DARE UK

DARE UK Phase 1 Sprint Exemplar Projects: Reflections and Lessons Learned

June 2023

UK Research and Innovation

HDR UK Health Data Research UK

ADR UK Data-driven change

DARE UK Phase 1 Sprint Exemplar Projects: Reflections and Lessons Learned

Table of Contents

1. Introduction

1.1. Background

The DARE UK (Data and Analytics Research Environments UK) programme is funded by UKRI (UK Research and Innovation) as part of its Digital Research Infrastructure portfolio of investments, which supports the development of a coordinated vision for digital research infrastructure in the UK. Part of the DARE UK programme's ambition is to enable the development of a national interoperable network of secure digital research infrastructures or Trusted Research Environments (TREs) that lays the foundation for the next generation of secure, cloud-enabled environments for advanced data research for the public good.

In 2022, as part of Phase 1 of the DARE UK programme, UKRI provided nine inter-disciplinary research teams with over £2 million in funding to deliver a portfolio of Sprint Exemplar Projects with the purpose of investigating the potential for the development of a joined-up and trustworthy national data research infrastructure for sensitive data research for the public good. These projects were the result of an open call for proposals that ran from September to November 2021. A total of 25 applications were received, and nine successful projects were chosen based on excellence, novelty, and diversity by an independent panel of experts, including members of the public. The projects ran for eight months, from January to August 2022.

The funded projects were intended to cover different research domains and UKRI research council remits and aimed to

1. investigate potential scientific use cases for joining up data from different sources,
2. demonstrate technology approaches,
3. explore best practices, particularly around governance and ethics, and
4. include Public Involvement and Engagement (PIE) activities

All within the context of the DARE UK programme's vision for cross-domain sensitive data research.

1.2. Purpose of this report

The purpose of this summary report is to provide a short, non-exhaustive overview of some of the strategic reflections from the nine Sprint Exemplar Projects. The report will summarise some of the key problems that the projects aimed to address, highlight some of their progress and challenges against their objectives, and how these fit together as a portfolio that informs future decisions on the development and implementation of a national data infrastructure for sensitive data research.

1.3. Overview of the nine Sprint Exemplar projects

There is a thriving - and growing - network of TREs already in existence, but interoperability across them is hampered by a lack of an agreed set of standards and protocols. Given the foundational requirement for some level of standardisation to enable ecosystem join-up, the majority (seven out of nine) of the Sprint Exemplars focused on testing different solutions to challenges around standardisation across the three foundational levels of infrastructure, data, and governance.

At the level of infrastructure, this was typically achieved by taking an 'infrastructure as code' (I-a-C) based approach to enable rapid stand-up of a TRE (*TREEHOOSE, 'Virtual' TREs*) onto cloud infrastructure using a technology stack typically assembled from the existing open source or off-the-shelf (proprietary) technologies.

These projects demonstrate clear benefits to a cloud-based approach, which could simplify secure network connectivity across the TREs and equalise access to computational resources through the flexible and scalable deployment of computing resources across the network, as well as enabling much fuller audit and management reporting.

Even if TREs can be joined up securely, ensuring data held within them is suitable for analysis requires 'data harmonization', which is the process of integrating and reconciling data from different sources or formats to ensure their consistency, compatibility, and coherence. These activities are typically very resource intensive and challenging, especially when considerable governance barriers must first be overcome to access the data for any manual curation. One potential solution to this problem is standardisation to a Common Data Model (CDM), an approach that establishes a uniform structure and format for data across multiple systems or sources according to a standardised framework. Two of the Sprint Exemplar (*Federated Genomics*, *FED-NET*) projects tested a popular CDM called the Observational Medical Outcomes Partnership (OMOP) CDM to standardise their data, highlighting instances where this process works well and where alternative solutions (particularly for cross-domain federated analytics) are required.

After data has been successfully linked and analysed, there is still a need to ensure that the privacy of individuals is preserved in any outputs from any (federated) research. Therefore, two other Sprint Exemplar projects were concerned with standardized protocols and best-practice for the assessment and management of privacy risk for both analytical as well as emergent predictive AI/ML types of research outputs. More generally, there is growing consensus (including from Phase 1 DARE UK recommendations, the Goldacre Review, the UK Health Data Research Alliance TRE Green paper, and the DHSC Data Saves Lives policy paper) that all sensitive data should only ever be accessed and analysed by researchers within a TRE. As TREs expand or proliferate to accommodate the increasing amounts of valuable, sensitive data available, and this data becomes amenable to state-of-the-art analytical and predictive methods, the types of people that use TREs will also need to grow. These include 'non-traditional' TRE users, such as third and private sector researchers, as well as AI researchers, which has the potential to present new governance issues. The *TRE Governance for Non-traditional Users* project focused on how the Five Safes framework could be developed to include such non-traditional users in a way that is acceptable to data owners and the public.

Finally, central to all nine of the Sprint Exemplars was Public Involvement and Engagement (PIE), which were mandated in the DARE UK Sprint Exemplar funding call to ensure that the views of the public fed into the design and the delivery of each of the projects. While it is increasingly recognised that public involvement and engagement is critical to ensure the ethical and trustworthy use of data, involving the public meaningfully in research using sensitive data remains challenging. For example, engaging with and recruiting public participants from harder-to-reach populations, such as children and young people (especially those from disadvantaged groups), is essential, particularly when these groups could potentially benefit the most from research done using their data. The *STEADFAST* Sprint Exemplar project explored the best ways to inform, engage and involve young people, their families and the wider public in important issues around the use of their sensitive data for research.

Among their findings was the importance of engaging with established and trusted community groups and charities for successful outreach and recruitment, particularly for the most disadvantaged groups, a finding which was also corroborated by the *FAIR TREATMENT* project's PIE activity reflections. Although the project specifically took children and young people with diabetes as its use-case, the public engagement toolkit that was produced by the project also informs and supports best practices for public involvement of children and young people with other health and social needs and builds upon a growing body of research to support more meaningful public engagement in research using sensitive data. The companion report "Public Involvement and Engagement in Sensitive Data Research: Lessons Learned from the DARE UK Sprint Exemplar Projects" provides a summary of the different sprint project PIE activities and lessons learnt from them as a collection.

Table 1. Sprint Exemplar Project summaries

#	Project Name	Lead Organisation	Main Project Objectives
1	Federated Genomics	University of Cambridge	• UK-first demonstration of federated analysis of genomic data in the UK by creating a secure connection between the Trusted Research Environments (TREs) of the NIHR Cambridge Biomedical Research Centre and Genomics England, allowing for joint analysis of consented clinical-genomic data from patients in both environments, without moving any original data.
2	FED-NET	University Hospitals Birmingham	• To develop and test a federated analytical environment that can combine data from diverse sources (clinical, meteorological, translational, and environmental pollution data) to enable federated analysis, using a study of asthma as proof of concept.
3	FAIR TREATMENT	University of Cambridge	• To demonstrate proof-of-concept federated analysis and predictive model training for early detection of mental health issues in adolescents using health, social care, and education data held in different trusted research environments while preserving individual's privacy.
4	TREEHOOSE	University of Dundee	• To develop and release open-source software and supporting documentation ('infrastructure as code') for building and operating standardised TREs on public cloud infrastructure. • Proof-of-concept for Secure Enclave (cryptographically secure environment) capability within the TREEHOOSE TRE, enabling proprietary algorithms to run within the environment without revealing any sensitive IP.
5	'Virtual' TREs	Francis Crick Institute, in partnership with BT, the Institute of Cancer Research and the Rosalind Franklin Institute	• Determine requirements from different stakeholders of what a 'virtual' TRE should be able to do to overcome existing barriers to research involving cross-domain collaborations. • Creation of a cloud-based technology stack to enable stand-up of 'on-demand' TREs, with novel features such as: i) experiment-level budget control, ii) audit and real-time management reporting, and iii) standardised (HIPAA and ISO27001) security compliance.
6	PRiAM	University of Southampton, University of Warwick and Privitar Ltd	• To develop a standard privacy risk assessment approach for cross-domain access and reuse of sensitive data for research purposes.
7	GRAIMatter	University of Dundee	• Examine the risks associated with using artificial intelligence (AI) and machine learning (ML) to train models on sensitive personal data and develop methods and tools to assess and mitigate these risks from ML models as outputs from TREs
8	TRE Governance for non-traditional users	University of Edinburgh, DataLoch team	• Address the technical and governance barriers to the use of Trusted Research Environments (TREs) by non-traditional researchers, for example, from the third and private sectors.
9	STEADFAST	Cardiff University and Diabetes UK	• Explore the best ways to inform, engage, and involve young people with chronic health conditions, their families, and the wider public in important issues surrounding the use of sensitive data for research.

2. Strategic Reflections

2.1. Use cases

2.1.1. The Sprint Exemplar use cases give a flavour of the enormous scientific potential that federation can enable

The incidence and morbidity of many diseases that constitute a large proportion of the global disease burden (including cancer, cardiovascular disease, mental health conditions, diabetes, and many others) often depends on complex interactions between genetic, behavioural, and other environmental factors. Although there has been exponential growth in the collection, and increasingly the availability for research, of 'big' data that can quantify different components of these factors in unprecedented detail, the research potential of these data when analysed in isolation is limited. It is only when this data is linked together that it becomes possible to gain deep insights into the complex interactions that drive disease incidence and outcomes, which in turn can inform policy and interventions that reduce their prevalence and severity.

Although the power of data linkage for research (especially in the social and economic sciences) is beginning to be realised through the growth in the availability of a small number of national, pre-linked datasets (for example, the Longitudinal Education Outcomes dataset, and the ONS' Public Health Research Database) which are accessible through single TREs such as the ONS Secure Research Service, combining and storing data in a single environment in this way is often not feasible for governance reasons (due to different legal or data protection policies for different types of data) or technical reasons (particularly for very large datasets including environmental, imaging, and genomics data whose size makes them challenging to move).

These barriers can make moving and storing this data in a single environment costly and time-consuming, and different kinds of data often require specialist tools for analysis that may not be easily accessible in all TREs. One alternative approach to moving and combining data within a single environment for analysis is instead to simultaneously query data in different TREs and then combine the results from these analyses, typically in a separate secure environment. This not only eases the federated analysis of bulky datasets but also presents a solution that may be more acceptable to data owners and the public because the data itself never needs to move, only the (usually less sensitive) federated query or algorithm and their aggregated outputs.

Three of the Sprint Exemplar projects explored the feasibility of this approach, highlighting just some of the potential scientific use cases that such a federated solution can provide, in these instances, for insights across different diseases. The *FED-NET* project focused on combining analysis of cross-domain data from administrative health records, translational data, and environmental (e.g., meteorological and pollution) data for insights into how these interact to explain the incidence of acute hospitalisations related to asthma.

The *FAIR TREATMENT* project looked to enable the joint analysis of health, social care, and education data for children and young people for the identification of early markers of mental health disease in adolescents to support earlier intervention. In contrast to the federated analytics of cross-domain data of these two projects, the *Federated Genomics* project was a UK-first demonstration of the joint analysis of very large (and therefore difficult to move or copy) clinical genomics (e.g., cancer) datasets. Common to all these use cases is that none of the intended analyses would have been feasible in the absence of a federated solution.

2.1.2. Different use cases require different solutions for federated analysis: adoption of Common Data Models can support federated analytics, but key challenges remain, particularly for cross-domain data analytics

Two of the Sprint Exemplar projects tested the use of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), an increasingly popular CDM designed for the standardisation of health informatics and genomics data, to facilitate the simultaneous running of queries (federated analysis) on data held across multiple TREs. The *Federated Genomics* project found that transformation of the data held within the Genomics England and the Cambridge TRE into OMOP prior to analysis was essential for successful federated analysis, and used ETL (Extract, Transform, Load) pipelines to automate the large-scale raw data conversion to the standardised, analysis-ready data without the need for data egress.

By contrast, the *FED-NET* project, which included both health as well as meteorological and Index of Multiple Deprivation (IMD) data, mapped the limitations of the OMOP CDM vs. use of native languages across their cross-domain data. They found that only just over half of the FED-NET asthma data fields could be readily transformed into OMOP, and that around a quarter of the data (corresponding to fields from the meteorological datasets, which OMOP wasn't designed to handle) could not be held in OMOP at all even with additional wrangling – highlighting the challenges around adoption of a CDM or standards across data from different domains for federated analytics. In this instance, considerable manual curation to derive bespoke schema and labelling for matching, which requires extensive experience in both wrangling and anonymising complex data, may be more amenable to approaches where the data itself needs to be brought together. It is therefore important that a federated ecosystem of TREs should enable both federated analytics (where data stays put and queries/aggregated results only move), but also support data pooling where data can be brought together into a single environment (in some cases only temporarily) where the need requires.

2.2. Technology demonstrators

2.2.1. 'Infrastructure as code' model brings progress towards an open-source reference TRE architecture for TRE interoperability and lowers the barrier for new TREs to join a federated network.

As the case for the universal adoption of TREs as the mechanism for opening up access to sensitive data strengthens and the collection of data (as well as computational methods to analyse this data) proliferates, the TRE infrastructure to store and make this data available to access for research for the public good will also need to grow. This will likely occur through some combination of the expansion of existing TREs to accommodate more and different types of data and analyses, as well as the emergence of new TREs with those capabilities.

Building such infrastructure on-premises (using local IT infrastructure) can be both technically challenging and costly while making the barrier to entry for new TREs prohibitively high. In addition, bespoke solutions can hamper interoperability, where variation will inevitably arise as different TRE builders opt for or are constrained by different local capabilities. One potential solution to this problem is the adoption of an 'Infrastructure as code' (I-a-C) approach, delivering TRE configuration and provision as software. Defining and managing TRE deployments programmatically (powered typically using virtualisation technology common in the public cloud – the so-called "cloud-first" approach) can make it easier to implement, provide greater application versatility, and facilitate scalability.

Both the *'Virtual' TREs* and the *TREEHOOSE* project developed and tested this I-a-C approach, demonstrating a massive potential reduction in the amount of time and expertise needed to stand up a TRE from several weeks to a matter of days or even hours 'on-demand', typically using technologies that work straight 'out-of-the-box'. For example, the open-source *TREEHOSE* TRE can be deployed automatically via I-a-C while still providing the ability for the TRE manager to tailor the template code to suit their particular TRE use cases, demonstrating the flexibility of software-based infrastructure deployment.

The importance of developing and adopting 'out-of-the-box' technologies was emphasised in several of the Sprint Exemplar projects, to enable greater replicability and interoperability between different TRE implementations - an important requirement for successful federation across them. Additionally, a number of the projects highlighted an emerging need for a community-driven consensus on a common set of TRE minimum standards or a reference architecture - precisely the objective of the DARE UK-funded SATRE project.

2.2.2. Cloud technology can equalise access to scalable computational resources, and simplify auditing across a federated network

Even if TRE infrastructure can be flexibly implemented in software, there is still a need to provide the computational resources to power the storage environment and analytical capability of the infrastructure. The Sprint Exemplar technology demonstrators focused on the opportunities provided by powering TREs on public cloud infrastructure, highlighting several compelling advantages to adopting cloud-based technologies to support TRE federation. The cloud-first approach to I-a-C brings advantages to both public cloud and on-premises deployments, but use of a public cloud platform enables the scalable and elastic deployment of computational resources where needed across the network. Because of the economies of scale this provides through the outsourcing of this to an external provider, where the user only pays for what they use, this can help to equalise access to computing and storage environments for smaller or less digitally mature TRE operators that might lack the in-house expertise required or face prohibitive start-up costs in standing up this infrastructure on their own.

This in turn can lower the barriers to the formation of research collaborations across different institutions, which can be hindered by high friction in bringing together different local digital infrastructures (in addition to the knottier problems of different domain expertise, ethics, and governance adherence requirements). This can be compounded if resources, both in terms of infrastructure and computational resources, as well as grant funding, are unevenly distributed among research consortia members. Adoption of a cloud-based solution flexibly equalises access to these digital infrastructure resources across the consortia, especially if controls for billing for the project-specific use of those resources can be supported (as demonstrated in the *'Virtual' TREs* project).

A network of TREs sharing standardised cloud infrastructure can also bring other important capabilities from the perspective of a federated network, most strikingly through fuller audit and management reporting enabled by the detailed, automated logging of metadata available by default in many cloud-first technologies. This information can include timestamped user and access logs of what data is being accessed, by whom, and what queries are being run against that data for what purposes. Not only does the collection of such detailed metadata increase security by enabling algorithmic identification and reporting of any activity that is deemed "risky", but it also opens the door for greater transparency for patients and the public about how their data is being used, by whom, and for what purpose, potentially on-demand and across the entire federated network.

A proposal during the PIE workshop in the *'Virtual' TREs* project to potentially take this one step further and adapt the technology platform to create a 'patient data account', much like a bank account, that could log the real-time use of individuals' data, was met with enthusiasm from public participants. Such a system could also support the tracking of any research outputs using individuals' data that evidences the benefits that providing their data has brought, providing a powerful incentive for individuals to do so. Although the realisation of adapting Sprint Exemplar technology to enable such capability is some way off, it nevertheless illustrates a very compelling use case for the power of this sort of technology to provide individuals with unprecedented transparency in the use of their sensitive data.

2.3. Best Practice

2.3.1. A standardised risk assessment framework will expedite and support consistency in governance decisions on risk management

The Sprint Exemplars illustrated some of the tremendous scientific potential for understanding complex disease through the linkage or federated analysis of cross-domain data, but bringing together data or analytics in this way also presents new privacy risks, especially when combined with the application of novel methods (such as ML/AI) that have much less well-understood risk profiles. In particular, the increased complexity of data flows in federated projects, expansion of the number and types of different stakeholders, and the different types of research collaborations required for federated projects (including multi-disciplinary teams) all present potential additional privacy challenges that need to be assessed and effectively managed.

Defining and assessing privacy risk, however, remains somewhat subjective. For example, the Five Safes framework developed by the ONS has emerged as a simple and powerful tool for categorizing different high-level requirements for providing access to sensitive data, centred around ensuring Safe People, Safe Projects, Safe Data, Safe Outputs, and Safe Settings. These categories are inherently subjective and not binary, and data controllers must decide for themselves whether their assessment of the degree of safety in a particular category satisfies their appetite for risk. This in turn can impose higher governance barriers for enabling the use of linked data if different data owners assess risk differently (since the approval of all data owners is required).

Two of the Sprint Exemplar projects aimed to address some of these issues, with the *PRiAM* project developing a standardised risk assessment framework based on the Five Safes as well tools to facilitate the automation of risk assessment, and the *GRAIMatter* project focusing on evaluating existing or else developing new methods, tools, and training for assessing privacy risk posed by releasing trained 'black box' ML models outside of TREs. *PRiAM*'s standardised risk tier classification framework was designed to help decision-makers, patients and the public assess and manage privacy risk, particularly for newer forms of data access and reuse, such as federation.

The adoption of this risk assessment framework by data controllers stands to bring several important benefits. Firstly, the framework will enable data custodians to explicitly define the criteria that they apply when assessing risk, which can then be tied into defining at a granular level discrete levels of assessed risk that can form the basis and justification for decision-making. This will not only increase consistency but will also increase transparency in the decision-making process, demonstrating trustworthiness to the public by increasing the visibility of the measures that have been taken to protect individuals' data. Secondly, the use of a standard and repeatable approach for defining and assessing risk should expedite data owner decision-making, particularly in commonly occurring or routine instances (which may also be amenable to more automated approaches), which is a critical bottleneck for the timely approval of projects to use data for research.

Importantly, however, such a framework by itself is insufficient to solve inherent challenges in the interoperable assessment of risk across different domains, especially within the context of uncertainties around the legal and regulatory guidelines on the use of federated data (for example on who becomes the controller of a linked dataset). However, such a framework does present an important starting point for a more systematic assessment of risk if it can be prototyped and tested by a set of data providers operating across a network of TREs.

2.3.2. Progress towards methods to assess privacy risks posed by novel research outputs, such as ML models

Output checking is a critical component of a TREs role in protecting the privacy of individuals from the use of their data by ensuring that individuals cannot be identified from the research outputs that leave the secure environment, typically through the application of statistical disclosure control methods (for example, to identify and prevent the egress of aggregated research outputs with potentially disclosive small group sizes).

These methods are well understood for traditional statistical analyses carried out on sensitive data, but very little is known about the potential for attackers to query trained ML models to identify individuals whose data was used to train them. As the scientific motivation to apply powerful AI and ML methods on large-scale sensitive, linked data grows (including by Sprint Exemplar projects, where *FED-NET* and *FAIR TREATMENT* both looked for capabilities to apply such methods for the early detection of disease), the need to mitigate against any privacy risk posed by any outputs from these methods becomes paramount.

Although much more research is needed to continue to develop and refine methods to both quantify and control risks posed by releasing trained ML models 'into the wild', the GRAIMatter project has made considerable progress in mapping different types of risk posed by releasing different types of trained models under different circumstances, as well as developing methods to start quantifying the degree of re-identification risk under these different conditions and how to protect against this.

This is of critical importance because without such protections TREs are unlikely to provide the capabilities in their environments to use such methods, which would be a missed opportunity. Finally, output checking requires considerable resourcing from TRE operators, and is a bottleneck to the timely publication of research results. The DARE UK-funded SACRO driver project is looking to ease this burden through the development of tools to enable semi-automated output checking, to enable more efficient output-checking while still ensuring the privacy security of research outputs.

3. Conclusion

Strikingly, the successful Sprint Exemplar projects were overwhelmingly partnerships with members across multiple public and private organisations, which were assembled to bring together the required infrastructure, expertise, and data together, and without exception included public involvement throughout to ensure alignment to public priorities and concerns. Given this demonstrated need for partnership working, the DARE UK programme launched its Community Groups initiative in April 2023, following a similar model to the Research Data Alliance and IETF groups and offering endorsement and logistical support for groups to come together around solving problems in the sensitive research infrastructure space.

By themselves, however, partnerships are not sufficient for the development of a federated research infrastructure that can enable cross-domain research for the public good. There is also a critical requirement for the 'right purpose' to bring a partnership together, motivated by the specific benefits and the new science that becomes possible. The use cases demonstrated by the Sprint Exemplar projects, mostly around multi-factorial diseases such as cancer, mental health disorders, and asthma, give just a flavour of the potential for large-scale linked data to lead to a step-change in our understanding and predictive power around global challenges – from global pandemics, climate change, and growing social and health inequality - that can, in turn, drive evidence-based interventions and policy change.

Even with the 'right partnership' and 'right purpose', it is also essential that there is also the 'right TRE security' in place. This is to ensure that sensitive data is being used by the right people for the right purposes within an environment that is robust against both inadvertent and intentional potential security breaches and where the risk of identifying individuals from any outputs is minimised. As the *Federated Genomics* project explained, all three of these are required for the success of any endeavour to join up TREs for cross-domain data analysis.

These considerations will also remain front and centre in the planning for the incremental delivery of a growing federated network of TREs during Phase 2 of the DARE UK programme: first, identifying compelling scientific use cases to motivate the join-up of data or analytics across different types of sensitive data, finding the right partnerships to realise this, and all the while ensuring that the security of the infrastructure and its governance demonstrates to the public that the custodianship and use of their data for research protects their privacy and is used for purposes which are acceptable to them.

4. Acknowledgments

We are grateful to Balint Stewart from the DARE UK programme team for their valuable work on this report. We also appreciate Nour Mohanna from Carnall Farrar and Fergus McDonald from the DARE UK programme team for their contributions on an initial version of the document, which helped to inform the structure and content of this report.