DARE UK (Data and Analytics Research Environments UK) Sprint Exemplar final report

Multi-party trusted research environment federation: Establishing infrastructure for secure analysis across different clinical-genomic datasets

This work was funded by UK Research & Innovation [Grant Number MC_PC_21026] as part of Phase 1 of the DARE UK (Data and Analytics Research Environments UK) programme, which is delivered in partnership with Health Data Research UK (HDR UK) and ADR UK (Administrative Data Research UK).

DOI: 10.5281/zenodo.7085536

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

Contents

Executive Lay Summary

Combining data assets for research has historically involved the movement of data between organisations. Trusted Research Environments (TREs) now offer an alternative to data sharing by providing secure environments for approved researchers to access and analyse data. However, data are still held separately. Even where researchers have permission to use data held in two separate TREs, moving data between organisations from one TRE to another can be costly, complicated and time consuming. The purpose of this project is to demonstrate how TREs can be enabled to 'talk to each other' to facilitate analysis across separate databases as if they were one, a process known as 'federation.'

This project aimed to deliver the UK's first demonstration of genomic data federation by bridging the TREs of the NIHR Cambridge Biomedical Research Centre and Genomics England without moving original data, only the combined analysis results.

Both TREs contain rich, secure, governed sources of fully consented clinical-genomic data and we designed and implemented a system to enable federation of data between the two TREs. In a live demonstration, we showed that a researcher could query data within the two separate TREs to find individual records to create a group or 'cohort' of interest. For example, a group with certain characteristics such as the same mutations in their tumour. Federated analysis was then run across this joint cohort. The anonymous, non-patient level results were combined in a secure environment known as a 'Safe Haven,' before being released to the researcher through an Airlock which checked it was not possible to re-identify individuals from the results. Once optimised, this system has huge potential to lower barriers to research, simplify data use between organisations, and avoid the cost of moving large datasets.

Patient and public involvement has been embedded throughout the project through our patient partner sitting on the project board and in technical meetings, and three meetings with public contributors to shape thinking and governance structures.

This report contains learnings from the project and sets out opportunities for future work to unlock unprecedented possibilities for research and discovery for long-term patient and public benefit.

UNIVERSITY OF CAMBRIDGE

Lead organisation

NIHR | Cambridge Biomedical Research Centre

A partnership between the University of Cambridge and Cambridge University Hospitals to deliver ground-breaking research that benefits patients

Eastern AHSN

CAMBRIDGE UNIVERSITY Health Partners

Project Management

lifebit

A UK enterprise providing the software platform and federation capability

Genomics england

A public-sector clinical research endeavour, founded to deliver the 100,000 Genomes Project

1 Project aims

The multi-party trusted research environment federation Sprint Exemplar Project aimed to deliver a proof-of-concept live demonstration of federated analysis between two Trusted Research Environments (TREs); A 'patient-centred dataset' in the TRE of the national genomics endeavour, run by Genomics England, and a 'discovery research dataset' in the NIHR Cambridge Biomedical Research Centre (Cambridge) TRE called CYNAPSE (under the tenancy of the University of Cambridge, in partnership with Cambridge University Hospitals NHS Foundation Trust). The project would utilise Lifebit's CloudOS technology to enable joint querying and analysis to address defined research questions. We wanted to learn about the concerns of patients and the public regarding federation whilst navigating the steps of a federation process, obtaining the necessary approvals, consider the universally accepted standards to ensure extensibility, and consider the methodologies for performing distributed analyses.

2 Achievements against the DARE UK programme objectives

DARE UK aims to:

1. Design and deliver a novel and innovative UK-wide data research infrastructure that is joined-up, demonstrates trustworthiness and supports research at scale for public good.
    - This project has designed and demonstrated a novel and innovative federation capability between a UK Higher Education Institution (HEI), the University of Cambridge, and the national Genomics Endeavour, Genomics England
    - The project demonstrates privacy preservation, meaning that patient level analysis can be conducted across both datasets, by either Genomics England or the University of Cambridge, without disclosing potentially re-identifiable patient information from one environment to researchers in the other
    - It has worked with the Genomics England Design Authority to demonstrate trustworthiness and is the first stage of supporting optimised research at a scale that has previously been restricted by the Information Governance, security and cost barriers associated with moving exceptionally large datasets from one environment to another
    - This federation capability, once optimised, will speed up both initial discoveries and the time it takes to validate results, meaning a reduction in the time it takes for translational research to make a difference to patient lives

2. Establish the next generation of Trusted Research Environments (TREs) that will enable fast, safe, and efficient sharing, linkage, and advanced analysis of data, where it is legal and ethical to do so
    - A set of learning points from this project have been drawn up, which in combination with the technical architecture and open-source APIs may serve as a platform agnostic blueprint for organisations to consider when embarking on further projects in large-scale genomic data infrastructures with federation capabilities

3. Enable UK researchers and innovators to securely and efficiently harness the full power of linked datasets, modern digital platforms, tools, techniques, and skills
    - The federation capability demonstrated through the use case presented as part of this project brings to life the ability for researchers to harness the power of parallel cohort discovery and analysis over disparate datasets whilst integrating a choice of modern tools to facilitate the analysis

4. Enable research and analysis on a broad range of potentially sensitive data from across the UK research and innovation spectrum

- This project has focused on demonstrating a use case showcasing disparate analysis across genomic and health data. This has future cross-council application, such as the Medical Research Council (MRC) through health-related research, Engineering and Physical Sciences Research Council (EPSRC) through high-performance compute support and Biotechnology and Biological Sciences Research Council (BBSRC) through research into normal, healthy tissues and the ageing population

3 Progress against Sprint Exemplar project objectives

3.1 Develop a reference architecture for multi-party federation between two TREs, informing federated data infrastructure standards

The consortium has developed a reference architecture (*Figure 1)* which demonstrates the mechanism by which two TREs can federate. It has been configured to be customisable to the individual requirements of the organisations involved. The design allows a researcher to access the "Connect" workspace environment, which enables federated querying of both TREs to generate the sample cohort of interest and initiate the required primary analysis within each. Once this primary processing is complete, the results from the analysis are pooled and further processed in a secure environment known as a "Safe Haven," before release through an Airlock process. The design requires the movement of the aggregated results only, eliminating the previous need for source or individual participant data to be ingested into a single environment. The location of the Safe Haven is movable depending on the Information Governance (IG) requirements of the individual data controllers involved. It could sit within a third-party neutral environment (*Figure 1*), or within the environment of one of the collaborating organisations (as demonstrated during this project, (*Figure 2*). The reference architecture, therefore, provides a blueprint for federation between multiple environments which have been appropriately configured to facilitate communication with others.

Here, we have successfully demonstrated the implementation and application of the architectural design using the TREs owned by Cambridge and Genomics England, respectively. In this instance, owing to the short timescale of the project, two TREs were chosen that are powered by Lifebit's CloudOS platform, built using AWS public cloud infrastructure. To demonstrate the use case within the 8-month project and avoid potentially lengthy user access requests and Data Access Agreements between the two organisations, we selected a large fully anonymised and publicly available dataset which the University of Cambridge houses within CYNAPSE on behalf of Prof Serena Nik-Zainal. Prof Nik-Zainal's research group have also already been granted access to the fully consented whole genome sequences held by Genomics England from patients with cancers.

The consortium worked closely together with the Design Authority at Genomics England to ensure they were comfortable with the architectural design. Genomics England have well-established Information Governance (IG) policies and processes, and their Design Authority considered the original proposal where the Safe Haven would sit in a neutral environment. Genomics England are keen to enhance their IG framework to facilitate federation with other TREs in future, but unfortunately the timescale of this project was too short to complete these types of amendments. The project time frame was also not conducive to fully perform penetration testing of a federated system with a neutral Safe Haven. For these reasons, the Design Authority requested that the Safe Haven would be located within the boundaries of the Genomic England TRE infrastructure (*Figure 2*).

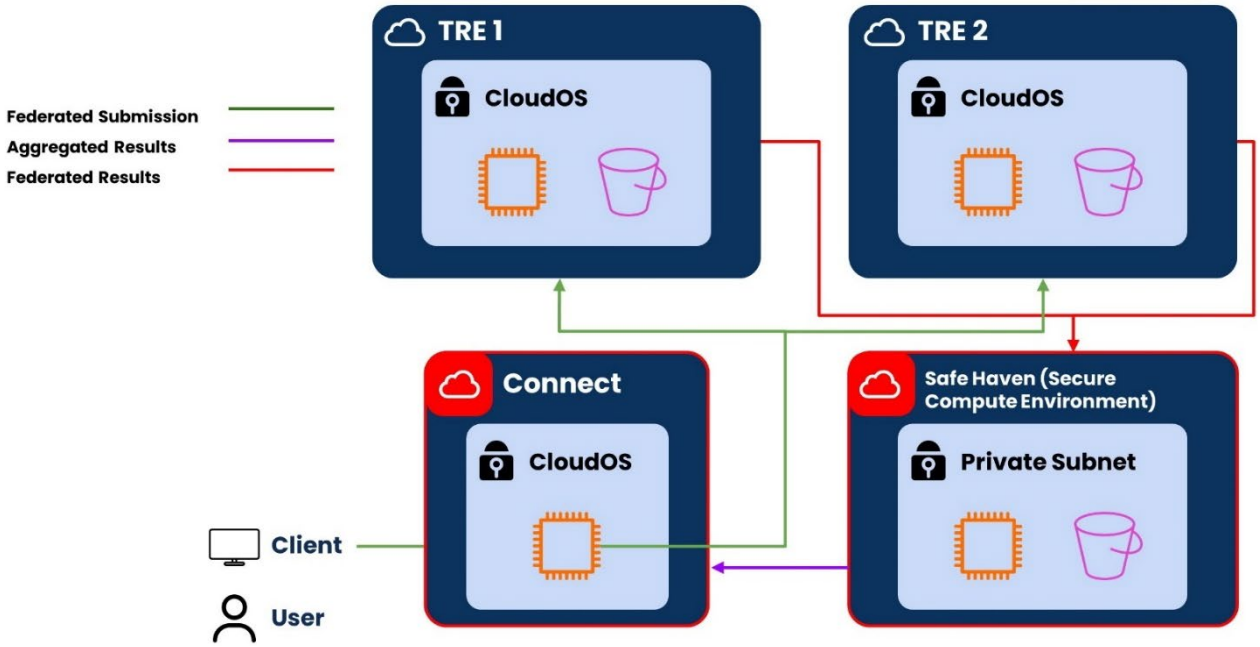

Figure 1: Reference architecture for federation between two TREs

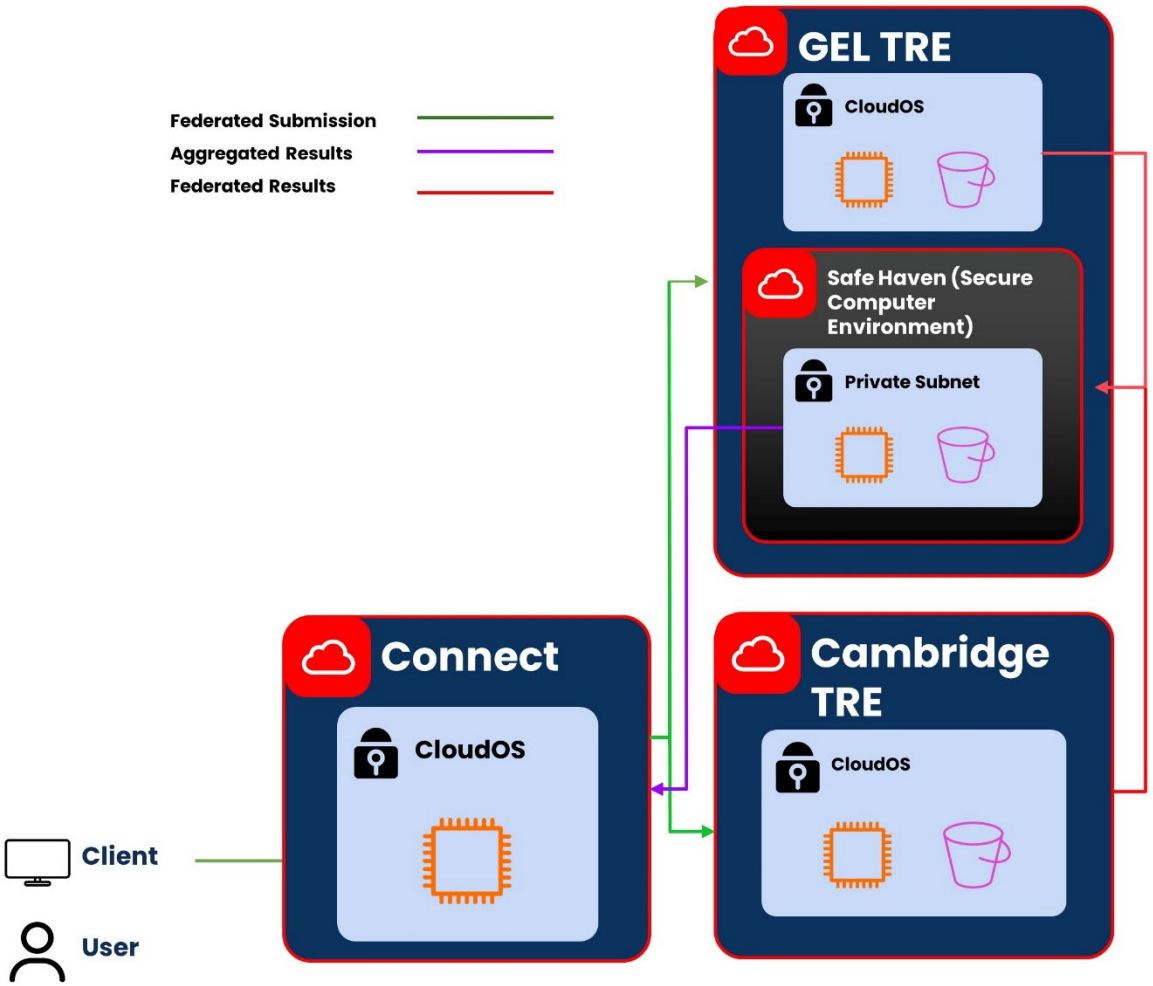

Figure 2: Reference architecture for federation between the TREs of Genomics England and the University of Cambridge

3.2 Develop open-source APIs required to facilitate secure communication and computational coordination between the TREs of Genomics England and Cambridge

Application programming interfaces (APIs) have been developed and tested during this project to orchestrate the Connect system depicted in *Figures 1 & 2*. We have aligned with Global Alliance for Genomics & Health (GA4GH) standards for interoperability. The APIs interact with almost all other services involved in the system, allow interaction with multiple TREs, GA4GH Passport broker at Genomics England and the Safe Haven to control the flow of job submission and results access. The APIs we have used fall into several categories:

- User management/workspaces
    - This API is essential to manage the users accessing the system. It checks the status of the user approval process, sends a request to approve the users into the application and creates a workspace where users can share cohorts, tools, and analyses
- Service information
    - This API shows information about the service
- GA4GH – Tool Registry Service (TRS)
    - This API provides a standardised way to describe to availability of tools and workflows, having a constant way to interact, search and retrieve information from Docker-based tools and various workflows (Nextflow in this case) which allows these tools to be easily portable across systems using the same standard
- GA4GH – Workflow Execution Service
    - This API describes a standard programmatic way to run and manage workflows. Having this API supported by multiple execution engines will allow the running of the same workflow (analysis) across different platforms and clouds/environments. It supports bringing a researcher's analysis to data controlled by an external organisation, rather than needing to make a copy of the data
- Results aggregation
    - This API describes how to orchestrate within Safe Haven the aggregation of results after a job is finalised across the multiple TREs, and returns the status of the results aggregation

Details of the code used for the APIs have been made openly available on SwaggerHub here.

By using these APIs, we have enabled secure communication and computational coordination between Genomics England and Cambridge to enable parallel and joint data querying in the respective local TREs over their local data. These APIs complement the pioneering work of the GA4GH, as well as other organisations establishing these standards.

3.3 Develop the requisite blueprint and governance standards for secure federated data interoperability amongst TREs

The consortium has reflected upon the shared learning points gained during the project (**Appendix 1**). These, in combination with the reference architecture designed and the list of open-source APIs used could serve as a blueprint for any future project looking to develop federation capabilities between large-scale genomic data infrastructure.

During this project, we found that data standardisation to a Common Data Model (CDM) was of paramount importance for enabling joint (federated) queries of data under the custodianship of different organisations. Here, we standardised the data to the Observational Medical Outcomes Partnership (OMOP) CDM, a globally recognised CDM in the field of genomics and health informatics. Both Genomics England and Cambridge data were transformed into the OMOP CDM before the federated testing could commence.

3.4 Present a live use-case bridging the University of Cambridge and Genomics England TREs, underscoring the value of federation

On Wednesday 27th July 2022 Professor Serena Nik-Zainal (Principal Investigator) and Dr Pablo Prieto (CTO, Lifebit) presented a live use case demonstration on real data at the DARE UK Final Sprint Exemplar Day in London. The demonstration involved the journey that a user would take to login to Connect securely, browse the cohort contained within the disparate datasets in the TREs of the two organisations, select a cohort of interest, run a federated analysis, request access to the aggregated results and view some results visualisations. This journey is depicted in *Figure 3*.

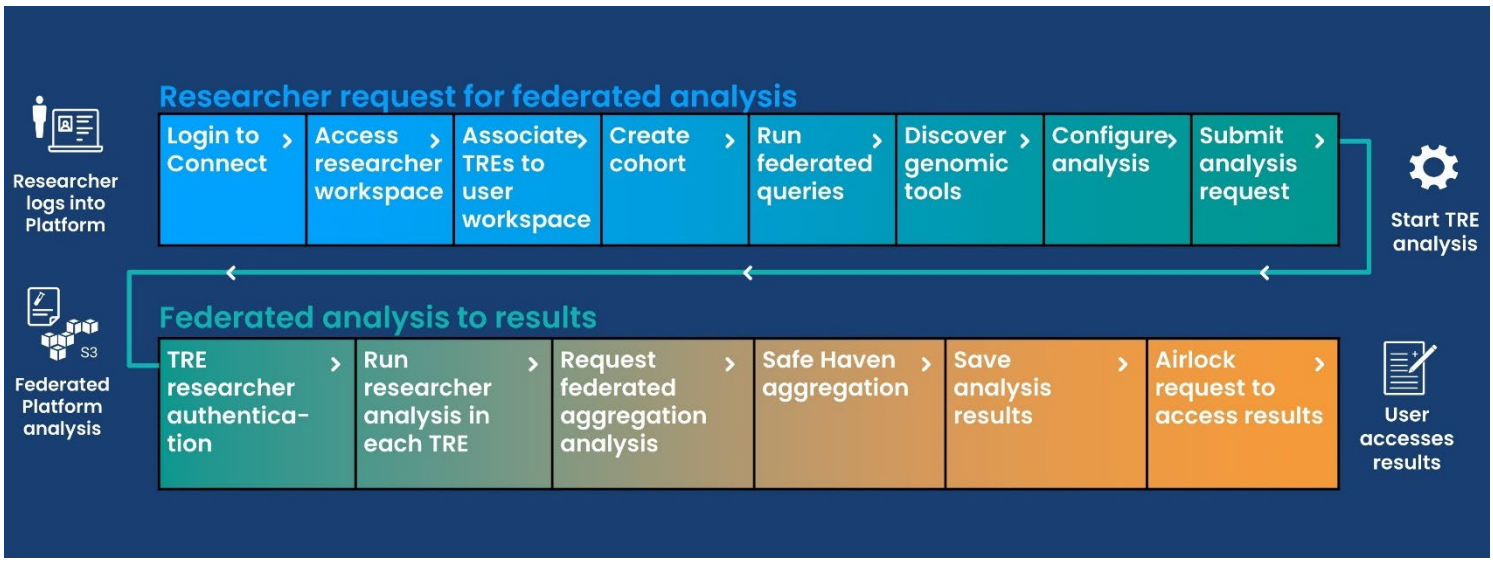

Figure 3: Depiction of the user journey for using the Connect system to run analysis over disparate datasets held within two different Trusted Research Environments.

As part of our use case demonstration, we performed a joint, parallel query over the Genomics England and University of Cambridge datasets. Leveraging the value of having normal and tumour whole genome sequence samples for cancer patients of what may be, when combined, the largest cohort analyses of whole genome sequence cancers worldwide, we performed a tumour analysis to explore somatic mutational signatures. Given the federated TRE architecture and strict Airlock Policy, there were several requirements:

1. Successful implementation of GA4GH passport authentication process
2. Parallel analyses performed within the secure environment of the respective Genomics England and University of Cambridge TREs. Distributed analysis pipelined to run using Nextflow workflows
3. Only aggregated results (non-identifiable summary statistics) can be exported via Airlock from the Genomics England TRE

The live use case was a UK first demonstration of distributed analytics through a federated system between a higher education institution (HEI) and the national genomics endeavour (Genomics England). It serves as a blueprint for a federation capability that, once optimised, will heavily reduce the current time burden on researchers to conduct their analysis over integrated cohorts by avoiding the Information Governance, security and cost barriers associated with moving exceptionally large datasets from one environment to another.

3.5 Work towards designing a novel, secure, role-based access control (RBAC) and scalable Airlock process for federated analysis between TREs

One of our most ambitious aims was to work towards establishing a secure, role-based access and scalable Airlock process at Genomics England which can account for federated analysis. This has not been proposed previously. A new Federated Airlock process would have two underlying goals. First, to be a more scalable, automated, and continuous process than the current manual design. We discussed the possibility of using workflows to perform the analysis which the Genomics England Design Authority had pre-approved, so they could be sure that any results generated by that workflow would be fully aggregated, with no risk of reidentification. Secondly, the development of automated federated governance rules which are required to enable programmatic triggering of the enhanced Airlock processes.

Genomics England has a long-established IG policy, which demands that all results are reviewed by an Airlock process (currently manual) before their release is granted to the researcher. After early discussions, the Genomics England Design Authority decided that with the short time frame for this Sprint Exemplar it would not be possible to amend the Genomics England IG processes to accommodate a fully automated Airlock process. They therefore requested that the results of the use case go through the manual Airlock process to authorise release. This means retrieval of the federated analysis results requires an export request to be raised through the GoAnywhere system, the current solution used by Genomic England to manage the Airlock process. However, the Design Authority approved a request to automate the process of raising a ticket on the Genomics England Jira system to trigger an Airlock request, before proceeding to the manual review step (*Figure 4*).

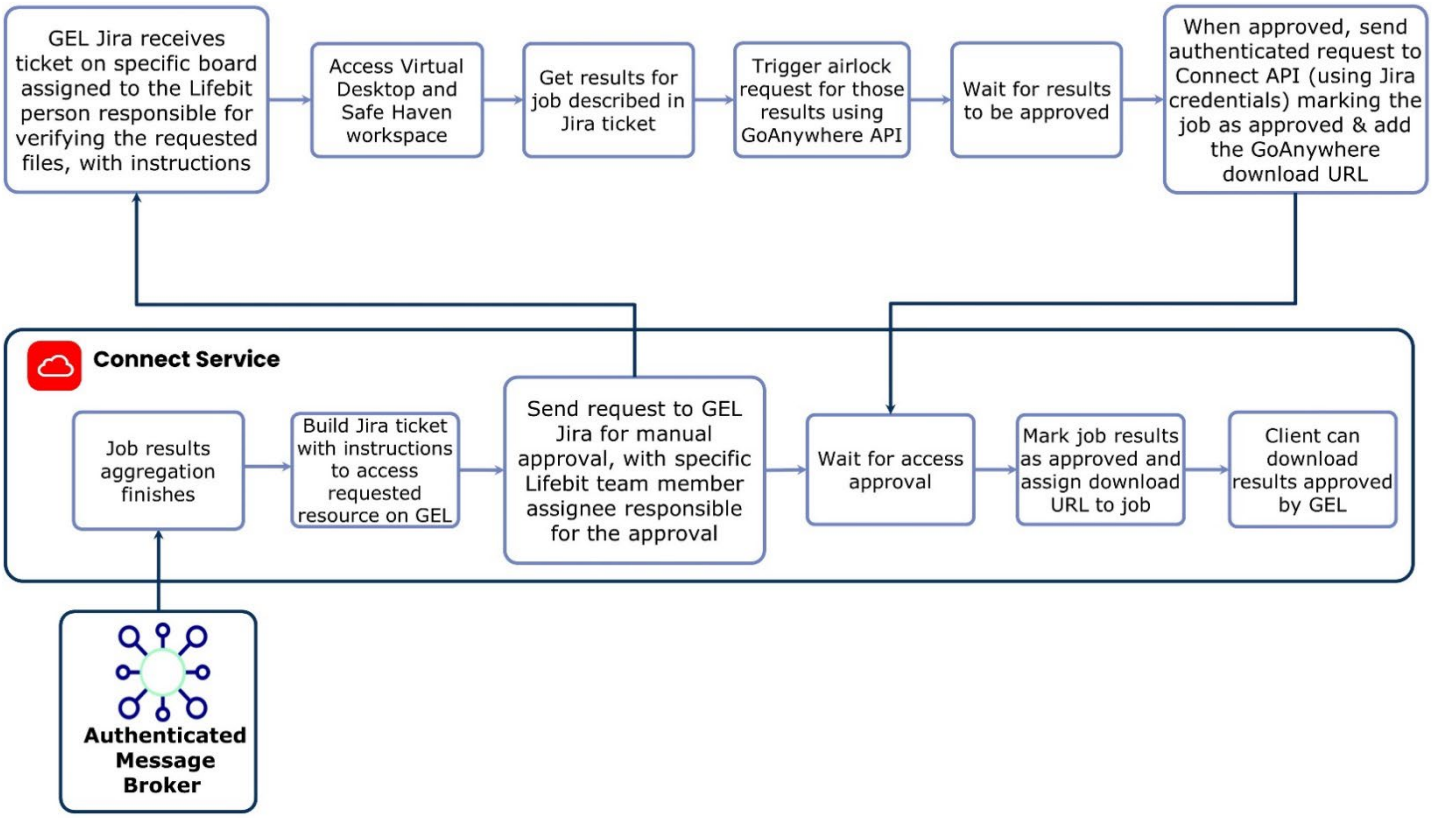

Figure 4: Schema for semi-automated Airlock triggering of the Genomics England Airlock process

3.6 Embed PPIE within the project

Patients and the public have been integral to our Sprint project from the very beginning, with our patient partner, Rosanna Fennessy, as a collaborative partner Ms Fennessy reviewed the original project proposal and advised on embedding PPIE activities within the project from the outset. When the project commenced, she became a member of the formal governance boards, was involved in technical scoping discussions with Lifebit and sat on the Patient Public Involvement & Engagement (PPIE) workstream throughout.

As the project involved TREs of Genomics England and NIHR Cambridge BRC, we focused our initial PPIE work on the PPIE groups that represent these organisations, the Cambridge University Hospitals PPI panel and Genomics England Participant Panel. Three separate activities were held with these groups over the course of the Sprint. A Q&A session to build on a shared explanation of the project, a focus group to gather patient and public preferences on data access requirements, and an information session on federation. Additionally, the team have been working with patients, the public and GEL study participants across these panels to answer their questions about the work underway and co-write a Frequently Asked Questions document.

The members of the public that we have been working with now have a shared understanding of what federation is and many members have expressed their interest in following this journey. A key learning point for the team is that the concept is still unknown to many, but once patients and the public fully understand that by using federation, source data does not move, only the results, they heavily encourage this approach and express frustration that federation is not already common practice.

The project has achieved engagement with patients and the public about the concept of federation, demonstrating a methodology that strengthens the safety of patient data.

"I have been delighted to be involved with this exciting project as a patient partner since the beginning. This has enabled me to better understand the potential that federation brings, both in terms of opportunities for researchers using health data and for patients and the public in terms of ensuring the safety of their data. Many patients and members of the public have worked in collaboration with the Sprint team to shape this project, both to the stage it is at now and with how it plans to move forward in the future. Involving us at this early stage will undoubtedly benefit both researchers and the wider public so that we can ensure the safe and fair use of health data to maximise improved outcomes for all."

Rosanna Fennessy, Patient Partner

---

4 Learning points

During our monthly meetings with the DARE UK team, it was made clear that a particularly valuable output of the Sprint projects would be communication around any learning points that arose over the course of the project, which could influence future projects being undertaken in similar areas. *Table 1* provides a high-level overview of some of the challenges we faced. **Appendix 1** provides a deeper dive into key insights and learnings from building a federation capability between the TREs of Cambridge and Genomics England.

Challenges	Key learnings	Possible future solutions
Misalignment of understanding of basic terms such as 'data,' 'linking' & 'federation' is not conducive to constructive dialogue	Leading with clear communication	Define and communicate key issues, agreeing a common lexicon
Recognising that PPI focus groups and workshops are not enough to shape the project as it moves through its lifecycle, we needed to ensure that our patient partner felt embedded and had true project oversight	Defining involvement of a patient partner throughout the project governance structure empowered them to feel confident in making meaningful contributions, permissive of rapid progress	Fully embed PPI representatives at all levels of project governance from the outset
Data Custodians agreeing to a federation partnership is not a trivial commitment, requiring full commitment to navigating complex design, implementation, and testing processes	Federate to collaborate: Right partners, right purpose, right levels of TRE security	Adopt a federated approach with appropriate partners to enable secure collaboration
The time required by Data Custodians to review the proposal and security arrangements was considerable, particularly given the pilot nature of the project	Federation is a complex multi-step process and requires rigorous security & Information Governance reviews	Allow additional time for review and approval process whilst federation is still in the early stages of its application. Once the end-to-end process is optimised, the time required is likely to reduce
Different user authentication systems presented challenges to streamlined, user-friendly federation	Unify the authentication processes, such as GA4GH passports	Prepare for future TRE extensibility by aligning to an industry-recognised authentication system
Automated processes that harmonise and integrate disparate datasets were required	Use of ETL (Extract, Transform, Load) pipelines can permit efficient large-scale raw data conversion to standardised, analysis-ready data. Reusability by design	Wide adoption of a common data model, such as OMOP, could be considered on a large-scale by the community to permit future federation capabilities
A location for the Safe Haven was not agreed by the Data Custodians from the outset	Alternative models of location of a 'Safe Haven' needed to be developed	Architecture permissive of the Safe Haven location being moved according to the partners involved, with enough time allocated to performing security checks/penetration testing activities
TRE firewalls were not necessarily permissive of scaling up security measures for federation across multiple TREs	Security controls need to be permissive of scaling up, whilst maintaining security of data and systems	Modification and scaling of security controls is needed to manage large-scale data across TREs

Table 1: High-level level challenges faced during this project and the key lessons we learnt from working closely with partners to find solutions

5 Impact of the project and future potential

5.1 Impact

This project is a first demonstration of distributed analytics through a federated system between a UK higher education institution (HEI) and the national genomics endeavour (Genomics England).

It has provided an architecture, Airlock design and open-source APIs permissive of replication for federation of other NIHR Biomedical Research Centres which are often (like the one in Cambridge) a partnership between a university and an NHS Trust.

This report details our learnings from the project which can be taken forward to facilitate a smoother journey towards federation between other sites.

The use case has demonstrated a federation capability that, once optimised and rolled out, will heavily reduce the current time burden on researchers to conduct their analysis over integrated cohorts, avoiding the significant costs and risk of data leaks associated with moving large datasets between environments.

This capability, once optimised, will speed up both initial discoveries and the time it takes to validate results, meaning a reduction in the time it takes for translational research to make a difference to patient lives.

DARE UK support was critical in accelerating the federation aspirations in Cambridge, which as part of the CYNAPSE project was intended to begin in 2024. The Sprint Exemplar funding expedited a proof-of-concept demonstration by more than 18 months, and firmly placed federation analytics as an area to cultivate going forward. The use case was an example of performing research across a 'discovery research dataset' and 'patient-centred dataset', emphasising how federation could contribute to reducing the gap between discovery research and clinical translation, through analysis and immediate validation in such cohorts.

Following the 27th of July meeting, a health-related agency has reached out to consider further funding support to enable federation between academic/hospital sites and GEL, and a non-medical agency has also reached out to seek collaborative potential uniting health and environmental data.

5.2 Future potential

- Optimisation and full implementation of an automated Airlock design at Genomics England
- Develop federation capabilities with multiple sites using a different TRE provider and/or public cloud platform
- Extension of federated capabilities beyond UK TREs to define and test international standards
- Demonstration of federated machine-learning
- Collaboration with the East NHS Genomic Medicine Service (GMS) towards a federated network of TREs
    - Norfolk & Norwich
    - Nottingham
    - Leicester
    - Papworth
    - Cambridge
- Implementation of a federated machine-learning clinical decision support tool to enable decision-making and increase efficiency for clinical service delivery for the NHS GMSs
- Wider dialogue with the public to promote a shared understanding of federation and its future potential

Appendices

Appendix 1: Learning points

Learning points

Multi-party trusted research environment federation: Establishing infrastructure for secure analysis across different clinical-genomic datasets

Purpose

The DARE UK project provided a unique opportunity for the exemplar demonstration of multi-party federation between the Trusted Research Environment of the NIHR Cambridge Biomedical Research Centre, and the Genomics England Research Environment, which houses NHS whole genome data. Here, we discuss our insights and learnings from the experience of building federation capability between these sites.

Introduction

The opportunities for data-driven research and innovation have never been greater. Trusted Research Environments (TREs) (NHS Digital, 2022) (Genomics England, 2022) have become established as a means of housing large datasets in a secure way, by only providing access to authorised researchers (UK Data Service, 2022). Yet, as a community, we could increase the potential of siloed datasets in different TREs, through aggregation and/or combining different data modalities, increasing power to accelerate health-related discovery, ultimately enhancing our understanding of how to detect, prevent and treat disease. Federation provides an exciting solution for how to analyse such disparate datasets *in situ* without the need to move source data.

Through DARE UK support, we performed a proof-of-concept study creating a federation link between a 'discovery research dataset', the NIHR Cambridge Biomedical Research Centre (Cambridge) TRE called CYNAPSE (under the tenancy of the University of Cambridge, in partnership with Cambridge University Hospitals NHS Foundation Trust), and a 'patient-centred dataset' in the TRE of the national genomics endeavour, run by Genomics England. Here, we report summary insights gained from patient-public preparatory and consolidation events, planning, approvals, technical considerations, and the use-case distribution analysis demonstration performed through these two TREs that had agreed to a federation partnership. This sprint project involved the University of Cambridge, Genomics England, technical platform input from a UK enterprise called Lifebit, project management support from Eastern Academic Health Science Network (Eastern AHSN) and Cambridge University Health Partners (CUHP) and had patient-public involvement throughout. Lessons learned during this project should be further considered for future projects working on the federation of large-scale genomic data infrastructures.

Learning point 1: Leading with clear communication

Define and communicate key issues and agreeing a common lexicon

Basic terms such as 'data,' 'security,' 'privacy' and 'linking' mean different things to different people and can conjure strong emotive responses. For example, there are many types of 'data' with differing levels of privacy attached (personally vs identifiable vs deidentified vs) and thus carrying varying levels of sensitivity. This is often

not defined at the start of conversations, which can cause misunderstanding culminating in onerous discussions that may be irrelevant. Through our patient and public involvement and engagement (PPIE) experience, it was clear that defining the discussion topics was key to ensuring constructive engagement. It could be in the interest of the data community to engage in a campaign to further inform the public on often misunderstood areas of concern, encouraging dialogue on defined issues. Some recurrent (non-exhaustive) topics that we met were as follows:

- The 'data privacy' areas that most concerned the public were reasonable but not always applicable in the context of federation between these specific TREs. Some were concerned with re-identification, while others were worried about collection of data by stealth. The major concern lay with how local Data Custodians ensured that *bona fide* researchers were performing legitimate projects
- The mention of commercial entities using genomic data often triggered alarm. A subject to explore could be the true risks of commercial entities using genomic data in the UK, and the distinction between bona fide research commercial entities (e.g., pharma) and other commercial entities such as insurers or consulting firms
- An area fraught with concern was misuse of genomic data by insurance providers and/or governments. It could be enlightening to discuss factors used by insurance providers today and to reflect on the Code on genetic testing and insurance (UK Government, 2022), an agreement between the government and the Association of British Insurers (ABI), on the use of genetic test results in underwriting insurance policies

In addition, data infrastructure projects add layers of technical terminology – data lakes, Trusted Research Environments, airlocks, controls, passports. As researchers and scientists, we could improve how we use these technical terms and agree a common lexicon between us, to minimize confusion and maximise approachability.

Learning point 2: Patient/public/participant involvement

Active involvement of patient/public representatives in development and governance

This project has benefited from the involvement of a patient/public/participant representative at every step. Active PPIE representation at every governance level, including at monthly meetings of the Strategic Partnership Board, Information Governance and PPIE workstreams meant that our patient partner had ongoing, consistent oversight of the project, often helped to re-focus discussions, and asked pertinent questions to ensure sight of goals. Additionally, having a PPIE representative present in meetings helped to maintain discipline amongst technical and academic teams, ensuring that discussions were held in language that was accessible, resulting in all partners understanding the issues and actions raised, and were ultimately able to progress rapidly.

Learning point 3: Federate to collaborate: Right partners, right purpose, right levels of TRE security

Adopt a federated approach with appropriate partners to enable secure collaboration

Federated approaches involve a computing paradigm whereby linking technologies permit data analysis to occur at multiple independent sites such as TREs. It is not however, a trivial commitment between Data Custodians to agree a federation partnership. For our two sites (Cambridge and Genomics England) to federate, the benefits had to outweigh the risks; the two sites were expected to adhere to the highest standards of data security, the reasons why federation would be academically and clinically valuable were clear, and other factors that made federation the best solution were fulfilled.

Right partners: The Genomics England TRE houses > 130,000 whole genomes derived from de-identified NHS patient-centred data. The NIHR Cambridge BRC TRE (CYNAPSE) contains a disparate collection of de-identified and

anonymised datasets, some from patients, others not, representing a wide array of discovery research datasets. Lifebit brought their unique experience in federation and TRE technology, as the providers of the TREs of both Genomics England and CYNAPSE (Lifebit CloudOS).

Right purpose: The data in these two TREs are not amenable to pooling; they are both too large to move and cannot be physically pooled/relocated for legal, regulatory, and practical reasons (Cell Genomics , International federation of genomic medicine databases using GA4GH standards, 2021) (Gardeni, HawkIinsii, & Winickoff, 2021). Yet both partners benefit from the collaboration with each other because the discovery and validation opportunities offered by combining subsets of data from the two TREs are considerable. The ability to increase power for discovery and the potential for seeking validation in NHS patient cohorts is immediate.

Right levels of TRE security: Our purpose here is not to elaborate on expected standards for TRE structure of participating federated sites exhaustively. That should be revisited in future dedicated projects. Here, we note some (non-exhaustive) basic expectations between our two sites:

- TREs should be compliant with industry-wide standards beyond General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), for example to standards expected through ISO 27001 certification and/or the UK government-backed scheme, Cyber Essentials Plus. (ISO, 2022) (National Cyber Security Centre, About Cyber Essentials, 2022)
- TREs should implement the Five Safes framework: Safe People, Safe Projects, Safe Settings, Safe Data and Safe Outputs (UK Data Service, 2022)
- TREs should apply trusted data controls to maximise security at each stage of the data lifecycle including Deidentification, Encryption, Airlock (a security process to manage movement of data into and out of the TRE which must be approved by an authorised team), Role-based Access Control (Regulating which users can view or use resources within the TRE), Tiered Access Levels (data access is tiered by levels according to end-user type), Segregation (ability to segregate datasets and workspaces to meet compliance requirements and restrict user access).

With the right partners, right purpose, and right levels of TRE security in place, in the short timeline of this project, we have been able to perform a distributed analysis through a federated relationship between a university/hospital TRE and Genomics England. Through this use case demonstration, we were able to:

1. Determine the target architecture of the federation process including detailed management of the authentication process, return of data through an airlock and Safe Haven location
2. Develop the APIs associated with performing a federated query using GA4GH standards

Learning point 4: Allow time for review and approval process

Federation is a complex multi-step process and requires rigorous security review

Maintaining data security is paramount for Data Custodians, even if dealing with deidentified patient data. In presenting proposals for target architecture, authentication plans and airlock designs as examples, multiple rounds of review and assessment were critical to guarantee that data governance expectations for the relevant TREs had been adhered to. From the perspective of the Data Custodian, for every update, a Data Protection Impact Assessment (DPIA) is carried out which covers risk assessment, looks for evidence of continued improvement of the TRE's security posture and a plan to address vulnerabilities longer term. Following completion of the DPIA, penetration or vulnerability testing should be carried out, where the DPIA detects a weakness or a threat.

In this project, one TRE contained fully consented deidentified NHS patient data and the other contained anonymised data. Yet, the time required for review was considerable, and we would expect the process to be lengthened if TREs contained more complex or patient identifiable information.

Learning point 5: Unify the authentication processes

Prepare for future TRE extensibility by aligning to an industry-recognised authentication system

A common or singular Authentication process across Data Custodian organisations involved in a multi-party TRE system is essential to allow universal, secure access to researchers. Selecting authentication processes that are established and well-validated in the industry, such as GA4GH passports, can help set up for future interoperability with other platforms and systems. A singular authentication process saves time for researchers too, removing the need for multiple authentication points across different TREs.

Where it is not possible to unify authentication processes, efforts should be made to provide an authentication translation service. Future work could include building APIs to support translations between other systems and the GA4GH passport.

Learning point 6: Reusability by design

Adopting a common data model aligned to FAIR Data principles

FAIR Data Principles (Wilkinson, 2016) are 'designed to enhance the reusability of data holdings', focussing on activities that facilitate the Findability, Accessibility, Interoperability and Reusability of scientific data assets.

A TRE needs to support the full data quality lifecycle, including ingestion, curation, harmonisation, and quality control. TREs are frequently acquiring data from multiple data sources, calling for automated processes that harmonise and integrate disparate datasets. Established ETL (Extract, Transform, Load) pipelines within the platform can permit the efficient flow of large-scale raw data and conversion into standardised, integrated analysis-ready data. (UK Health Data Research Alliance & NHS X, 2021) (Cell Genomics, 2021).

Federated queries need to be performed on clinical, phenotypic and omics data items (or fields) that are aligned to a common data format or model. Therefore, the alignment of these data to a common standardised format, at source, prior to ingestion is also of high importance. Effective harmonisation and standardisation align health data across organisations so that data resources can be queried more quickly and efficiently.

Transforming the data into a common format (in this case, OMOP Common Data Model) using a standardised set of vocabularies and metadata that aligns with industry standards, means it can meet the Findable, Accessible, Interoperable, and Reusable (FAIR) principles that are vital for enabling joint queries across diverse sites. (Observational Health Data Sciences and Informatics, 2022) (Wilkinson, 2016)

Learning point 7: Alternative models of location of a 'Safe Haven'

Identify the most appropriate location for aggregation of results post distributed analyses/federated learning

A 'Safe Haven' is the environment where aggregated results are placed following a distributed analysis or a federated learning exercise that has been run within the different Data Custodian TREs. This location should be an area of 'minimal data risk.' The Safe Haven could be in a neutral zone and this model would work for two or more sites that are federating with each other. An alternative model, depending on governance, security, consent, or operational requirements, is to place the Safe Haven within the TRE of a single Data Custodian. For this sprint project, we opted for the second model, the selected Safe Haven was embedded within the Genomics England TRE, as this TRE was more mature, contained over 130,000 whole genome datasets of NHS patients, and had

more advanced vetting systems and security measures already in place. A one-size-fits-all approach is not necessarily the best option and some flexibility in location of Safe Haven may be beneficial depending on security, information governance, or operation considerations.

Learning point 8: Security controls that are scalable

Modification and scaling of security controls is needed to manage large-scale data across TREs

Our work has flagged two feature alterations that may be needed if scaling security measures to multiple TREs. Providing a researcher has been approved for access to respective TREs through local mechanisms:

We suggest that **TRE firewall configuration** could be permissive of all ingoing operations/requests (i.e., query and analysis requests) and be restrictive of outgoing operations/requests, limiting them to essential ones only, following a "least privilege" framework (national Cyber Security Centre, 2022). The least privilege framework restricts access rights to users, accounts, and computing processes, to resources that are absolutely required to perform authorised, routine activities. This configuration is critical to protect against unauthorised data leakage.

We also suggest that an **automated Airlock process** is critical to enable a more scalable approach for federated analysis across distributed TREs. There is potential to explore the concept of "approved workflows" in designing an automated airlock process. These would require a single human authentication to be granted at the start of a study for the workflow (analysis) itself because the code has been written to aggregate (and therefore anonymise) results as part of the process. It has also been configured to abort the process if data are insufficiently aggregated to proceed, so no potentially identifiable data could get through. The workflow could therefore be deemed as safe to export results without any further review, which means all remaining airlock steps could be automated. This design may prove very useful for highly standardised and well-known processes such as GWAS. Analyses like these often require the researcher to run them multiple times, which under the current airlock process would require human authentication to release the results on each occasion. Handling in a pre-approved way, with the approval given for the code at the start of a study, would reduce the need for human intervention (and therefore approval or data release delays), in the Airlock process.

Appendix 2: Acknowledgements

University of Cambridge & NHS East Genomics
Serena Nik-Zainal
Cambridge BRC Executive Council
Miles Parkes
Eamonn Maher
Victoria Hollamby
Andrea Degasperi
Yasin Memari

Patient/public/participant representation
Rosanna Fennessy
Amanda Stranks
Cambridge University Hospitals PPI Panel
Genomics England Participant Panel

LifeBit
Maria Chatzou-Dunford
Pablo Prieto
Thorben Seeger
Eleni Kyriakou
Chiara Bacchelli
David Ardley
Sangram Keshari Sahu
Bruno Goncalves
Ismael Peral
Filippo Abbondazo
Christina Chatzipantsiou
Rafael Zochling
Bevinson Binu
Jonathan Amir
Ingrid Knarston

Eastern AHSN & Cambridge University Health Partners
Mark Avery
Eleanor Hall
Keiran Raine
Charlie Clarke
Celia Kelly
Simon Day
Tony Taylorson

Genomics England
Francis Carpenter
Geoff Coles
Graham Binns
Parker Moss
Mark Tilley
Dominic Stait
Yufan Chen
Simon Wilde
Carmelo Velardo
The Design Authority at Genomics England

DARE UK
Elizabeth Waind
Susheel Varma

References

Cell Genomics . (2021). *International federation of genomic medicine databases using GA4GH standards*. Retrieved from https://www.cell.com/cell-genomics/pdf/S2666-979X(21)00039-2.pdf

Cell Genomics. (2021). *GA4GH: International policies and standards for data sharing across genomic research and healthcare*. Retrieved from https://www.cell.com/cell-genomics/fulltext/S2666-979X(21)00036-7

Gardeni, H., Hawklinsii, N., & Winickoff, D. (2021). *Building and sustaining collaborative platforms in genomics and biobanks for health innovation*. Retrieved from https://www.oecd-ilibrary.org/science-and-technology/building-and-sustaining-collaborative-platforms-in-genomics-and-biobanks-for-health-innovation_11d960b7-en

Genomics England. (2022). *Genomics England launches next-generation research platform central to UK COVID-19 response*. Retrieved from https://www.genomicsengland.co.uk/news/research-environment-covid-19-lifebit-aws

ISO. (2022). *ISO/IEC 27001 INFORMATION SECURITY MANAGEMENT*. Retrieved from https://www.iso.org/isoiec-27001-information-security.html

National Cyber Security Centre. (2022). *About Cyber Essentials*. Retrieved from https://www.ncsc.gov.uk/cyberessentials/overview

National Cyber Security Centre. (2022). *Preventing Lateral Movement*. Retrieved from https://www.ncsc.gov.uk/guidance/preventing-lateral-movement

NHS Digital. (2022). *Trusted Research Environment service for England*. Retrieved from https://digital.nhs.uk/coronavirus/coronavirus-data-services-updates/trusted-research-environment-service-for-england

Observational Health Data Sciences and Informatics. (2022). *OMOP Common Data Model*. Retrieved from https://www.ohdsi.org/data-standardization/the-common-data-model/

UK Data Service. (2022). *What is the Five Safes framework?* Retrieved from https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/

UK Government. (2022). *Code on genetic testing and insurance*. Retrieved from https://www.gov.uk/government/publications/code-on-genetic-testing-and-insurance

UK Health Data Research Alliance, & NHS X. (2021). *Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems*. Retrieved from https://zenodo.org/record/5767586#.YvOwc3bMKUk

Wilkinson, M. D. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Nature Sci Data*.