Policy briefing

PHG FOUNDATION
making science work for health

The federation of trusted research environments for genomics and health

The problem of data sharing at scale

Data sharing can be a laborious and bureaucratic process for health researchers and is fraught with technical, legal and practical challenges. The *Goldacre Review* in 2022 provided a critical reframing of the way that data should be shared by recommending the use of "trusted research environments" which reorient the data sharing process away from direct transfers of data to strictly controlled remote access.

However, a significant drawback of this model is the added complexity that researchers face if they need access to large-scale and multimodal data which sits across multiple different research environments.

Terminology

Several terms are used interchangeably in the literature to encapsulate privacy enhanced data sharing spaces such as “trusted research environments”, “data safe havens” and “secure data environments”. Whilst there are some contextual differences i.e., trusted research environments are predominantly used in research only spaces, they are by and large synonymous.

Note also that the appropriateness of the term ‘trusted’ has been heavily debated in the literature which has also contributed to the creation of these other titles.

The challenges of having multiple entities with their own data sharing environments include:

- **Complex negotiation:** Researchers may need to negotiate access to multiple environments and complete several ethics approval processes whilst undertaking analysis according to the set technical functionality that each enables. These can take months to complete.

UNIVERSITY OF CAMBRIDGE

- **Closed environments:** To enhance data security, data movement in and out of these environments is generally restricted, making analysis across environments challenging.
- **Differing functions and governance restrictions:** Their differing functionality may also mean that even if extraction is permitted, results are presented in differing formats meaning that combining the results of each analysis could be a costly and time-intensive process.
- **Environmental impact:** In genomics and genome sequencing, scaling up health data production is also driving exponential growth in data storage requirements, which can reach into the petabytes of data storage exacerbating the known CO2 impact of server farms.

Therefore, despite the huge advances already made by the introduction of the trusted research environment model, data sharing at scale is insufficiently future-proofed.

One proposed solution is "federation". Whilst trusted research environments are not yet widely federated ("federated environments") we outline what federation is and why it may be necessary to future proof genomic and health data sharing.

What is federation in the data context?

The literature identifies three definitions of what is meant by federation in the data context:

1. **Metadata model:** Defined as creating a database that makes the trusted research environments' associated metadata centralised and searchable, while the infrastructures of the individual research environments remain isolated from each other.
2. **Federated infrastructure:** National or international scale infrastructures that are linked by centralised nodes that oversee research sharing between different virtual environments in the network. Take, for example, the European Health Data Space's HealthData@EU. Several initiatives are looking at federation at differing scales i.e., national and international trusted research environments (see Table 1 below).

Table 1: Examples of national and UK/EU-wide initiatives

Examples of national initiatives	Examples of UK or EU-wide initiatives
Honest Broker Service Northern Ireland	European Health Data Space (not just for research)
Scottish National Safe Haven	ELIXIR Data Platform (across Europe)
NHS England National Secure Data Environment	The Secure Research Service (SRS)
SAIL DATABANK (Wales)	Health Data Research Innovation Gateway
	Data and Analytics Research Environments UK (DARE)

3. **Federated data analysis:** No data is copied or moved, instead researchers have virtual access to several environments at the same time by 'using software that enables numerous databases to work as one'.[1] This third model is more sophisticated than a purely "federated infrastructure" in the sense that this additional software enables researchers to analyse data across multiple disparate data environments at the same time. Table 2 provides some real-world examples.

Table 2: Real world examples of federated data analysis models[2]

Real world examples of the federated data analysis model for genomics	About
Genomics England and NHS England's collaboration during COVID-19	This collaboration enabled large-scale genomic and medical data analysis in the search for why some patients were more susceptible to COVID-19.
Genomics England (GEL) and the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC)	In a UK first, both institutions' trusted research environments were federated to enable secure analysis across these environments without moving the original data. This collaboration arose out of the *DARE UK Sprint* project.
Greece's national newborn sequencing program, "First Steps"	Federated analysis is helping to link global cohorts of data for secure analysis in Greece for their newborn sequencing program using Lifebit's federated technology. They aim to detect around 400 early onset and actionable genetic conditions in newborns.

Case Study: Genomics England and University of Cambridge

Under this collaboration, federation meant that a researcher could query data held across two separate trusted research environments to find individual records and create a group of interest i.e., a group with certain characteristics such as a specific genetic mutation.[3] A set of application programming interfaces (APIs) i.e., software, were produced and tested to enable the trusted research environments to talk to each other. Once complete, analysis could then be undertaken across this group. Finally, to retrieve the results, '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 checks that reidentification of individuals is not possible.'

Benefits of federated data environments for genomics

All these understandings of federation will enhance the capacity for genomics to generate new insights and future proof genomics research by increasing:

- **Scalability:** Whether it is by linking environments through centralised nodes, through metadata that makes data repositories more searchable or via software that enables researchers to concurrently run analysis across several environments, federation in any of these senses better facilitates research capability and capacity.
- **Cost-effectiveness:** Federated environments are also cloud-based meaning that researchers will only pay for what they need. This eases economic disparities to an extent and possibly opens the "market" to smaller research cohorts to create their own secure data environments or access others on the assumption that access fees are fair.
- **Environmental sustainability:** Using cloud-based infrastructures enables many to use and access valuable data sources and as such, reduces the need for every single research cohort to copy, duplicate, download and/or store vast amounts of data. Consequently, this model also reduces the CO2 impact of server farms.
- **Feasibility:** This unique combination of cloud technology, data environments and APIs enable cross-environment analysis at a scale and in a way that has not been possible before.

However, whilst all three enable greater searchability and accessibility than non-federated trusted research environments to differing extents, it is the third model which arguably enables the features that facilitate and meet the analytical demands of future genomic research (i.e., multimodal, large-scale, cross-environment data analysis). The rest of this briefing will therefore focus on analysing the "federated data analysis" model.

Challenges of federated data analysis

Application programming interfaces can and have been used to provide the necessary infrastructure for cross-environment analysis. However, to better streamline this process, further factors need to be considered. These include mitigating the variation in how data are stored and curated, and whether there is anything novel about this infrastructure that gives rise to new privacy and security risks. We outline some of these considerations below and note some of the efforts that have been taken to address them.

The need to reduce variation

Common data models

Federated data analysis will require homogenisation of how data are described and stored across genomics and health data research environments. Such environments are increasingly interdisciplinary and researchers may need to combine health, genomic and other data from varying sources and specialisms. To do this, such environments may need to adopt common data models (CDMs) to ensure interoperability for APIs drawing data from differing environments and sources. The Observational Medical Outcomes Partnership (OMOP) has created a common data model to harmonise disparate genomic data sources by standardising the data format and vocabulary.[4] However, whilst these models may mitigate such variation, we are yet to know just how feasible this will be. There will be limits to just how much each disciplines variables can be homogenised due to their differing requirements and nuances.

No or low code tools

Not all researchers will have the necessary software language literacy to run certain analyses in and across these environments. In answer to this, some trusted research environments offer no or low code tools that provide a customisable user interface for researchers to “create” tools that facilitate their analysis without needing to code these applications themselves. A common example of this are 'beacons' which is an interface that enables researchers to query the data and to be given a response in a 'yes' or 'no' format.

The 'Beacon Network' is a federated form of this where genetic researchers can discover data on genetic mutations across several genetic datasets held globally. There are also limits on the number of questions that can be asked to ensure statistically safe responses that ensure privacy and data protection considerations are upheld.

Harmonisation of governance standards

To use software that enables federated analysis, governance standards will need to be streamlined where possible to reduce the governance burden that can impede research. In the UK context, some have already sought to tackle these challenges such as the UKRI funded Data and Analytics Research Environments UK (DARE UK) programme, HDRUK and the TRE Community. This task will be even more challenging at international scales considering the varying accreditation processes adopted in different countries and increasing geopolitical tensions over genetic and genomic data, which is viewed as both a potential national security threat and asset.[5]

Attitude shift

As some level of harmonisation is required to achieve federation, researchers may have to accept greater restrictions on how they would ordinarily undertake their research, which may impede its uptake. Researchers may therefore need encouragement to accept greater limitations on how they conduct their research (i.e., moving away from traditional methods of data sharing) for the promise of the greater reward offered by federation or risk not having access to the data at all.

Emerging technology in these environments

Advances in technology have continued at a rapid pace in the areas of artificial intelligence; supercomputing power, cloud and edge computing; in data generation (e.g., synthetic data); and privacy enhancing technologies (e.g., differential privacy and privacy accounting). Keeping abreast of these intersecting technologies within these environments will be a constant challenge for privacy specialists and regulators.

Many data innovations are purportedly privacy-enhancing but in such increasingly linked environments, with ever-growing variables for privacy specialists to consider (e.g., more easily accessible and linked data, advanced technologies that can makes these links with increasing ease, new forms of cybersecurity attacks etc.), proving that these environments and their data are sufficiently secure will be a near impossible task.

There is plenty of evidence from the roll out of previous data initiatives that lack of consultation and data breaches can have a disproportionately chilling effect on research and innovation. Drawing lessons from this experience may help to shape the evolution of federated data environments in a way that secures the greatest benefit achievable while maintaining an acceptable level of security and enjoying public confidence.

Regulatory and liability challenges

Federation provides a novel infrastructure unlike what has previously been available to researchers, and consequently may challenge how liability and responsibility for data protection breaches are determined.

Whilst the software capacity is there to enable, for example, beacon software which utilises externally facing APIs across several environments concurrently,[6] it is unclear whether these software capabilities could change how data protection impact assessments need to be undertaken. As these environments and their capabilities are linked, it may be unrealistic to assume that individual TREs can sufficiently assess reidentification risk in light of the increasing volume and accessibility of data in wider circulation.

It is also unclear how determining liability for data breaches, data protection responsibilities and what amounts to misconduct will be determined – it may be that a range of actors will be considered 'joint controllers' under data protection law. However, such considerations are highly complex and further guidance will likely be necessary.

Merging care and research

The literature focusses on the use of trusted research environments for research. However, similar environments are also likely to be used in the context of clinical care, as is the case under the European Health Data Space model. The emergence of such spaces demonstrates that the traditional divide between care and research is becoming more blurred and increasingly so since technologies, such as wearable devices, have meant that lifestyle data can be used to inform research and enable innovations such as virtual wards. This blurring is sometimes referred to under the rubric of learning health systems, which marks out another emerging context into which trusted research environments may become incorporated.

Conclusion

In just two years since the *Goldacre Review* was published, huge advances have been made in data technologies and sharing environments. Federated data analysis is a vital development for genomic research. It provides a way of overcoming the notable challenge of analysing data held across several trusted research environments.

However, as with any innovation, new challenges may arise. The literature suggests that there is variation in what is meant by 'federation' and differing degrees of federation exist. The risks presented by emerging technologies in these environments and how regulators think about liability, responsibility and reidentification risks may also need to be reconsidered.

References

1 Lifebit, 'Four key requirements to enabling federated data analysis' (Lifebit Blog, 16 June 2023). Available at: <https://www.lifebit.ai/federated-data-analysis/four-key-requirements-to-enabling-federated-data-analysis> accessed 30 August 2024.

2 Lifebit, 'How is federated data analysis boosting genomics research?' (Lifebit blog, 19 July 2023). Available at: <https://www.lifebit.ai/federated-data-analysis/boosting-genomics-research-federated-data-analysis> accessed 30 August 2024.

3 DARE UK, 'Multi-party trusted research environment federation: Establishing infrastructure for secure analysis across different clinical-genomic datasets,' (DARE UK blog, 25 October 2022). Available at: <https://dareuk.org.uk/multi-party-trusted-research-environment-federation-clinical-genomic-datasets/?utm_source=twitter&utm_medium=social&utm_campaign=dare_uk_genomic_federation_blog> accessed 30 August 2024.

4 Lifebit, 'Four key requirements to enabling federated data analysis' (Lifebit blog, 18 June 2023). Available at: <https://www.lifebit.ai/federated-data-analysis/four-key-requirements-to-enabling-federated-data-analysis> accessed 30 August 2024.

5 See for example: Hansard Report, 'Genomics and National Security' (UK Parliament Hansard Reports, Volume 729, 8 March 2023). Available at: <https://hansard.parliament.uk/commons/2023-03-08/debates/3F7E5903-596F-492A-B130-A4503928CA7F/GenomicsAndNationalSecurity> accessed 26 July 2024; The White House, 'President Biden Issues Executive Order to Protect Americans' Sensitive Personal Data' (28 February 2024). Available at: <https://www.whitehouse.gov/briefing-room/statements-releases/2024/02/28/fact-sheet-president-biden-issues-sweeping-executive-order-to-protect-americans-sensitive-personal-data/> accessed 26 July 2024.

6 UK Health Data Research Alliance, 'Building Trusted Research Environments. Principles and Best Practices; Towards TRE Ecosystems' (HDRUK Report, 8 December 2021). Available at: <https://www.ed.ac.uk/sites/default/files/atoms/files/5_safe_principles_-_building_trusted_research_environments.pdf> accessed 19 September 2024.

Author: Dr Elizabeth Redrup Hill

Published: October 2024

UNIVERSITY OF CAMBRIDGE