economic Forum outlines an eight-step approach to building federated networks, namely: establishing trust between contributing institutions; jointly determining the problem a federated approach can solve; aligning incentives; defining resources; identifying institutional gaps; creating a governance model; structuring the data; and deploying the technology [10,11]. Creating federated networks between institutions initially requires a substantial amount of investment from funding, data standards, data governance, technology, training and resource perspectives. An established federated network also requires a robust infrastructure for the operational and maintenance life cycle, cybersecurity, enforcement of governance and standards, and data management [8].

There are many examples of projects that have implemented and used federated networks within the health research and care sector, as discussed in Section 2.2. While unlikely to be a suitable solution for all research projects, setting up trusted and transparent federated networks between institutions enables the potential to access large volumes of health-relevant data on a global scale for analytics, through offering a trusted, interoperable and secure data access approach [12].

This paper outlines the technical details and methodology of developing a prototype federated network and piloting the use of federated analysis, using test data from a COVID-19 research project. This work was a collaborative effort by ICODA, Aridhia Informatics [13], Secure Anonymised Information Linkage (SAIL) Databank [14] and researchers from the International Perinatal Outcome in the Pandemic (iPOP) Study [15,16]. An overview of the partners and components used in ICODA federated network is provided under Appendix A. This paper also highlights other examples of federated analytics and presents some of the benefits and challenges of this approach.

1.1. Project aim

The primary aim of this project was to develop, implement and test a federated network to support federated data analysis across multiple repositories. This involved:

- Developing a technical solution to integrate a Trusted Research Environment (TRE) with one or more data repositories, enabling the exchange of federated analysis tasks and retrieval of screened aggregate results
- Implementing the solution within the ICODA Workbench (TRE) and the SAIL Databank (data repository)
- Testing the full integration and workflow of the solution using aggregate-level data from the iPOP Study which consisted of aggregate level (non-individual level) birth data sourced from multiple countries. The iPOP Study aimed to investigate the impact of the pandemic on preterm births and other birth outcomes.

The wider aim of this work was to demonstrate the potential of federated approaches to securely analyse sensitive health data from multiple repositories across countries and regions, providing critical insights for health research.

1.2. Success criteria

The key success criteria for the project included:

- Installation of the federated solution within both the TRE and data repository
- User acceptance testing of each step in the federated solution, as illustrated in Fig. 4
- Successful execution of federated analysis queries based on the iPOP Study test data and use case

Meeting these success criteria involved both technical and non-technical actions, such as:

- Developing a tailored technical solution to support the requirements of the iPOP Study's use case
- Facilitating training for researchers and staff on how to use the federated data analysis solution
- Managing data governance and information security considerations for the data used

These are expanded on in Table 3.

2. Methods

These methods outline the technical process of developing and implementing a federated network prototype and subsequently testing federated analytics, which is the practice of applying basic data science methods to decentralised data node(s) and returning aggregate level analysis results [12]. This approach accelerates health research insights by enabling large-scale data analysis across multiple repository sites, while reducing the risk of disclosure for sensitive datasets. By allowing data to remain in situ, it provides researchers access to data that otherwise may be subject to access constraints.

2.1. Data use-case and requirements

The iPOP Study was one of the first ICODA driver projects and sought to catalyse rapid discoveries about preterm birth, stillbirth and perinatal health during the COVID-19 pandemic [15,16]. It used existing, aggregate level birth data from 26 countries, collected at national, regional or facility level. The year-long project aimed to determine and compare changes in preterm birth outcomes during COVID-19 lockdowns, with

future plans aimed at using individual level data from different countries. As this study could potentially seek to set up distributed data nodes in multiple countries that would be subject to their own data governance and data privacy laws, it was seen as a use case well suited to pilot the technical implementation of a federated network and demonstrate the potential for federated analysis within this network. Successful testing of this solution would then enable this approach to be used for the iPOP Study's immediate and longer-term data analysis and allow other ICODA and related health research studies to set up federated networks to access data for their projects, perform federated analysis and accelerate research insights.

To overcome challenges in accessing and analysing data from many countries and facilities, the iPOP Study focused heavily on data contribution agreements (DCAs). It provided a comprehensive protocol that included outcome definitions and data collection templates for data contributors to populate [15,16]. This further strengthened the choice of using the iPOP Study's data to test federated analysis, as the structure and utilisation of the data was standardised, minimising heterogeneity and bias within the study. Data contributors from around the world securely uploaded their data directly into SAIL Databank via a secure link. Data collection, quality assurance, curation, standardisation, and modelling were performed in the SAIL Databank. An interrupted time series model was applied to the data. The model outputs were aggregated, and stored in a database within the SAIL Databank, providing the test data for piloting federated analysis for this project.

The SAIL Databank acted as a data node in the federated network, connected to the ICODA Workbench, a TRE used to perform federated analysis. Whilst the SAIL Databank held the aggregate data for all the countries involved in the iPOP Study, a key advantage of federated networks is the ability to connect to multiple data nodes hosted remotely, removing the need to move and store the data centrally. For this solution, a user was first authenticated using their credentials by the SAIL Databank. Secure, authenticated application programming interface (API) requests are used to transfer the analysis tasks from the ICODA Workbench to SAIL Databank, and to retrieve results. Functionality was built to allow results to be quarantined and reviewed in the SAIL Databank prior to the release and subsequent access in the ICODA Workbench [13]. This trustworthy approach prevented any source data leaving the SAIL Databank, or researchers having direct access to the data. Additionally, the transferred results were used to test a *meta*-analysis tool developed by Cytel [17] hosted on the ICODA workbench for the iPOP Study. While the federated network was established, federated analytics was not used in the final analysis of the iPOP Study itself, due to challenges in areas of data availability, access to technology, training, project funding and duration. These factors are detailed in 4. Discussion. However, the project demonstrated the potential of implementing a federated network and developed capabilities for future research studies.

2.2. Statistical methods and examples of federated analysis networks

Federated networks enable researchers to request dataset details at the metadata level, and obtain analysed, aggregated and approved results from the source data, which itself can be aggregated or individual-level data. A researcher has the flexibility to request results at a descriptive statistics level (e.g. means, frequencies, standard deviations) to gain an overview of the dataset's characteristics or apply inferential statistical methods (e.g. *t*-test, analysis of variance, regression models, time-series analysis) to explore relationships, test hypotheses, and make predictions using the data. The results from each data node can be further analysed to gain an understanding of the differences between the results of each dataset. Furthermore, federated networks can be expanded to enable federated learning whereby researchers can incorporate remote datasets into training runs, improving model accuracy without compromising data privacy [18]. A summary of different federation projects and networks are provided in Table 1.

2.3. Technical development

The federated network architecture detailed in this paper was developed to support the iPOP Study and wider ICODA initiative, enabling federated analysis of data from many countries through remote queries and computation. Our solution was based on the infrastructure and concepts developed by the Alzheimer's Disease (AD) Data Initiative project, as outlined in Appendix A.

Some key features of the federated network solution are:

- Ability for data controllers to define different levels of federated analysis access
- Ability for researchers to see the structure of the dataset, through federated metadata queries
- Analysis queries are automatically run on the remote dataset and do not require manual intervention
- Option for federated results to be quarantined for manual release approval dependent on the requirements of the data access agreements and output checking policies of the data owners

A data node represents data repositories at health facilities, research institutions, or individual devices. Contributing data to research projects may be difficult, due to data governance or technical constraints. The solution piloted for the AD Data Initiative project allowed data contributors the flexibility in choosing how their data can be used and provided a framework for standardised access through the development of a common API for federated data sharing [37]. During the technical development from the AD Data Initiative project, three levels of data sharing were proposed, defined in Table 2 and shown in Fig. 1:

The federated network architecture discussed in this paper uses level 2 data sharing, as shown in Fig. 2.

For this solution, a Docker container [38] containing the analysis script is first uploaded to the data node and stored in a registry. A federated analysis task is then sent to the data node using the Common API. The federated task is constructed of a data selection query which details the data required for the analysis, and a reference to the Docker container which contains the analysis script. A process is run within the container to perform the analysis. The container is provisioned with an input folder containing a read-only, temporary copy of the extracted data selection. The container can only write to a temporary output folder. Once the container has been run, the output folder may move to a quarantine review step before being released to the user. When the output is approved, the approved results are available to be retrieved via the Common API.

2.4. Implementation

Security is a primary concern in developing a federated network. Sites providing federated access must ensure that: data are only accessed securely, by approved researchers and for the approved and intended purpose; that executing third party containers does no harm to their systems; and no disclosive output results are shared. Additionally, researchers may have a concern that their analysis code remains confidential. To address these concerns, the following measures were implemented during this pilot:

- The source of the container was restricted to specific registries in accordance with any related governance or project restrictions
- Several approved base images were provided for common frameworks (R, Python) in the approved container registry accessible within the ICODA Workbench
- Users could only build containers from these base images
- Federated data nodes can specify what container registries they accept: SAIL Databank was given read-only rights to the container registry operated by the ICODA Workbench

Table 1
Examples of projects and networks using federated analysis.

Example	Description
Open-source initiatives from the Observational Health Data Sciences and Informatics (OHDSI)	OHDSI have created an Rshiny application called Atlas [19] which allows researchers to easily select statistical model settings (e.g. Lasso Logistic Regression, Random Forest, Gradient Boost machine, Ada boost) as part of their analysis plan. The open-source project DataShield [20] has also developed several R packages with a number of statistical modelling functions based on Generalised Linear Models on data from single to multiple sources for in the area of pharmacoepidemiology
Examples of initiatives that have implemented federated networks within that provide trustworthy data access for health research and patient care	These include the Global Alliance for Genomics and Health (GA4GH) [21]; Alzheimer's Disease Data Initiative (AD Data Initiative) [22]; Norway's precision medicine initiative, BigMed [23]; Canadian Distributed Infrastructure for Genomics (CanDIG) [24]; Australian Genomics [25]; Autism Sharing Initiative [26]; European-Canadian Cancer Network (EUCANcan) [27]; German Medical Informatics Initiative [28]; Patient-Centered Outcomes Research Network (PCORnet) [29]; and European Health Data & Evidence Network (EHDEN) [30].
Examples of initiatives that have focused primarily on large homogeneous units of federation, such as at the level of healthcare systems.	The TriNetX system allows users to conduct customised search queries of over 100 million electronic health records [31]. In the SCAlable National Network for Effectiveness Research (SCANNER) platform, existing health research datasets were made more FAIR (Findable, Accessible, Interoperable, and Reusable) and a federated machine learning architecture was applied on top of the FAIRified datasets from different health research performing organisations [32]
Efforts in adopting Common Data Models (CDM)	These include the Observational Medical Outcomes Partnership (OMOP) CDM and, Patient-Centered Outcomes Research Network (PCORnet) CDM, For example the COVID – Curated and Open Analysis and Research Platform (CO-CONNECT) project uses data standardised to the OMOP CDM format thus enabling data discovery through federated requests to many data repositories in the UK [33]. Further community efforts by the OHDSI initiative provide tools and infrastructure for distributed data sharing. that allows for federated query functionality and enables healthcare institutions with different policies and operating under different state laws to permit federated access to data.
Other examples of federated approaches	As a federated data network, The Data Analysis and Real World Interrogation Network (DARWIN EU) standardised anonymised real-world data using the OMOP CDM to ensure consistency across the data sources. Data partners also use standardised analytical methods provided by DARWIN EU to examine their data locally, enabling the rapid conduction of larger, multi-database studies [34]. Other examples of federated approaches include the TRE-FX [35] and Teleport [36] projects which have developed solutions that connect TREs to enable federated analysis of health data for research.

- Additional network controls were in place to restrict communication from the ICODA Workbench and the federated node at SAIL Databank. All traffic was encrypted between components: SAIL Databank used Singularity to further secure the execution of containers in their environment [39]
- To identify authorised users, project-specific access tokens were generated by a SAIL Databank service for approved researchers which were then used to submit the analysis tasks through the Common API. These access tokens determined what tables and variables were available to this researcher at the time the analysis was performed, and the system ensured no other data was accessed outside of what was agreed upon by the governance controls in place for the specific project
- Within SAIL Databank, a query engine was developed that implemented the agreed Common API for federated analysis
- In addition, a test node was deployed on the ICODA Workbench, based on the reference implementation of the Common API developed by AD Data Initiative [37]. This allowed users to train in the use of the protocol independently of SAIL Databank, using dummy data.

The Common API is a constrained version of the GA4GH Task Execution Service [40] and is made up of three parts – metadata browsing, remote data selection and federated computation, see Example of task API for more details. The SAIL Databank implementation provided external access to two API parts based on the access permission granted:

- All authorised users had access to the metadata API and could remotely query metadata

Table 2
Three levels of data sharing.

Level of data sharing	Description
Level 0 – Centralised data access:	Data and metadata are directly transferred for hosting to a TRE, and researchers access data directly within the TRE
Level 1 – Distributed data access:	Data and metadata are hosted in distributed data nodes or repositories. Researchers can query the metadata and data remotely from the TRE and results or full copies of the data can be retrieved. This manner of querying the data from the TRE is made possible through a direct query to the data node hosting the database.
Level 2 – Federated data analysis:	Data and metadata are hosted in distributed data nodes or repositories. Researchers can query the metadata remotely from the TRE. Researchers can send federated tasks (made up from a selection query and a containerised script) to be executed within the node where the data resides. Based on the agreed setup by the data providers, the results may require a manual review process at the quarantine step or can be configured to be automatically approved by using secure containers. Access, transfers or copies of the source data itself directly to the TRE are not permitted, only approved aggregate results can be transferred to a TRE.

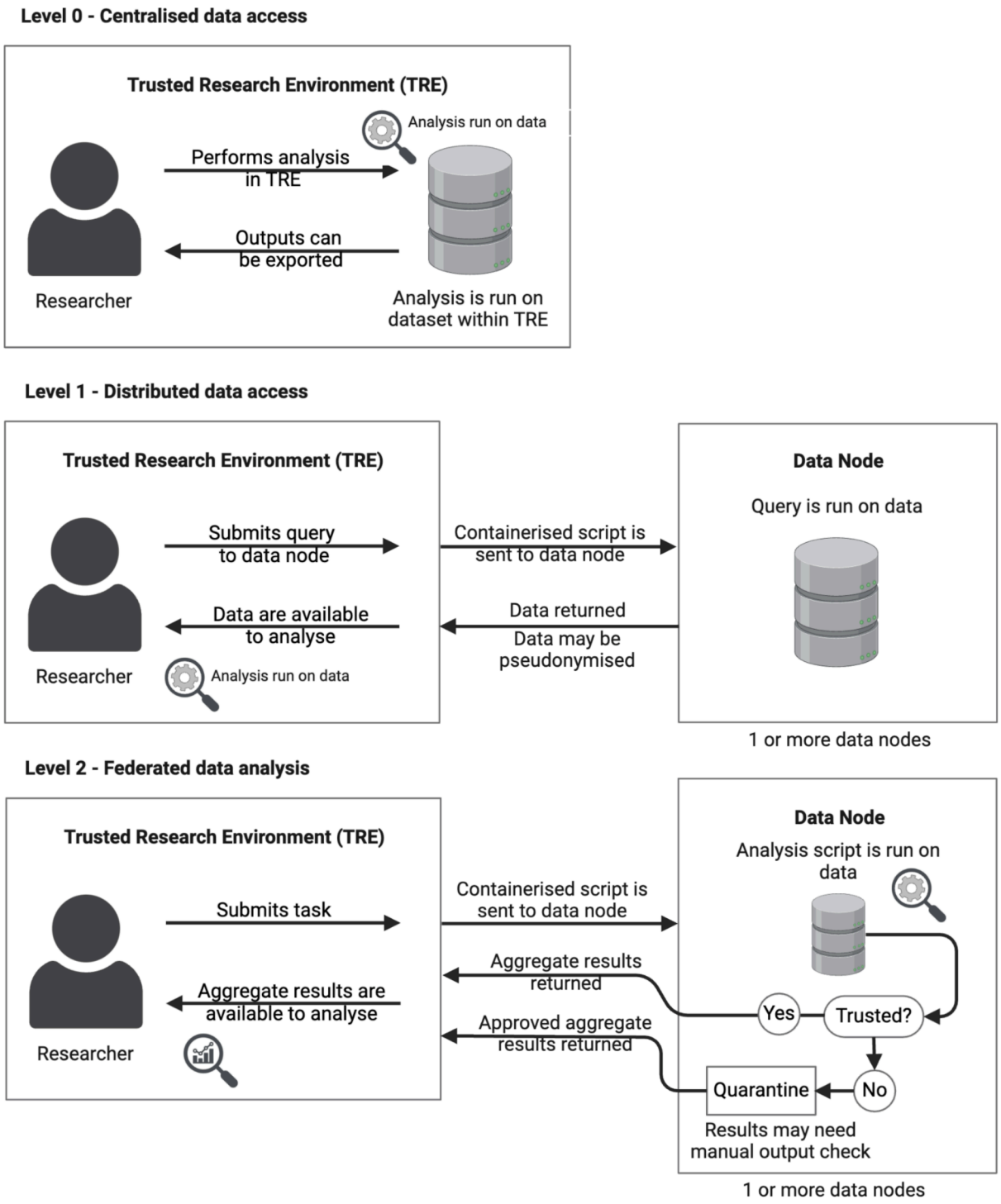


Fig. 1. *Three different modes of data sharing.* **Level 0: Centralised data access** – Users perform analysis within the TRE, where data never leaves the secure environment, but outputs can be exported. **Level 1: Distributed data access** – Users submit queries to a remote data node, where data are analysed, and full results are returned to the TRE for further analysis. **Level 2: Federated data analysis** – Users submit tasks to a remote data node, where analysis scripts are run without direct access to source data. Only aggregate results are returned to the TRE, and outputs undergo a manual or automated check for trustworthiness before release.

- Users with Level 2 permissions to a dataset had access to the remote compute API but not the selection API. The actual data stored in the databank's database is never transferred, only the results of the executed, containerised script
- Manual output checking was required for level 2 outputs prior to release to the requesting user, who could then download the results.

3. Results

Following the technical development, test data from the iPOP Study was used to successfully test the federated network between the ICODA Workbench and the SAIL Databank. This work has demonstrated how federated networks can be used to enable secure access to health data for research, even when timescales are short. It has provided materials and lessons learned on how this approach can be used by other research teams.

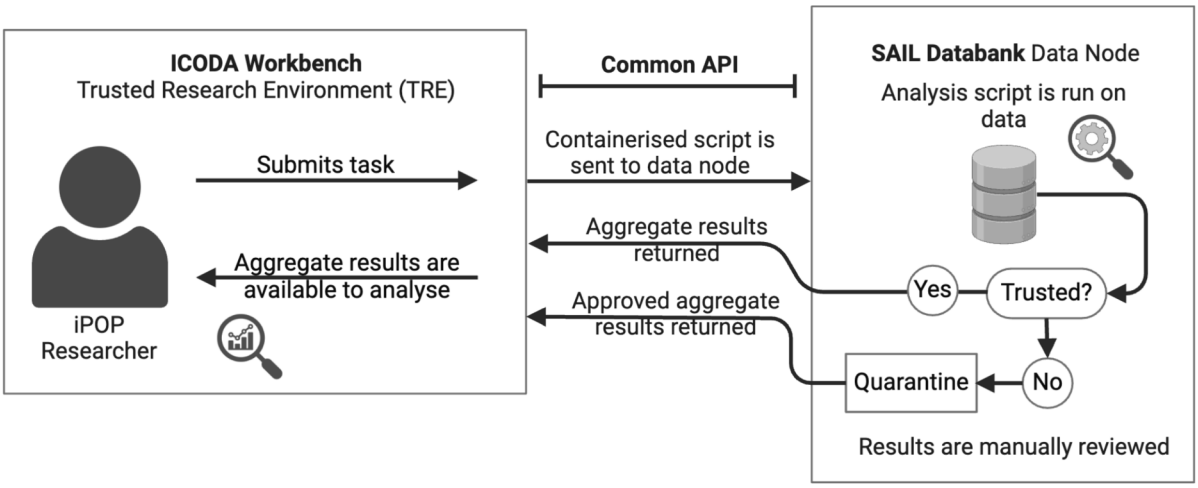


Fig. 2. High-level data-sharing infrastructure piloted during the project.

3.1. User training

To facilitate federated analytics, hands-on training was provided to guide users on the technical architecture and execution steps. [41]. The common-api-examples GitHub repository [42] provides worked examples of how to use the Common API [43]. For simplicity, the word "script" was used however this could refer to any programme that can be run in a Docker container, including complex programmes with encapsulated library or package dependencies. The examples demonstrate a three-step approach shown in Fig. 3:

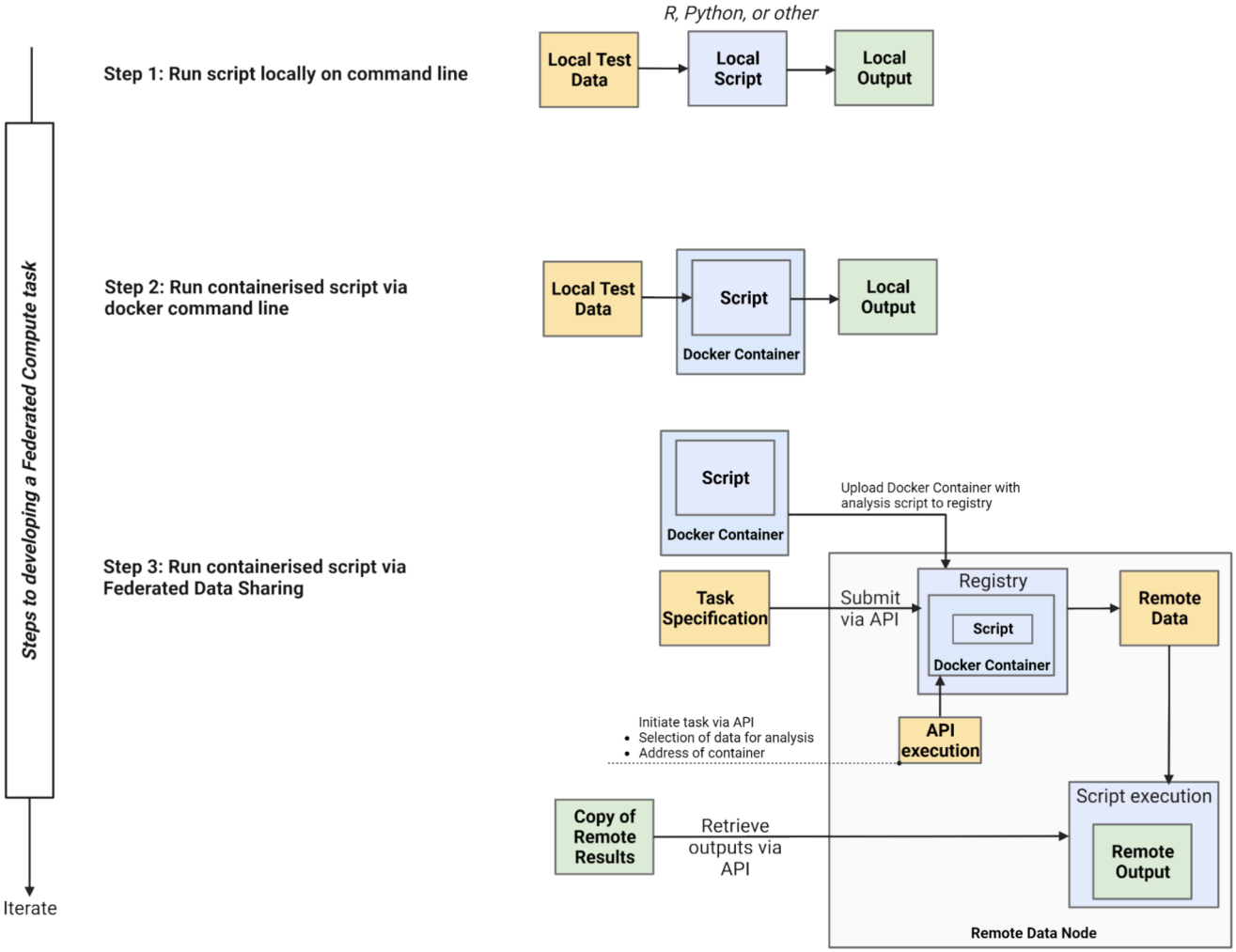


Fig. 3. Schematic outlining the three steps to develop a federated compute task.

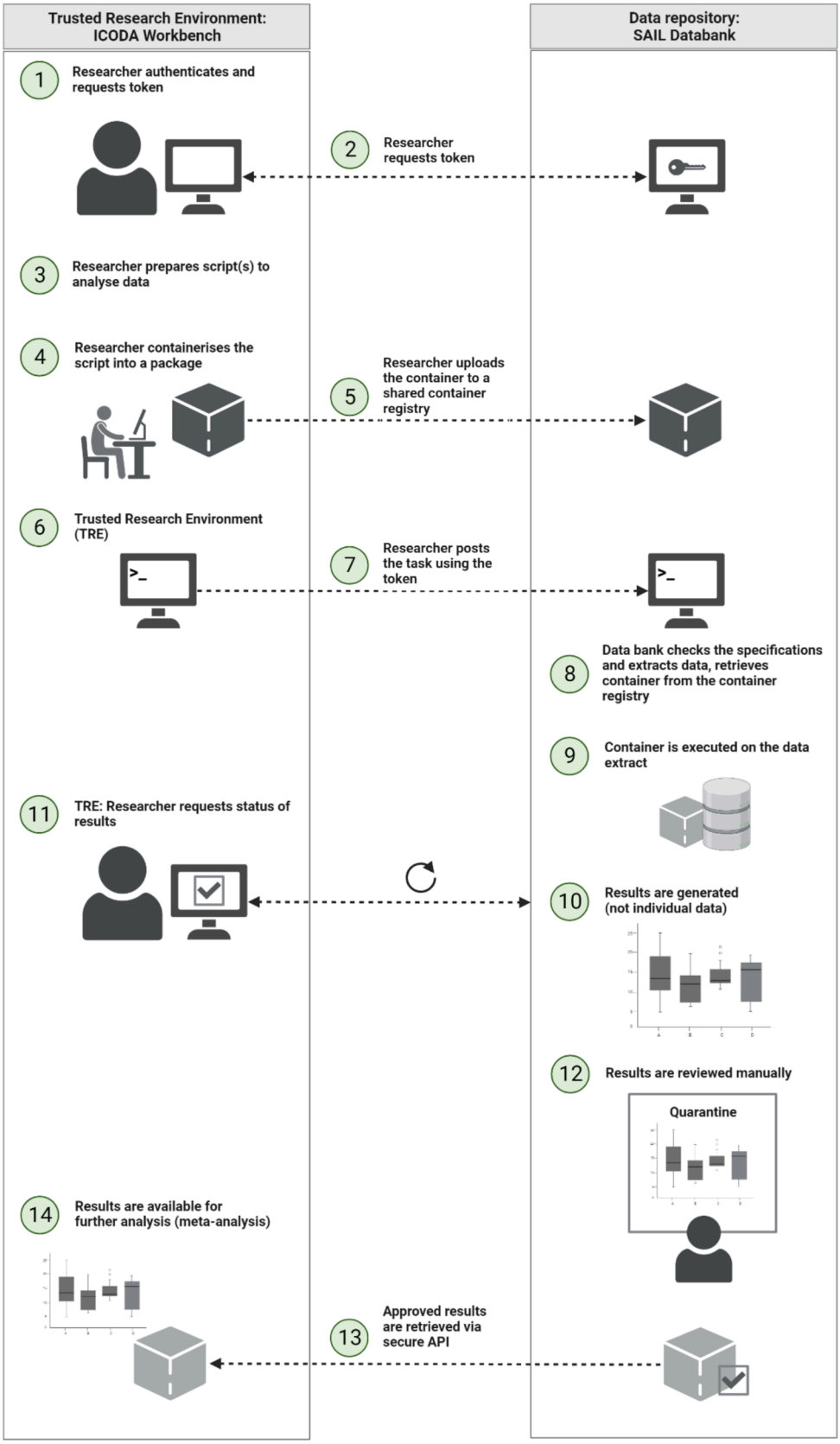


Fig. 4. Schematic of steps listed in the order of executions between the research environment and the databank.

Table 3
List of challenges encountered during implementation of a federated network for ICODA, potential enablers to help address these and responsibilities of team members involved.

Categories	Challenges	Potential Enablers	Delivery Team Members
Data standards	An additional curation step to standardise some datasets was required, despite having a clear data dictionary. Due to the global nature of the project, challenges such as interpretation of data requirements and dialogue with data providers was needed.	Maximise data harmonisation and minimise heterogeneity and bias in the data by providing clear documentation in different languages. Setting up data drop-in sessions to support data providers with an understanding of the data needed for the project. Adding metadata and data dictionaries for source datasets to data discovery services such as FAIR Data services [52] and the HDR Gateway [53], allowing researchers to gain an understanding of the underlying data. Facilitate data interoperability and reusability for international projects.	The iPOP leadership team designed the iPOP Study research project. [15,16] The analysis team designed the standardised data templates, data onboarding, ingestion, curation, quality assurance, and analysis of data within the SAIL TRE. Furthermore, they worked with the SAIL Databank team and technical partner Aridhia to create the agreed test data structure for the federated data node.
Data governance	Questions were received from data contributors requesting clarification on wording and terminology in the DCA. This was often due to English not being a first language for many contributors and may have contributed to a delay in data onboarding.	Translation of key documentation such as project mission, and DCA from English to support data contributors, and provide forums or drop-in sessions to provide clarity. Providing a FAQ page on terminology and DCA to support data contribution onboarding	ICODA and SAIL Databank information governance experts worked together with the iPOP leadership team to create multiple template DCA for both centralised and fully federated models.
Technology	General understanding of federated analysis approaches, and the training and tools required to support this. Challenges, complexity and effort associated with integrating multiple systems. Wait time due to manual output checks for disclosive data by SAIL Databank meant the need to regularly check back to see if the step had been performed.	Further development to user-interfaces and tooling making the system more user friendly or an existing orchestration framework could be layered on the Common API. Having the Common API reduced the complexity of discussions and design. SAIL Databank was able to independently implement the API on their existing infrastructure To improve efficiency of task review in the manual approval step, the use of notification services could be introduced.	Aridhia led the technical design of federated protocol and APIs, deployment planning and testing, support and training of researchers in federated analysis ahead of integration testing. SAIL Databank team members provided leadership and expertise in federated analytics and use of TREs for data science in addition to implementation of a security token system which provided security for analysis request submissions and results checking.
Skills	Learning needs for team members involved in the development of the analysis plan in concepts and technologies for federated analytics. Tooling for end-users was a challenge because the Common API assumed a level of programming competency. This required some helper scripts to be developed in R and Python. Technical knowledge and training using the Common API, tokens, docker, command prompts.	Hands on learning sessions and worked examples from technology partners on the background components which are needed to carry out federated analytics within a network. Developing user interfaces for abstraction, guides, training modules is recommended. This would require some helper scripts to be developed in R and Python	Aridhia conducted learning workshops with members of the analysis team in overview of federated analysis, script containerisation, construction of task execution specification and providing user guides with steps to send and retrieving a federated query from the ICODA workbench.
Project duration	The short project duration (one-year) did not provide sufficient time to test the implemented federated network with individual-level data or extend to other federated data nodes.	Assessment of whether setting up federated networks would fit a research project's short- and longer-term objectives. Implementation of federated networks may be best considered for projects with a longer duration.	

1. The scripts are run on dummy data through the command line on the user's local machine;
2. The scripts are packaged in a Docker container and run via the command line on the user's local machine;
3. The Docker container is uploaded to a remote registry in a federated network. The federated task, made up of the selection query and containerised script, is sent from the secure TRE to the remote data node. A copy of the results can be retrieved. The user can iterate on these steps and updates their scripts to improve outputs.

The training process tested federated approach assumptions. While Docker containerisation is widely known, its practical use was less common, necessitating enhancements to the user guide and examples. The Common API was targeted at users with programming skills, the training highlighted the need for a desktop or browser-based user interface. A prototype of helper functions in R and Python was developed to simplify running a federated analysis. For example, the Python functions could be used from within a Jupyter notebook, removing the need to manage low-level network requests [43].

3.2. Integration testing

The integration testing for completing a federated task between the ICODA Workbench and SAIL Databank involved multiple sequential steps shown in Fig. 4.

Overview of steps required to perform federated analytics:

1. Researcher authenticates and requests a token from the research environment
2. Researcher requests a project-specific access token
3. Researcher prepares script(s) to analyse dataset(s)
4. Researcher containerises the scripts into a package
5. Researcher uploads the container to a shared registry
6. Researcher prepares a task execution specification, referencing the container at the registry
7. Researcher posts the task using the project-specific access token from the research environment
8. Databank checks the task specification, extracts required data, and retrieves containerised scripts from the container registry
9. Container is executed on the data extract
10. Researcher requests the status of the computation and quarantine from the TRE
11. Results are generated
12. Results are manually reviewed during quarantine and approved or rejected by the databank
13. Approved results are retrieved via the secure API
14. Results can be downloaded and unzipped in the TRE ready to be further analysed and reviewed by the researcher.

4. Discussion

This project is the first example for successfully implementing a federated network for the ICODA initiative. The co-development and collaboration between the technology partners Aridhia, ICODA and SAIL Databank allowed for a successful pilot for the integration of federated analytics between the ICODA Workbench and SAIL Databank. Ultimately, the iPOP study did not use the federated data analysis network developed in this project for its final research analysis. This decision was primarily due to the short timeframes of the study and the parallel development of the federated solution, which could have delayed the research team's analysis and publications. Additionally, all the aggregate-level data required for the study was collected and stored within a single data node, the SAIL Databank. This allowed for direct analysis of all data within that node, corresponding to Level 0 in Fig. 1. However, the development of the prototype for federated analysis and the subsequent integration testing which made use of aggregate-level test data from the iPOP Study represented the first step in implementing the necessary technical and user experience functionalities, which could be expanded upon for future studies that use individual-level datasets from multiple data nodes.

The approach piloted in this paper is already enabling a wider range of health research projects, such as PHEMS, an international consortium of paediatric hospitals developing federated nodes based on the Common APIs open standard for federated data sharing. This technology will provide an effective, secure, and trustworthy way to access data for analysis purposes by keeping data hosted locally [44], building on the prototype solution described here.

Another project building on this work is the AD Data Initiative, where additional funding has enabled the development of the Federated Data Sharing Appliance [45]. This tool enhances the user experience by providing a graphical user interface for the Common API and federated analytics, moving beyond the command-line interface piloted in this paper.

4.1. Challenges and potential enablers

Setting up federated networks to address the challenges of linking siloed health datasets comes with many obstacles. The paper "Federated Networks for Distributed Analysis in Health Data" outlines potential challenges and enablers in categories such as cultural and organisational, technological, data standards, legal and regulatory, knowledge and competence, ethical and social, and financial and political [8]. The set of challenges, potential enablers encountered, and responsibilities taken by team members involved in the development and implementation of the ICODA federated network are outlined in Table 3.

5. Conclusion

This paper highlights the successful development of a federated network between the ICODA Workbench and SAIL Databank using the Common API and integration testing for federated analytics, focused on aggregate level test data from the iPOP Study. Guided by the principles of open science and the "Five Safes" framework, a successful implementation of a federated network was achieved through a collaborative effort between ICODA and its partners. This proof of concept highlights the potential of federated approaches in enabling secure access to health data for research both within and beyond the ICODA initiative.

The benefits of setting up a federated network include:

- Simplifying data governance processes and enabling safe access to research data that might not have been previously accessible
- Providing data custodians with greater control and oversight over their data, and reducing risk of unauthorised data access, as data always remains within the data custodian's system and is never accessed directly by researchers
- Demonstrating responsible and trustworthy management of sensitive data through the application of federated networks
- Enabling researchers access to data from multiple sources without needing to transfer or centralise data, ensuring analysis tasks are always run on the latest version of the source data
- Allowing researchers to analyse large datasets without transferring or storing source data
- Minimising costs and accelerating progress by removing the need for data custodians to transfer large datasets to researchers for analysis

Initial Implementation of a federated network requires an extensive investment from a funding, data standards, data governance, technology, training and resource perspective. However, over time, this is likely to reduce as technology and techniques are more broadly adopted and skills in this area increase. For health research to have global scale, reach and impact, there is a need to access national and international datasets, and federated networks provide a solution to access data which cannot

physically be brought together, owing to governance, security or practical challenges, such as size of the datasets.

The development and testing of this prototype has demonstrated the importance of technical infrastructure, as well as the information governance, standards and resources needed to set up a federated network. Once established, focus should also be given to the operation and maintenance of infrastructure, security and enforcement of governance and standards, and data management [13]. For future data scalability and to provide researchers with a secure and robust data analysis platform to perform joint multi-site collaborations, the establishment and management of federated networks should be considered in medium to long term planning.

Federated networks have an enormous potential in bringing together national and international health care datasets that may not be accessible due to cross-border governance and security reasons, providing a solution to data silos and aiding the collaborative research effort within healthcare and health research sectors to address major public health challenges.

6. Authors' contributions

Members from ICODA, Aridhia, SAIL Databank, iPOP and Cytel worked collaboratively during this one-year international project to support the successful implementation of the federated network and integration testing of a federated analytics. Specifically:

ICODA members – N.P., M.R., K.J.H., A.W. and R.B. provided project management, project funding and organisational support, coordination of technical partnerships for delivery, and working with the iPOP Study research team and data contributors assisting with information governance and data contribution.

Aridhia members – R.B. coordinated and led the technical design of federated protocol and APIs, deployment planning and testing, support and training of researchers in federated analysis ahead of integration testing.

SAIL members – S.T. and D.F. provided leadership and expertise in federated analytics and use of TREs for data science. S.T. designed and led the implementation of the federated analytics platform based at SAIL, with J.P. and H.G. providing the operational technical development of the platform. S.H. provided legal expertise and drafted and negotiated all data sharing agreements with iPOP data providers and provided expertise to support the development of ICODA governance structures. J.K. and C.M. provided support in developing governance structures for the iPOP data and the use of SAIL Databank as host of the data for the project. S.R. provided analytical support in providing data from Wales, hosted by SAIL Databank, for inclusion in the scientific study outputs and provided support for the instantiation of the project through governance review and data ingress validation. C.O. provided project management and coordination of SAIL/SeRP's role within the project and fed into governance structure development and data acquisition.

iPOP members – A.S. D.B., M.B., M.B.A., N.R., H.Z., and S.J.S., representing the iPOP leadership team designed the iPOP research project. S.E.O., R.M., C.C., and J.E.M., representing the analysis team, designed the data ingestion, quality assurance, curation and analysis of data within the SAIL TRE. S.E.O. performed the integration testing for the federated network. N.F. was the project coordinator for the study.

Cytel members – J.H., A.R. designed and developed the *meta*-analysis tool hosted on the ICODA Workbench for the iPOP Study. A.R. was also part of the iPOP analysis teams and developed the coding for the interrupted time series model for the iPOP project.

S.E.O. drafted the manuscript as an employee of Aridhia Informatics Ltd. M.R. and E.F. created Fig. 1, Fig. 2, Fig. 3. and Fig. 4 and supported the redrafting of the manuscript.

S.E.O., M.R., N.P., E.F., K.J.H, R.B. revised and reviewed subsequent versions of the manuscript.

Summary table.

What was already known on the topic?	• Data sharing in health research remains a major challenge. • Striking a balance between rigorous data governance processes, and data access has been critical in the creation of federated networks. • Federated networks support federated analytics, which is a type of privacy-enhancing technology enabling trustworthy data access and analysis
What did this study add to our knowledge?	• Demonstrated the successful creation of a federated network and piloted federated analysis for a COVID-19 health research project. • Source data harmonisation and standardisation, technical training and upskilling researchers in benefits and uses of federated data sharing networks should be a focus of a project adopting federated networks. • Assessment of whether setting up federated networks would fit a research project's short- and long-term objectives should be considered and may provide the right solution for long term research projects.

CRediT authorship contribution statement

Solmaz Eradat Oskoui: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation. **Matthew Retford:** Writing – review & editing, Visualization. **Eoghan Forde:** Writing – review & editing, Visualization. **Rodrigo Barnes:** Supervision, Methodology, Investigation, Conceptualization. **Karen J Hunter:** Writing – review & editing, Project administration. **Anne Wozencraft:** Writing – review & editing. **Simon Thompson:** Methodology, Conceptualization. **Chris Orton:** Validation, Methodology, Investigation, Conceptualization. **David Ford:** Methodology, Conceptualization. **Sharon Heys:** Project administration. **Julie Kennedy:** Project administration. **Cynthia McNerney:** Project administration. **Jeffrey Peng:** Validation, Investigation. **Hamed Ghanbariadolat:** Validation, Investigation. **Sarah Rees:** Investigation, Data curation. **Rachel H Mulholland:** Data curation. **Aziz Sheikh:** Supervision. **David Burgner:** Methodology, Conceptualization. **Meredith Brockway:** Supervision. **Meghan B. Azad:** Supervision. **Natalie Rodriguez:** Supervision. **Helga Zoega:** Supervision. **Sarah J Stock:** Investigation, Data curation. **Clara Calvert:** Supervision. **Jessica E Miller:** Data curation. **Nicole Fiorentino:** Project administration. **Amy Racine:** Resources. **Jonas Haggstrom:** Resources. **Neil Postlethwaite:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by International COVID-19 Data Alliance (ICODA), an initiative funded by the COVID-19 Therapeutics Accelerator and convened by Health Data Research UK (HDR UK). We acknowledge funding via the COVID-19 Therapeutics Accelerator from the Bill & Melinda Gates Foundation (INV-017293), and the Minderoo Foundation (INV-017293) and support from Microsoft's AI for Good Research Lab. Aridhia Informatics Ltd was funded by the Bill & Melinda Gates Foundation (INV-021793). Cloud hosting support was provided by Microsoft AI for Health. SAIL Databank and the Secure eResearch Platform (SeRP) UK, based at Swansea University, were funded by an award from Health Data Research UK (2020.112), supported by funds from the ICODA initiative, to develop the underlying infrastructure and providing

expertise in establishing the federated analytics platform and governance models. This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge the iPOP data providers who made their anonymised data available for research [15]. This work used data collected on behalf of patients as part of their care and support. This project was approved by the SAIL Information Governance Review Panel, under project numbers 1292 and 1299. Helga Zoega was supported by a UNSW Scientia Program Award during the conduct of this study. Sarah J Stock was funded by a Wellcome Trust Clinical Career Development Fellowship (209560/Z/17/Z). Meghan B. Azad is supported by a Canada Research Chair in the Developmental Origins of Chronic Disease. All authors approved the version of the manuscript to be published. This publication is based on research funded in part by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

Appendix A. Supplementary material

Outlined below is the background to each partner and components required to establish a federated network for ICODA to perform federated analytics:

Background on international COVID-19 data Alliance (ICODA)

ICODA was set up as a global collaborative effort to unite international health research data to both enable discoveries and empower researchers to access health data from around the world to address key research questions to tackle the COVID-19 pandemic. Convened by Health Data Research UK (HDR UK) [46], ICODA has adopted the Office of National Statistics (ONS) "Five Safes'' framework [47], encompassing Safe People, Safe Projects, Safe Settings, Safe Data and Safe Outputs. The ICODA Workbench, is a Trusted Research Environment (TRE) and data repository that includes a wide range of analytical tools and fosters collaborations between researchers and data scientists, to support high quality science. ICODA has developed a set of trustworthy information governance processes through its exemplar projects, of which the International Perinatal Outcomes in the Pandemic (iPOP) Study is one.

Background on international Perinatal Outcomes in the pandemic (iPOP) Study

The iPOP Study [15,16] aimed to provide new insight on the impact of the first COVID-19 pandemic lockdown on preterm birth and stillbirth rates. Over 26 countries, representing data from 52 million births occurring in the 5 years preceding and up to 4 months after the start of the first wave of the COVID-19 pandemic, provided aggregate level perinatal and birth data at a national, regional and hospital level. These data were stored within the SAIL Databank, and a comprehensive interrupted time series and *meta*-analysis was conducted within the SAIL TRE. Alongside the research objectives, this project was developed to test and pilot the creation of a federated network to demonstrate the potential for federated analytics using the aggregated data provided for the study. While the federated network was established and tested, federated analytics was not used in the analysis of the iPOP Study for the reasons discussed in 4. Discussion.

Background on secure anonymised information linkage (SAIL) Databank

As part of the data contributor agreement (DCA) for the iPOP Study, data were stored in the SAIL Databank. SAIL Databank [9] is a TRE, and robust secure data storage facility used for anonymised person-based data for research to improve health, wellbeing and services. Primarily the national data safe haven for de-identified data sets about the population of Wales, the SAIL Databank has a well-established and comprehensive data governance framework, data management framework, and project access processes [48]. The system holds several accreditations, such as the National Health Service (NHS) Data Security and Protection Toolkit, Cyber Essentials, ISO 27001, and is an accredited processing environment under the UK Statistics Authority Digital Economy Act. SAIL Databank runs upon the Secure e-Research Platform (SeRPUK) [49] which provides the technology to enable innovation such as supporting this work in developing a federated network. All datasets are made available through SAIL Databank housed on SeRPUK in the context of a project aligned to the 'five Safes'. The partnership between ICODA and SAIL permits accredited researchers with approved projects access to data. Through extending the capabilities of the databank, additional role profiles were developed to enable SAIL to define which projects, uses and datasets can participate in a federated network. The work discussed in this paper was to enable the integration between SAIL Databank and the ICODA Workbench to allow federated analysis tasks to be received, processed, outputs checked and returned.

Background on ICODA Workbench

The ICODA Workbench is provided by the Aridhia Digital Research Environment (DRE) [13] and is made up of two key components, a TRE and a repository for data and metadata discovery.

The TRE provides a collaborative way for researchers to manage, curate and analyse data. Initially developed as a secure research space for cancer researchers in Scotland, following the model of the Scottish Health Informatics Programme [50], it is now provided as a managed service. The platform was further developed as part of the AD Data Initiative [22]. Through the Dementias Platform UK (DPUK) consortium [51], the infrastructure on which SAIL operates, the Secure eResearch Platform (SeRP) UK, had already been participating as a federated site for AD Data Initiative and were able to build on this infrastructure to support the iPOP Study. When the COVID-19 pandemic started, the Bill & Melinda Gates Foundation [52], HDR UK and Aridhia combined efforts and contributed platform components to create a scalable trusted data sharing network to accelerate data-sharing, resulting in the creation of the ICODA Workbench. Once data are authorised as available and accessible, the project team can access and start working with the data using a workspace within the ICODA Workbench. A workspace is a safe, secure environment where data relevant to the project can be accessed by accredited researchers invited by the project leads. Researchers within the ICODA Workbench can utilise a range of preexisting tools as well as develop their own tools in their chosen coding languages. The ICODA Workbench was used by the iPOP Study team to send test federated analytics tasks to the SAIL Databank. Output results were manually reviewed at the databank and were securely transferred to the workspace to test the implemented federated network infrastructure. These data were further used to test a *meta*-analysis tool developed by Cytel [17] hosted on the ICODA workbench for the iPOP project.

Additionally, the workbench offers metadata creation, discovery, and data request capabilities, as well as connecting out to other data repositories throughout the world, such as the Health Data Research Innovation Gateway in the UK [53]. This is a vital component as it promotes FAIR data principles [54].

Background on Common API

An API is a connection between computer programmes [55]. It acts as an intermediary that enables organisations and companies to open their data and functionalities so they can communicate and share through a documented interface without compromising security. The Federated Data Sharing Common API ("the Common API") [37] was created to provide a mechanism to support federated analysis across individual data platforms when accessing and utilising data for research purposes. The use of the common API was initially piloted for the AD Data Initiative project to be able to analyse data consistently, despite

some of the data not being able to travel from a data repository. The primary goal was to simplify the decisions for data owners to participate in a federated network by offering standardised levels of participation. A secondary goal was to reduce the variety of interfaces and general complexity faced by researchers wishing to integrate data from multiple federated sources. This proposed solution allowed data users the flexibility in choosing how they want to use the data, abstracting this from the end-user by developing a Common API for federated networks. By providing trusted data sharing networks between research environments and data providers, the Common API aims to support collaborative science for data users and researchers.

The Common API is based on open internet standards World Wide Web Consortium (W3C) [56], Data Catalog Vocabulary (DCAT) [57], OAuth2 [58] and the Global Alliance for Genomics and Health Task Execution Service (GA4GH TES); an approach to federated bioinformatics [40]. The successful pilot leading to the development of the Common API brought together leading groups in the Alzheimer's Disease research community including Dementias Platform UK [51] and their Data Portal at SeRPUK [49], Swansea University [59], Critical Path Institute (C-Path) [60], Global Alzheimer's Association Interactive Network (GAAIN) [61] and technical partners including Aridhia Informatics. The development approach used was to adopt or adapt existing standards for the domain of distributed clinical research. To implement the federated network and perform federated analytics, the Common API was used to securely link the ICODA Workbench to SAIL Databank where the data for the iPOP Study was hosted. It was during this implementation of the federated network, the Common API was further developed by using the GA4GH TES API [62] for compute purposes. Additionally, clarifications on how workspaces within the TRE were referenced by the Common API calls were added. The Common API is made up of three sections [37]:

- Metadata – the component used for discovering metadata.
- Selection – the component which utilises GraphQL [63] formatted queries to interact with the database holding the project data. GraphQL was chosen as an abstraction over existing query approaches to optimise data extraction, not analysis.
- Federated compute – the component used to execute remote tasks on federated data. The permissions at execution time are defined by the permissions of the user on the stated project. The task-related API methods in GitHub [43] enables a user to submit a task for execution, monitor its progress and eventually retrieve the approved output from this task.

A standard workflow is recommended to manage secure remote computation. A user submits a task specification, and the recipient system executes the task in the background in an asynchronous model. Users can check the status of their task at any time. When the task succeeds, they are provided with a link to download the output. The first version of federated computation was limited to selecting structured data and there was no option to specify compute resources required. These limitations are expected to be addressed in future work.

Example of task API.

The task-related API methods in GitHub [43] enable a user to submit a task for execution. A federated task can be formulated in two sections:

1) The data selection uses the Selection API (internally) to format the data to be used or returned. Identify the dataset table name and variables of interest and construct a GraphQL compatible query as below:

query = "{table_name{varname1 varname2 varname3 …}}".

2) The containerised analysis script is to be run on this data. The name, tag and registry of the containerised script needs to be provided as below:

container = "{.
"name": "name of the container image",
"tag": "tag of the image",
"registry": "URL for container registry"
}"

Combining the two sections above, the task body will look like:

"task": {
"name": "name of the task",
"description": "small description of task",
"project_id": "project id received from data provider",
"queryInput": {
"selectionQuery": ***query***
},
"container": ***container***
}

The fields name, description and project_id are optional within the task body.

References

[1] S. Pooransingh, R. Abdullah, S. Battersby, R. Kercheval, COVID-19 highlights a critical need for efficient health information systems for managing epidemics of emerging infectious diseases, Front. Public Health 9 (2021) 767835, https://doi.org/10.3389/fpubh.2021.767835.

[2] N. Rieke, et al., The future of digital health with federated learning, Npj Digit. Med. 3 (1) (2020) 119, https://doi.org/10.1038/s41746-020-00323-1.

[3] A. Sadilek, et al., Privacy-first health research with federated learning, Npj Digit. Med. 4 (1) (2021) 132, https://doi.org/10.1038/s41746-021-00489-2.

[4] D. B. Satija, 'Supporting the adoption of privacy-enhancing technologies', Centre for Data Ethics and Innovation Blog.

[5] T. Desai, F. Ritchie, and R. Welpton, 'Five Safes: designing data access for research'.

[6] 'ICODA - Home', ICODA - A globally coordinated, health data-led research response to tackle the COVID-19 pandemic. Accessed: Dec. 20, 2023. [Online]. Available: https://icoda-research.org/.

[7] L. Dron, et al., Data capture and sharing in the COVID-19 pandemic: a cause for concern, Lancet Digit. Health 4 (10) (2022) e748–e756, https://doi.org/10.1016/S2589-7500(22)00147-9.

[8] H. Hallock, et al., Federated networks for distributed analysis of health data, Front. Public Health 9 (2021) 712569, https://doi.org/10.3389/fpubh.2021.712569.

[9] E.A. Voss, et al., Contextualising adverse events of special interest to characterise the baseline incidence rates in 24 million patients with COVID-19 across 26 databases: a multinational retrospective cohort study, eClinicalMedicine 58 (2023) 101932, https://doi.org/10.1016/j.eclinm.2023.101932.

[10] 'Sharing Sensitive Health Data in a Federated Data Consortium Model: An Eight-Step Guide'. World Economic Forum. Accessed: Jul. 30, 2023. [Online]. Available: http://www3.weforum.org/docs/WEF_Sharing_Sensitive_Health_Data_2020.pdf.

[11] 'World Economic Forum', World Economic Forum. Accessed: Jul. 30, 2023. [Online]. Available: https://www.weforum.org/publications/federated-data-systems-balancing-innovation-and-trust-in-the-use-of-sensitive-data/.

[12] 'Federated Analytics: Collaborative Data Science without Data Collection'. Accessed: Dec. 20, 2023. [Online]. Available: http://research.google/blog/federated-analytics-collaborative-data-science-without-data-collection/.

[13] 'Aridhia DRE | Trusted Data Sharing Network | Digital Research Environment'. Accessed: Dec. 20, 2023. [Online]. Available: https://www.aridhia.com/.

[14] 'Sail Databank - Home', SAIL Databank. Accessed: Oct. 29, 2024. [Online]. Available: https://saildatabank.com/.

[15] C. Calvert, et al., Changes in preterm birth and stillbirth during COVID-19 lockdowns in 26 countries, Nat. Hum. Behav. 7 (4) (2023) 529–544, https://doi.org/10.1038/s41562-023-01522-y.

[16] S.J. Stock, et al., The international Perinatal Outcomes in the Pandemic (iPOP) study: protocol, Wellcome Open Res. 6 (2021) 21, https://doi.org/10.12688/wellcomeopenres.16507.1.

[17] 'Cytel - Home', Clinical Trial Software & Data Analysis | Cytel | Contact Us. Accessed: Dec. 20, 2023. [Online]. Available: https://cytel.com/.

[18] 'Statistical methods in federated analyses | Swedish Medical Products Agency | Läkemedelsverket'. Accessed: Oct. 14, 2024. [Online]. Available: https://www.lakemedelsverket.se/en/about-the-swedish-mpa/reports-and-publications/federated-analyses/statistical-methods-in-federated-analyses.

[19] 'ATLAS'. Accessed: Oct. 14, 2024. [Online]. Available: https://atlas-demo.ohdsi.org/#/prediction/0.

[20] 'DataSHIELD'. Accessed: Oct. 14, 2024. [Online]. Available: https://datashield.org/.

[21] 'Global Alliance for Genomics and Health (GA4GH)'. Accessed: Dec. 20, 2023. [Online]. Available: https://www.ga4gh.org/.

[22] 'Alzheimer's Disease Data Initiative (ADDI)', ADDI. Accessed: Dec. 20, 2023. [Online]. Available: https://www.alzheimersdata.org/.

[23] 'BigMed Focus areas', Big Med. Accessed: Dec. 20, 2023. [Online]. Available: https://bigmed.no/focus-areas.

[24] 'CanDIG'. Accessed: Dec. 20, 2023. [Online]. Available: https://www.distributedgenomics.ca/.

[25] 'Home — Australian Genomics'. Accessed: Dec. 20, 2023. [Online]. Available: https://www.australiangenomics.org.au/.

[26] 'Autism Sharing Initiative', Autism Sharing Initiative. Accessed: Dec. 20, 2023. [Online]. Available: https://www.autismsharinginitiative.org.

[27] 'European-Canadian Cancer Network', EUCANCan. Accessed: Dec. 20, 2023. [Online]. Available: https://eucancan.com/.
[28] S. Gehring, R. Eulenfeld, German medical informatics initiative: unlocking data for research and health care, Methods Inf. Med. 57 (S 01) (2018) e46–e49, https://doi.org/10.3414/ME18-13-0001.
[29] M.R. Harris, L.A. Ferguson, A. Luo, Infrastructuring an organizational node for a federated research and data network: A case study from a sociotechnical perspective, J. Clin. Transl. Sci. 5 (1) (2021) e186.
[30] 'Data Partners', ehden.eu. Accessed: Dec. 20, 2023. [Online]. Available: https://www.ehden.eu/datapartners/.
[31] M.B. Palchuk, et al., A global federated real-world data and analytics platform for research, JAMIA Open 6 (2) (2023) ooad035, https://doi.org/10.1093/jamiaopen/ooad035.
[32] C. Alvarez-Romero, et al., Predicting 30-day readmission risk for patients with chronic obstructive pulmonary disease through a federated machine learning architecture on findable, accessible, interoperable, and reusable (fair) data: development and validation study, JMIR Med. Inform. 10 (6) (2022) e35307, https://doi.org/10.2196/35307.
[33] E. Jefferson, et al., A hybrid architecture (CO-CONNECT) to facilitate rapid discovery and access to data across the united kingdom in response to the COVID-19 pandemic: development study, J. Med. Internet Res. 24 (12) (2022) e40035, https://doi.org/10.2196/40035.
[34] 'Data Analysis and Real World Interrogation Network (DARWIN EU) | European Medicines Agency (EMA)'. Accessed: Oct. 04, 2024. [Online]. Available: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/real-world-evidence/data-analysis-real-world-interrogation-network-darwin-eu.
[35] D.T. Giles, et al., TRE-FX: Delivering a federated network of trusted research environments to enable safe data analytics, Zenodo (2023), https://doi.org/10.5281/zenodo.10055354.
[36] C. Orton, et al., TELEPORT: Connecting researchers to big data at light speed, Zenodo (2023), https://doi.org/10.5281/zenodo.10055358.
[37] 'common-api/doc/API_Overview.md at master · federated-data-sharing/common-api', GitHub. Accessed: Dec. 20, 2023. [Online]. Available: https://github.com/federated-data-sharing/common-api/blob/master/doc/API_Overview.md.
[38] 'What is a Container? | Docker'. Accessed: Dec. 20, 2023. [Online]. Available: https://www.docker.com/resources/what-container/.
[39] 'Container Checks — Singularity container 3.0 documentation'. Accessed: Dec. 20, 2023. [Online]. Available: https://docs.sylabs.io/guides/3.0/admin-guide/container_checks.html.
[40] 'GA4GH TES API: bringing compatibility to task execution across HPC systems, the cloud and beyond'. Accessed: Dec. 20, 2023. [Online]. Available: https://www.ga4gh.org/news_item/ga4gh-tes-api-bringing-compatibility-to-task-execution-across-hpc-systems-the-cloud-and-beyond/.
[41] 'Federated Data Sharing', GitHub. Accessed: Dec. 20, 2023. [Online]. Available: https://github.com/federated-data-sharing.
[42] federated-data-sharing/common-api-examples'. Accessed: Dec. 20, 2023. [Online]. Available: https://github.com/federated-data-sharing/common-api-examples.
[43] 'common-api-examples/src/data-profiler/README.md at main · federated-data-sharing/common-api-examples', GitHub. Accessed: Dec. 20, 2023. [Online]. Available: https://github.com/federated-data-sharing/common-api-examples/blob/main/src/data-profiler/README.md.
[44] 'How the Aridhia DRE is Enabling Federated Analysis in the PHEMS Consortium | Trusted Data Sharing Network | Digital Research Environment', Aridhia. Accessed: Oct. 14, 2024. [Online]. Available: https://www.aridhia.com/blog/how-the-aridhia-dre-is-enabling-federated-sharing-in-the-phems-consortium/.
[45] A.W. Toga, et al., The pursuit of approaches to federate data to accelerate Alzheimer's disease and related dementia research: GAAIN, DPUK, and ADDI, Front. Neuroinformatics 17 (2023), https://doi.org/10.3389/fninf.2023.1175689.
[46] 'Home', HDR UK. Accessed: Dec. 20, 2023. [Online]. Available: https://www.hdruk.ac.uk/.
[47] U. D. Service, 'What is the Five Safes framework?', UK Data Service. Accessed: Nov. 27, 2023. [Online]. Available: https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/.
[48] K.H. Jones, D.V. Ford, S. Thompson, R.A. Lyons, A profile of the SAIL databank on the UK secure research platform, Int. J. Popul. Data Sci. 4 (2) (2019) 1134, https://doi.org/10.23889/ijpds.v4i2.1134.
[49] 'SeRP UK Trusted Research Environment'. Accessed: Jan. 02, 2024. [Online]. Available: https://serp.ac.uk/serp-uk/.
[50] 'About | ScottisH Informatics Programme (SHIP)'. Accessed: Jan. 02, 2024. [Online]. Available: https://www.scot-ship.ac.uk/about.html.
[51] 'Welcome'. Accessed: Jan. 02, 2024. [Online]. Available: https://www.dementiasplatform.uk/copy_of_front-page.
[52] 'Bill & Melinda Gates Foundation', Bill & Melinda Gates Foundation. Accessed: Jan. 02, 2024. [Online]. Available: https://www.gatesfoundation.org/.
[53] 'Health Data Research Gateway'. Accessed: Jan. 02, 2024. [Online]. Available: https://healthdatagateway.org/en.
[54] 'FAIR Principles', GO FAIR. Accessed: Jan. 02, 2024. [Online]. Available: https://www.go-fair.org/fair-principles/.
[55] 'What Is an API (Application Programming Interface)? | IBM'. Accessed: Jan. 02, 2024. [Online]. Available: https://www.ibm.com/topics/api.
[56] 'About us', W3C. Accessed: Jan. 02, 2024. [Online]. Available: https://www.w3.org/about/.
[57] 'Data Catalog Vocabulary (DCAT)'. Accessed: Jan. 02, 2024. [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/.
[58] 'OAuth 2.0 — OAuth'. Accessed: Jan. 02, 2024. [Online]. Available: https://oauth.net/2/.
[59] 'Swansea University'. Accessed: Jan. 02, 2024. [Online]. Available: https://www.swansea.ac.uk/.
[60] 'Critical Path Institute - The Path Forward in Drug Development', C-Path. Accessed: Jan. 02, 2024. [Online]. Available: https://c-path.org/.
[61] 'The Global Alzheimer's Association Interactive Network'. Accessed: Jan. 02, 2024. [Online]. Available: https://www.gaain.org/.
[62] *ga4gh/task-execution-schemas*. (Oct. 28, 2024). Global Alliance for Genomics and Health. Accessed: Oct. 29, 2024. [Online]. Available: https://github.com/ga4gh/task-execution-schemas.
[63] 'GraphQL | A query language for your API'. Accessed: Jan. 02, 2024. [Online]. Available: https://graphql.org/.