

DARE UK



**FAIR TREATMENT: Federated analytics
and AI Research across TREs for
Adolescent MENTAL health**

**Dr Anna Moore, University of Cambridge, Dept of
Psychiatry**



**UK Research
and Innovation**

HDRUK
Health Data Research UK



ADRUK
Data-driven change

1. Executive summary

Globally, adult mental health problems continue to rise year-on-year. The onset of these problems is frequently in the first two decades of life, determined by a complex interplay of nature and nurture. Platforms including linked multi-agency data provide an important opportunity to understand the mechanisms of child mental ill-health and build digital tools to support early identification. However, accessing, linking and analysing such data is beset with significant technological and governance challenges. The public must also be engaged and find solutions acceptable. To address this, we worked with a diverse community of young people and parents to co-create technical and governance solutions. We integrated three existing technologies into an existing TRE provider to demonstrate one method of building a network of multi-agency data capable of privacy preserving federated analysis. Using synthetic data, we demonstrated its functionality. To support translation of this demonstrator to a working system utilising 'real' data, we worked with local information governance leads and experts to develop governance frameworks to support local data sharing and federation between sites.

We have developed recommendations describing how technologies can be deployed to enable federated analysis, including the ability to: automatically construct themselves based on an underlying data model and mirror this in their user and programming interfaces, automatically generate user interfaces and application programming interfaces (APIs) that allow efficient processing of queries not defined at the outset; support secure federated queries in a privacy-preserving fashion; enforce information governance controls, and fine-grained user authentication, and role-based access). These were developed/tested at scale using synthetic data generated using k-anonymous data derived from real data sources, leading us to recommend this approach to support rapid innovation.

Governance recommendations are: for the public to be meaningfully involved by co-creation along the data science pipeline from governance to interpretation and dissemination; that young people from 11y are able to meaningfully contribute; engagement groups must be diverse and should be recruited with third sector support; the ICO and HRA could helpfully align advice on the use of data for the purpose of creating research databases; and flexible 2-level federation frameworks are a means to enable analytics between TREs with differing local IG structures.

Next steps involve operationalising the governance frameworks, migrating to 'real' data, and scaling to federate with a non-AIMES provided TRE.

2. Introduction

Globally, four of the top ten medical disabilities are psychiatric and the prevalence of adult mental health problems continues to rise. The onset of these problems is in the first two decades of life and is determined by a complex interplay of nature and nurture. The most vulnerable and disadvantaged suffer the greatest and are least able to access help. Conventional, single-discipline research into the causes and treatment of child and adolescent mental health disorders has yet to yield a reduction in incidence or prevalence; rather, both are increasing. However, the evidence describing the mechanisms underlying mental ill health and resilience is rapidly evolving to reveal the role of biological factors such as (epi)genetics, immunology, inflammation, gut health and imaging. We need to understand how these interact with early life experiences and the environment, across diverse populations. This requires: (1) a system that is capable of providing better insights into the causes and triggers of illness and resilience, and (2) a platform that ensures that this knowledge can be translated into effective early identification and intervention tools, working within a system for child mental health service delivery. As its basis, this platform must include large, representative datasets of multi-domain data reflecting the relevant range of bio-psycho-social factors, to support research on how they converge to create resilience or susceptibility to psychopathology. The proposed data platform has strategic implications for predicting and preventing child

mental health problems early. However, there are particular challenges associated with accessing, linking and using these integrated datasets for analysis and predictive modelling. This includes governance, standardisation to a common data model, rapid data harmonisation, flexible querying, federated analysis between Trusted Research Environments (TREs) including multi-domain data, and high levels of security. Finally, it is important that the approach is acceptable to the public and patients, and that they have the opportunity to contribute to the development of any proposed solutions.

The FAIR TREATMENT project was designed to tackle these issues by building a demonstrator able to develop prediction models for child mental health using linked multi-agency data across a federated network. In each region a TRE including health, education and social care data was created, with the ability to carry out federated analysis across these. This provided the opportunity to describe how we tackled the challenges of (a) harmonising multi-agency data within a single TRE, (b) enabling privacy-preserving federated analytics across different TREs, (c) establishing a governance model to enable local and federated analysis, and (d) consulting with the public to understand the acceptability of the project and gather input into the development of the governance framework. The following report presents the findings from the technology and governance workstreams, and highlights what was learnt from public engagement work.

3. Technology demonstrator

Our approach was to integrate a mixture of existing open source and commercial technologies and deploy them to create three separate TREs in Cambridge, Birmingham and Essex. These were populated with synthetic data representing information collected by mental health, community, local authority, and social care services. Federated analyses were run over these platforms to demonstrate the ability to: (1) measure the incidence and prevalence of mood disorders and adverse childhood experiences in children, and (2) train logistic regression models predicting the incidence based on various risk factors. A video demonstrator of the technologies being used in practice can be found here <https://bit.ly/3QmfnS7> and a detailed description of the technical specification can be found in Appendix 7.1. We have uniquely brought together three existing technologies within a TRE infrastructure:

- **AIMES** provides the underlying infrastructure for the TRE, including multi-factor authentication and airlock mechanisms (<https://aimes.uk/tre/>).
- **CRATE**¹ (Cardinal et al., open source) provides de-identification tooling so that raw data can be pseudonymised at source before transfer into the TRE, and data without common identifiers can be accurately linked.
- **InterMine**² (Micklem et al., open source) provides an easy-to-use interface for standardising multi-agency data and for defining "Data Extraction Contracts" such that data for particular users is restricted to the desired subsets of data.
- **Bitfount**³ (<https://www.bitfount.com/>) enables federated analyses across the TREs, including SQL queries and machine learning tasks. The platform also provides privacy checks in the release controls, requiring any data leaving the TRE to have differential privacy at an approved level.

¹ **CRATE**: <https://crate.readthedocs.io/>; <https://github.com/ucam-department-of-psychiatry/crate>

² **InterMine**: <https://github.com/intermine/im-docs>. Code developed in this project is available: <https://github.com/intermine/intermine>. InterMine database: <https://github.com/intermine/camCHILDMine>

³ **Bitfount** <https://pypi.org/project/bitfount/>

Combining these technologies allowed us to create a novel environment which enabled multi-agency data linkage and federated analytics across several TREs. The AIMES TRE is designed to support arbitrary new services and so integration of additional components onto the AIMES TRE involved simple installation; no custom software was required. However, integration of Bitfount and InterMine together did require the development of several key new pieces of software. Firstly, a new type of Bitfount Data Source was created, supporting data in an InterMine service (previously only SQL databases and CSV files were supported). Additionally, new APIs were developed within InterMine for administrator- level access to list all the users' templates and associated data. This enabled the Bitfount service to act as the authorisation point for all cross-TRE analyses. The individual components within the TRE added the following functionality in order to support our use case:

- CRATE added support for de-identification via an application programming interface (API), additional features for free-text de-identification, and support for probabilistic de-identified linkage without a shared person-unique identifier (e.g. for linkage between health and social services data).
- InterMine added support for Bitfount integration, implementing new APIs for accessing and executing all the user templates and adapting an existing API to return the data types associated to the templates (Appendix 7.2).
- Bitfount added support for a new algorithm of federated SQL queries, with differential privacy disclosure controls.

The governance model that was developed for the federated analysis requires that each TRE controller could independently define which federated collaborations were to be allowed (see section 3 below). By bringing together these technologies, we have shown how it is possible for each data controller to:

- independently pre-process their data to pseudonymise it before adding it to the TRE,
- independently define the views of data that should be visible to users within their TRE, using a convenient graphical interface,
- independently develop suitable multi-agency data models for their data,
- independently define any pre-processing data transformations (separate from the federation),
- maintain independent access controls for direct access to their data,
- independently decide which external TRE users should be allowed to access the TRE data in a federated way, and
- run the approved federated queries, whether analytics or machine learning.

An important characteristic of this demonstrator is that the tools are all designed to work 'out of the box'. The next stage in the programme is to demonstrate that additional TREs could be added to the federated network to form a similar outcome with relatively little technical effort. The interfaces are user-friendly and do not require significant technical know-how to manage data access once the platform is set up. Table 1 provides an overview of our findings and recommendations.

Table 1: Overview of the principles, outputs and deliverables from the workstreams, and recommendations based on learning

TECHNOLOGY WORKSTREAM		
Principle	Outputs/deliverables	Recommendations
1) Data model driven systems	We demonstrated the value of automatically creating flexible high-performance APIs and user interfaces from a data model using the open source InterMine platform.	<p>Database systems should automatically construct themselves, as well as their user and programming interfaces, from the underlying data model. This is important because defining robust standards and data harmonisation is hard, will take time and will evolve; systems will thus have to be rebuilt many times as the standards and the data model evolves.</p> <p>This approach a) reduces the maintenance load associated with ongoing data harmonisation efforts and provides user interfaces and client library support for Python, R and other commonly used languages; and b) clearly separates the work for harmonisation from the work needed to build a useful data integration and query system.</p>
2) Query support for research data scientists	We demonstrated that it is possible to answer a complex query efficiently, spanning multi-agency data, of the type useful in a research context. This was enabled by exposing metadata and providing visualisations of the data structure and volume as well as ways to apply filters to the data.	Database systems should automatically generate user interfaces (UIs) and programming interfaces (APIs) that allow efficient answering of queries not known at the outset.
3) Federated analysis while	We enabled federated analytics/machine learning via Bitfount, using data supplied through InterMine APIs within TREs. Bitfount technology provides	TREs should support secure federated queries (within and across TREs), in a privacy-preserving fashion, to avoid the

<p>preserving privacy</p>	<p>privacy guarantees when presenting query results, via Data Extraction Contracts.</p> <p>Federated analyses functioned much as expected. Federated analytics and machine learning use cases were both successful. Differential privacy could be applied successfully before any disclosure.</p> <p>We learned that it is important for both analytical and machine-learning-based tasks to be available through the same interface. This is important because federated data cannot be directly seen by the data scientist, so data scientists in the federated setting need a mechanism to understand the data. We found that providing arbitrary SQL-based queries with differential privacy disclosure controls gives a good mechanism for doing this, while still ensuring that privacy is protected.</p>	<p>creation of data lakes and enable training of machine learning (ML) algorithms.</p> <p>Privacy-preserving queries should be available for conventional analytical purposes (including for data scientists to gain understanding of the data) as well as for ML.</p>
<p>4) Automatic enforcement of information governance constraints for analysis projects</p>	<p>We demonstrated InterMine-based "Data Extraction Contracts". These provide APIs, used by Bitfount, to permit flexible high-performance searches. Importantly, the contracts allow searches over only those data permitted by information governance requests.</p> <p>The technology we implemented provides scalability while preserving privacy through Bitfount federation over InterMine APIs: across multiple databases within one TRE and across InterMine databases in different TREs (e.g. Universities of Birmingham, Cambridge, Essex).</p>	<p>It should be easy to rapidly and securely discover and access data, within preset (and automatically enforced) information governance controls. The use of systematised and standardised information governance controls should become best practice, and "Data Extraction Contracts" are a useful concept for describing and defining access.</p>

<p>5) User access controls within and across TREs</p>	<p>Our pipeline provided fine-grained user-access controls to services (Bitfount API) and to defined data subsets within the TRE (InterMine API). For within-TRE analyses, it was possible to use a standard authorization system based on AIMES' implementation of Microsoft Active Directory. This enabled us to restrict within-TRE analyses to certain datasets as per standard practice. Cross-TRE analyses need a novel approach to authorization so that constraints could be set around the automated disclosure controls that would be required for users in other TREs.</p> <p>Bitfount's system for cross-TRE analyses suited this requirement well. The cross-TRE access controls could be set to require various controls around the data processing, for example a minimum required level of differential privacy. The challenge involved the integration of Bitfount's authorization system for cross-TRE analysis with AIMES' system for within-TRE access.</p> <p>For the Sprint project we simply decided to use AIMES's system for within-TRE analysis and Bitfount's for cross-TRE analysis. However, Bitfount, AIMES, and InterMine all support a standard protocol for authentication called OAuth and this could be used in future.</p>	<p>As well as project-level data-access contracts (4), systems should support corresponding fine-grained user access controls.</p> <p>The adoption of a common authentication protocol (such as OAuth) is recommended. This would allow a uniform authorization approach both for federation and for accessing the individual TREs.</p>
<p>6) Creation of synthetic data for pipeline testing</p>	<p>We aimed to create synthetic data conforming to multiple data standards for system development/ testing/ demonstration, and show its usefulness for cohort identification for a preliminary research use-case, by integrating data from four contributing organisations (mental health NHS Trust; acute Trust; community services; Local Authority, including education and social care data). Our use case was: "What is the incidence and prevalence of mood disorders and adverse childhood experiences in children 0-17y?"</p> <p>We built a synthetic data generation system that ensures high fidelity of the generated data relative to the original. This is excellent for ensuring testing of the technologies involved and means that the synthetic data can even be used for certain data analytics tasks. Unfortunately, it also had the unintended consequence of making the data</p>	<p>Systems should be developed/tested at scale using synthetic data generated using k-anonymous data dictionaries from real data sources.</p> <p>Intermediate (minimal) representations should initially be readily verifiable by humans as being truly anonymous, to allay information governance concerns; as trust grows, this process could be automated fully.</p>

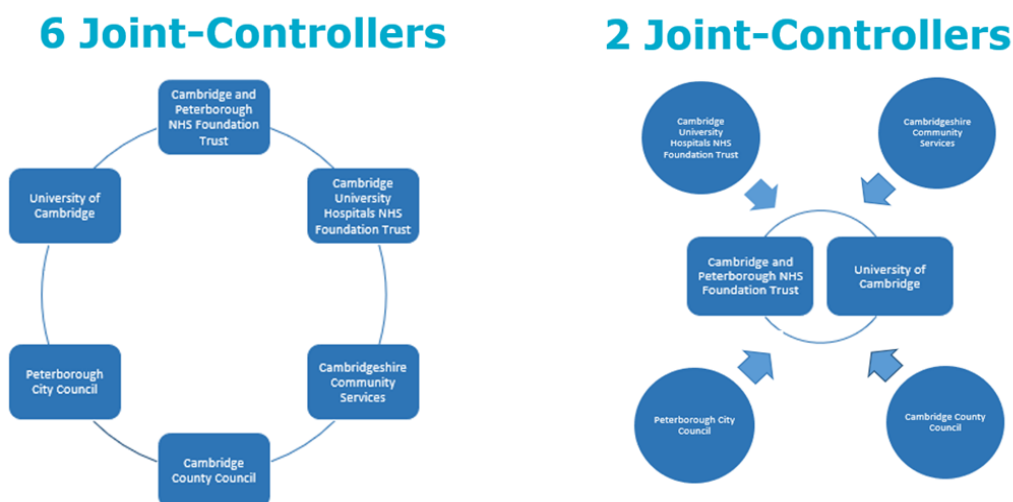
	<p>custodians feel less comfortable that the data was truly synthetic and anonymous, as the large-scale synthetic data was difficult for humans to review.</p> <p>We learned that when synthetic data are used it is important to also build sufficient verification systems that data custodians can be made comfortable that the data is indeed synthetic. Within the Sprint project timescale, full provision of the synthetic data was not achieved. This was not a significant issue in delivering this project as we were able to generate alternative low-fidelity synthetic data and continue the end-to-end testing of the technologies. If future projects depend on high-fidelity synthetic data, simpler intermediate data representations and/or better verification tools will become a necessity.</p>	
<p>7) The technical solutions must be acceptable to the public</p>	<p>Our PPI workshop participants had no substantial objections to federation; however, this was predicated on the data being de-identified. Participants considered federated analyses to be helpful for improving the generalisability of research, and help understand regional differences and local service requirements.</p> <p>We were able to explain how we deployed the Six Safes model. Workshops suggest that participants do not expect or require the risks to be zero (e.g. data breaches). Rather, they expect a plain explanation of the safeguards in place, a plan for what happens when an adverse event takes place, and transparent communication.</p> <p>Participants as young as 11 could understand complex technical concepts (such as virtual desktop infrastructure, access controls and federated analytics) and contribute.</p>	<p>Involving the public in the co-creation stage of the technological solutions is critical.</p> <p>Co-creation workshops should be used for testing and refining explanations of the technology in “plain English”.</p>

4. Governance

The aim of the Governance workstream was to explore how the FAIR TREATMENT project can be realised under current legislation to allow lawful data linkage and processing of multi-agency data for a research purpose, including analysis across a federated network. We aimed to propose an information governance (IG) model to ensure that the legal and ethical obligations placed on data controllers are robustly complied with. Information Governance Services Ltd (IGS) were commissioned to assist in the development of this IG model, and to support the data controllers in the implementation of the database. The model was co-created with a diverse group of members of the public supported by the Anna Freud National Centre for Children and Families. This involved nine workshops including young people (11 to 24y) and parents/guardians. Details of the workshop materials used to support this process are found in Appendix 7.5.

As part of the process, three similar examples of de-identified databases of patients' data were examined. These enable research and other secondary purposes to take place within TREs, taking into account the "Five Safes". Building on these findings, IGS recommended the adoption of the IG documentation to ensure that the FAIR TREATMENT project complies with data protection legislation (summarised in table 1 and more information provided in Appendix 7.3 and 7.4). The biggest challenge was finding a suitable model to govern the sharing of data at a local level. Two legally sound models were identified, depicted in Figure 1 below.

Figure 1: Illustrating two viable governance model options



Data Federation Framework (DFF): A two-level governance framework was developed to support federated analysis (Figure 2, Appendix 7.3). This allows TREs to be part of a wider federation and to participate in federated data analysis projects regardless of the local IG model deployed, provided that certain interoperability criteria are satisfied. A key learning point that emerges is that, when a governance framework is intended to support federated data analysis across multiple TREs, it is important to consider whether this governance framework allows for local flexibility. The DFF and other template documentation developed by IGS (see Appendix 7.4) will be made freely available, so that they can be used to inform future projects involving federated data analysis. Next steps for the Governance workstream will be to identify and assess the IG challenges arising from federating with other (i.e. non-AIMES) TREs. The key findings and recommendations from the governance workstream are found in Table 2.

Figure 2: Showing the two-level governance model for federation

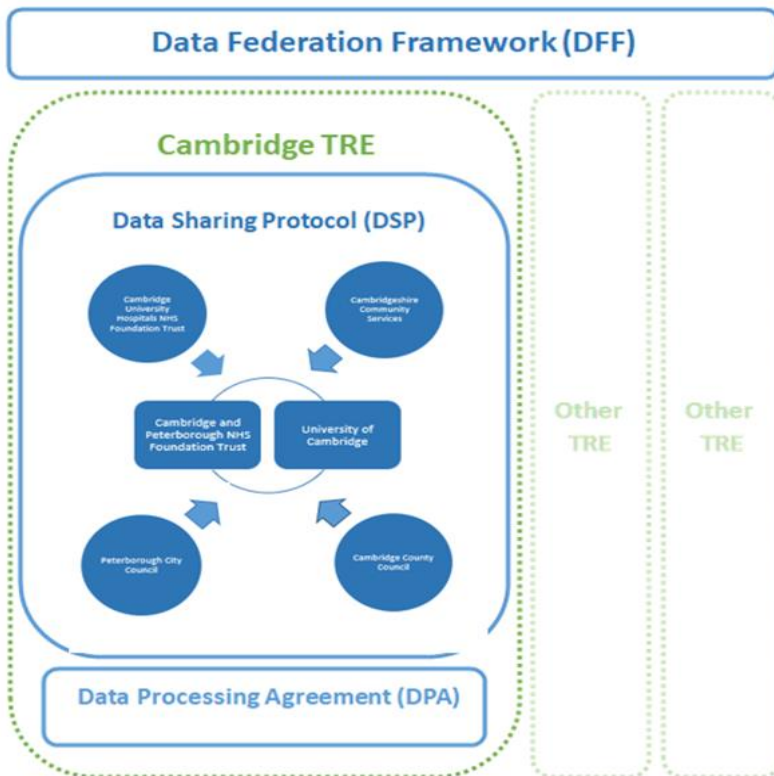


Table 2: Overview of the principles, outputs and deliverables from the workstreams, and recommendations based on learning

GOVERNANCE WORKSTREAM		
Outputs/deliverables		Recommendations
8) Public acceptability	<p>Involving the public in the co-creation stages was incredibly valuable - our participants consistently asked thoughtful, insightful questions and were able to make helpful suggestions. These ranged from high-level suggestions on approaches to governance (such as the composition of the Data Access Committee and how it should approach decision-making) to specific wording within the communications materials.</p> <p>We found enthusiastic support for building secure, scalable TREs for the purposes of research to benefit public health. Once understanding the challenges of building accurate prediction models and the value that the data could have for service improvement, participants wanted to ensure that as much data would be included as possible. Furthermore, that processes would be put into place to ensure data quality.</p> <p>Involving a diverse group of individuals was valuable, with emphasis on including those from underserved groups. This provided important insight into particular issues, for example the inclusion of sensitive data, who should be part of governance groups and ensuring that tools and models do not widen existing inequalities.</p> <p>People wanted to be sure that the data about them was accurate, and ideally participants should have the right to check their information and contextualise it by making additions, corrections, and be included in its interpretation. Any research</p>	<p>Similar projects must ensure that a diverse group of people are included. Recruitment of PPI participants should be via organisations for under-served groups.</p> <p>Projects should meaningfully demonstrate how public feedback had been incorporated (or an explanation as to why some suggestions were not possible to include); participants want to see that their contributions are valued and taken on board, and that PPI work is not a “box-ticking exercise”.</p> <p>Open and transparent communication and co-creation with the public is critical for building trust and confidence. This can be achieved by involving people with lived experience and patient/public advocates at all stages of the project. This includes throughout the pipeline, from co-creation at the design stages of a governance model, to ongoing involvement in</p>

	<p>outcomes should be interpreted by trained, trusted professionals. This was particularly important to underserved groups, who felt that mis-interpretation was a greater risk.</p> <p>Participants wanted reassurance that only the “right” organisations or people would be permitted access to the TRE. This included motivations (i.e. improve public health) and those with a proven track record to demonstrate competence. Organisations with poor track records with minority groups were not favoured.</p>	<p>TRE management (including the Steering Committee and Data Access Committee) and subsequent analysis, interpretation and communication of findings.</p>
<p>9) Viability of legal frameworks</p>	<p>From an IG standpoint, the biggest challenge we faced was finding a suitable model to govern the sharing of data at a local level. To create the research database and ensure that it could be operated effectively within a secure TRE, we needed to establish a robust model, upon which all participant organisations could agree.</p> <p>Two legally sound models were identified for the creation of the Cambridge research database. The models were designed to be able to support the data controllers in the implementation of the research database, and to address some of the specific difficulties associated with the governance of multi-agency data.</p> <p>A Data Federation Framework was developed to regulate the data federation across the TREs. A major benefit of this DFF is that it allows TREs with different local models to be part of a wider federation and to participate in federated data analysis projects, provided that certain interoperability criteria are satisfied.</p>	<p>When a governance framework is intended to support federated data analysis across multiple TREs, it is important to consider whether this governance framework allows for local flexibility. The DFF and other template documentation developed will be made freely available, so that they can be used to inform future projects involving federated data analysis.</p> <p>We additionally recommended the adoption of the following IG documentation to ensure TRE compliance with data protection legislation, the details of each are in Appendix 7.4: (1) Data Sharing Framework, (2) Data Processing Agreement (3) Terms of Reference (4) Data Access Request Form (5) Terms of Use (6) Data Pseudonymisation, Anonymisation and Extraction Policy (7) Information Security Model - Standard Operating Procedures (8) Transparency material and (9) Data Protection Impact Assessment (DPIA).</p>

<p>10) Best practice for the development of governance frameworks and engagement with local data donators</p>	<p>Each model was assessed for acceptability and viability, including reviewing against guidance available from regulatory authorities.</p> <p>We found that the current guidance from the Information Commissioner’s Office (ICO) relating to the use of data within the field of research is brief. Furthermore, guidance available from the Health Research Authority (HRA) is centred upon research projects, and provides limited information about the governance of research databases, especially research databases within federated networks. The HRA provided the following clarification about their guidance on the subject of controllership:</p> <p>“Research databases, research tissue banks and other biorepositories do not have a research sponsor. The controller will be the organisation responsible for the management and oversight of the resource. You can find further guidance on defining a data controller and their responsibilities in section 11 of the Standard Operating Procedures (SOPs).”</p> <p>This left substantial room for interpretation and this led to some divergence in opinion about the ideal approach between local stakeholders. This wording was found to differ somewhat from the wording used in data protection legislation, which defines a data controller as a person or organisation which “determines the purposes and means of the processing of personal data” (Article 4(7) of the UK GDPR). This led to delays in reaching a decision about which of the organisations contributing data to the research database should be data controllers. We noted the ICO is currently drafting more complete advice about the use of data for research purposes, which is welcomed.</p>	<p>Alignment between HRA and the emerging updated ICO guidance specific to research would be an important contribution. This would substantially support and expedite the decision-making process for local IG groups regarding the adoption of suitable governance models for multi-agency data sharing for research purposes.</p>
---	---	---

5. Phase 2: next steps

Access controls: although acceptable for the purposes of the current project, these would be more effective through deeper integrations and technical improvements: (1) integration with the governance system, where users could sign up to relevant collaboration agreements and terms within the platform, and (2) between the technology platforms' authentication systems simplifying the associated governance processes.

Adding real data & operations: we will submit a research ethics application to the Health Research Authority (HRA) before creating the database. Stakeholders need to agree on specific provisions, for instance, how third-parties can request to join the federated network.

Building federated network: Essex and Birmingham need to submit ethics applications. Preliminary discussions have also taken place with existing databanks with the aim of demonstrating federation with a non-AIMES provided TRE, which will require adaptation of our technology platform.

Model building and validation: preliminary predictive models have been developed in the welsh SAIL databank. The next stage is to validate and refine these in the Cambridge and then federated network.

Public and patient engagement: PPI must be meaningful, ongoing, and include a diverse group of people. We have already started work to build a long-term PPI Community of Engagement, which to date has over 200 members from across the UK willing to contribute to child health research. We will work with this community in delivering these next steps, as well as creating communication materials to support transparency and dissemination.

6. Conclusion

Working with a diverse community of members of the public, and a multi-disciplinary consortium including health education and social care providers, academics, the third and commercial sectors, we have demonstrated the feasibility of federated analysis of multi-agency data and creating a governance framework to support it. While the Sprint has enabled us to lay the foundations required to create an innovative research network of this scale, further work will need to be carried out to put the governance framework into practice, test the technical architecture with real data, and continue working with the public to steer the database management.

7. Appendices

7.1. Appendix 1: TRE Security Model Structure



INFORMATION GOVERNANCE SERVICES

Helping you make the most of your data



Furlong House, 10A Chandos Street,
London, United Kingdom, W1G 9DQ



info@informationgovernanceservices.com
www.informationgovernanceservices.com

Cambridge Trusted Research Environment (TRE)

Cambridge Standard Operating Procedures and Security Model

Contents

1. Purpose	3
2. Introduction to the FAIR TREATMENT Project	3
3. Information Governance Framework	4
4. Technical Security	6
Technical overview	6
Data sources	6
Data extraction	7
Data Dictionary	7
Extract frequency	7
Data Transfer and Connecting to Cam-CHILD	7
HSCN and External Firewall	7

Hosting Provider and Security	8
Physical security	8
TRE infrastructure	8
Dedicated cloud solution	9
Data availability	9
System access to the hosted environment	10
Migration	11
Data backups and disaster recovery	11
Source Data Retention	12
Secure Disposal	12
Federated Analytics	12
Secure Aggregation	13
Differential privacy	13
5. Data De-identification	13
Clinical Records Anonymisation and Text Extraction (CRATE)	13
6. Role-Based Access Controls	14
Access to InterMine interface	14
Access to Bitfount environment	15
Audit	16
New user registration	16
Data export	16
7. Procedural Controls	16
Steering group and Data Access Committee	16
Access Form	17
Federated Queries	17
Terms of Use	17

Training	17
Data Quality and Risk	18
8. Operational Processes	18
Opt Out	18
Data Subjects' Rights	18
Subject Access Requests	18
Right to rectification	18
Right to be forgotten	19
Right to restrict processing	19
Right to object	19
Automated Decision Making and Profiling	19
9. Incident Management	19
Unauthorised Disclosure, Loss or Destruction of Personal Data	19
System security testing, audit and reporting	19

1. Purpose

This document outlines the technical and procedural controls used as part of Cambridge's Trusted Research Environment (TRE). It can be used for internal and informational purposes.

2. Introduction to the FAIR TREATMENT Project

FAIR TREATMENT ("Federated analytics and AI Research across TREs for Adolescent MENTAL health") is a project sponsored and led by the Department of Psychiatry at the University of Cambridge launched in 2022, which aims to: (1) combine two new technologies to demonstrate that it is possible to analyse data across TREs in different places and preserve the privacy of individuals; and (2) consult with patients, the public, organisations contributing data, and legal and ethics experts to agree the best way to oversee data use, ensuring that it is managed safely and fairly.

Indeed, the aim of this project is to uncover and test early thinking in the development of a joined-up and trustworthy national data research network to enable cross-research to spot patterns in mental health of young people where professional help is needed.

One of the main barriers for researchers to conduct useful analysis is that information is secured in different places, across health, education, social care records. Such fragmentation makes it difficult to build accurate predictive models, duplicates data, efforts, and resources, and is incompatible with an harmonised approach towards security, fairness and transparency.

Ethical permission has been granted to construct a linked whole-population, de-identified, database of electronic patient record data in Cambridge and Peterborough (NHS REC ID: 20/EM/0299) called Cam-CHILD. This includes data from five other organisations mentioned below. Cam-CHILD will be replicated in Essex and Birmingham, with equivalent ethics applications submitted for both. The Cam-CHILD database project aims to provide a secure way for approved professionals to use this information to find ways to improve local services, better understand the healthcare needs of young people, and find ways to get young people the right kind of help earlier on.

Different health and care organisations have been asked to participate in the project, by contributing with data for the creation of the research database and actively taking part in the decision-making process in regard to the information governance elements applicable to the database. This includes the following organisations: Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge University Hospitals NHS Foundation Trust, Cambridgeshire Community Services, Cambridge County Council and Peterborough City Council.

The project is funded through the Data and Analytics Research Environments (DARE) UK Programme and UK Research and Innovation (UKRI). In addition, it brings together several other organisations such as the Anna Freud National Centre for Children and Families, Intermine, AIMES, Bitfount, the Universities of Essex and Birmingham and Information Governance Services.

3. Information Governance Framework

Information Governance Services (“IGS”) has proposed an information governance framework to ensure that the data controllers involved in FAIR TREATMENT comply robustly with their legal obligations. The proposal is described in detail in a separate document.

In summary, the proposal envisages a two-level data sharing framework to regulate data sharing between the relevant stakeholders. First, at the top level, the proposal suggests a Data Federation Framework between the Cambridge TRE and other TREs. Second, at the lower level, the proposal suggests both a Data Sharing Protocol between the different data controllers, and a Data Processing Agreement between the data controllers and each data processor.

It is suggested that this data sharing framework will be overseen by three separate governing bodies:

- (i) a steering group;
- (ii) a data access committee for the Cambridge TRE; and
- (iii) an operational-level governing body for the data federation, each of which will operate in accordance with specified Terms of Reference.

In addition, the proposal suggests using a Data Access Request Form and a Terms of Use document. The former will be an application form, filled out by the researchers seeking to access the research database. It will capture all the relevant information that the governing bodies need to decide whether to grant access to the research database. The latter will set out the specific terms in accordance with which researchers will be able to receive access to the research database.

These documents will be supplemented by:

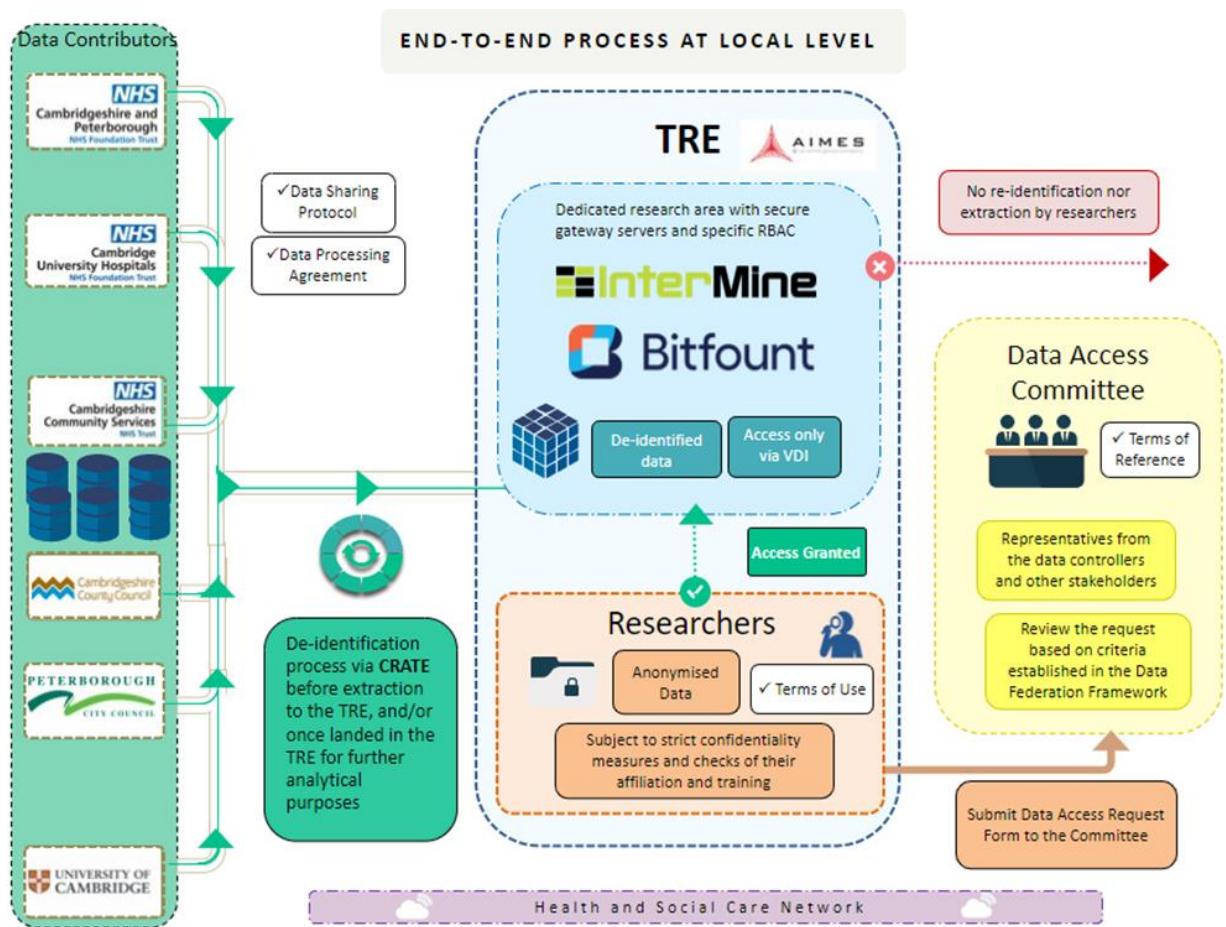
- Privacy notices to inform the public;
- A Data Pseudonymisation, Anonymisation and Extraction Policy; and
- A Standard Operating Procedures and Security Model, the latter of which are embodied in the current document.

A Data Protection Impact Assessment (DPIA) will also systematically review all processing activities relating to the FAIR TREATMENT project, contrasting their necessity and proportionality against the envisaged purposes, assessing the risks to the rights and freedoms of data subjects and the measures conceived to address the risks.

The level of risk attributed to both the impact on the rights and freedoms of the individuals and the likelihood of those rights and freedoms being compromised will be determined. In particular, we will conduct a thorough legal analysis of the justification for processing and sharing data for this specific project to ensure that such processing is compatible with the purpose for which the personal data was collected in the first place and that the data is being shared in a lawful, harmonised, safe, and secure manner by the organisations involved.

The data sharing process will be built with privacy in mind and according to the above-mentioned bespoke sharing framework. Proposed solutions and actions will be included in such assessment to result in the risks being accepted, reduced or eliminated.

The temporary diagram below describes how the process would look like at a local level, subject to changes depending on further discussions between the organisations involved on controllership.



4. Technical Security

The security of patient data is paramount for all the partners in building the TRE. A series of robust technical and procedure controls are being considered to secure data in transit and at rest, whilst providing a complete and granular level of control over access to data, at each step of the processing.

Technical overview

Technical considerations include restricting access to the research database and web site to computers within the institution’s secure network only; requiring that connections from researchers to the database computer use only encrypted HTTPS/SSL, even within the institution, to prevent “wire sniffing”; appropriate securing of the computer against other forms of access (using fire-walls and other aspects of operating system security); and physical security, power protection, and backup systems for the hosting computer(s).

Data sources

The de-identified database will include data from:

- Cambridgeshire and Peterborough NHS Foundation Trust
- Cambridge University Hospitals NHS Foundation Trust

- Cambridgeshire Community Services
- Cambridge County Council
- Peterborough City Council

Data extraction

Data Dictionary

Dr Anna Moore's team have conducted a review of the predictor variables required to model early identification of mental health. Their review highlights 196 candidate variables spanning 7 domains (physical health, psychological, social/environment, behaviour, education/employment, biomarkers, and service use patterns) and > 60% are available only from social care and education sources.

Synthetic data will be first used to demonstrate the performance of the database and the proposed solution. The technology demonstrator (WP1) will use synthetic data generated using k-anonymous data dictionaries from real data sources, for which ethical approval (NHS REC ID: 20/EM/0299, enabled by MRC and Turing Funding) has been granted.

De-identified data will then be used at the implementation phase (from September 2022), to construct a linked whole-population database of electronic patient record (health, education and social care) data on 0-17 years old. The Data Dictionary will be a mapping and security file for Cam-CHILD, describing the relationships between Cam-CHILD and the data sources. It will collect the attributes of the data from health services, schools and social services that will be used for this project. The de-identification process will use the Data Dictionary to identify the fields where masking should be carried out.

Extract frequency

Since only synthetic data will be used for the first stage, the frequency of extraction will be determined later on. How frequently extracts are run is also a decision for each controller to make. It will depend on how up to date each organisation requires their data to be, which will depend on the intended use of that information and IT capacity and resources available to complete the process.

Data Transfer and Connecting to Cam-CHILD

The Data Dictionary will be transferred to the hosted environment via HSCN (see below). End-users wishing to access the TRE will need to do so via a web browser from their organisation's terminal.

Information sent and received via the web browser will be secured with a HTTPS connection, authenticated by Secure Socket Layer Protocol (SSL), and will use a strong encryption key to encrypt all traffic.

HSCN and External Firewall

The Health and Social Care Network (HSCN) is a data network for health and care organisations. It provides underlying network arrangements to integrate health and social care services by enabling them to access and share information in a reliable, flexible, and efficient manner. HSCN is available to organisations whose purpose is the delivery, facilitation, or

support of health and/or social care in England. HSCN is a private network but does not provide security to prevent loss, tampering, authenticity, or inappropriate usage of the information transferred through it.

To protect data in transit, encryption, network protection and strong authentication should be implemented. As described in this document, all data is encrypted when in transit. To protect the network's perimeter against unauthorised access from the outside, firewalls are managed by AIMES.

Hosting Provider and Security

Security of data on the hosted environment is of paramount importance and the supplier of the infrastructure service was selected on this basis. AIMES is a secure digital platform which will host the data received from the data sources and provides the TRE infrastructure. There will be a separate AIMES TRE for each locality, and each locality will be responsible for managing its local TRE.

AIMES hosts in-house private cloud servers in the UK, which are ISO27001, ISO27017 and ISO2018 accredited, and comply with the NHS Data Security and Protection Toolkit and Cyber Essentials Plus. AIMES was the first data centre in the European Union to be awarded the new Data Centre Alliance (DCA) certification.

Physical security

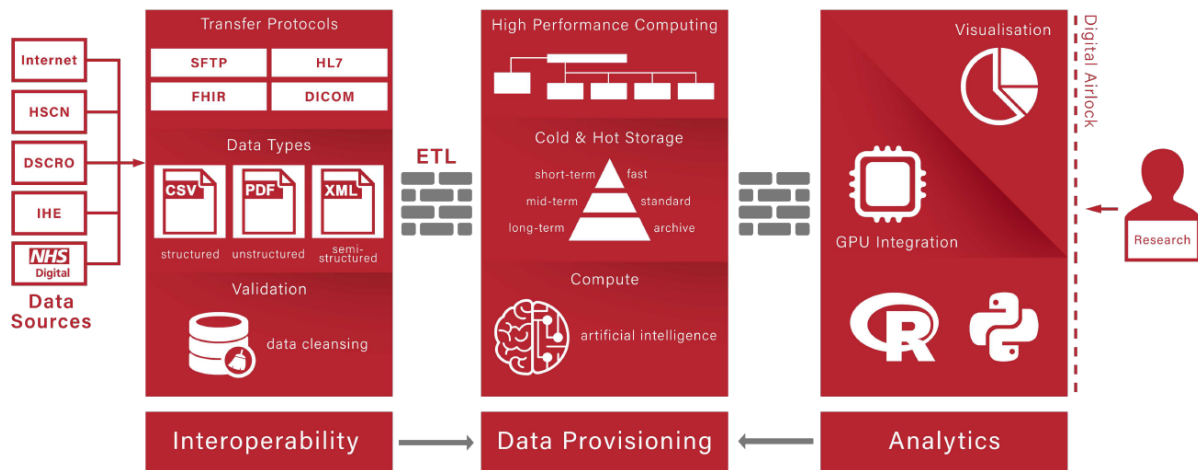
AIMES is located on the Liverpool Innovation Park Campus. The technology park is purpose-built, surrounded by secure fencing with a single point of entry. AIMES provides constant and complete CCTV surveillance across the site, with 10-day backup. Security guards monitor external CCTV and patrol the perimeter frequently. Visitors are presented with the multi-layered access control method (swipe card and a 4-digit passcode), with each passcode personalised to the individual. There are 4 layers of security for data centre access, the foyer, anti-tailgating area, data centre entrance and combination locks on each rack.

Full details can be found here: [20210513-Data_Centre_Security_Features_Diagram](#)

TRE infrastructure

The environment is accessible to researchers through secure gateway servers using multi-factor authentication, for both persistent and non-persistent virtual desktop infrastructure (VDI).

The TRE has 3 zones – Interoperability, Data Provisioning, and Analytics – which are individual securely separated areas. The high-level diagram below outlines the details of the hosted environment and how the infrastructure is set up:



The Interoperability Zone is primarily concerned with the “input” of the TRE. The primary role of this zone is to provide a secure healthcare gateway to receive data, in all formats, in an AIMES cloud platform. This zone is mainly HSCN-facing, however ingestion of data sets from other sources outside of HSCN can be facilitated. For the Cam-CHILD database purposes, the data will be ingested only from HSCN.

The Data Provisioning Zone can be described as the “data vault” and core of the TRE. It consists of high-performance computing (HPC), storage, and computing. The primary role of this zone is to provide storage of data sets of varying sizes and in varying formats, whilst computer nodes can be leveraged for AI, ML/NLP, databases analysis on these datasets. Along the periphery of this zone are customer/analyst driven tool sets. InterMine will sit within this zone.

The Analytics Zone is primarily concerned with the “output” side of the TRE. It provides analysts and researchers with an appropriate view of approved data with an agreed tool set. In this zone, researchers are able to run queries/analysis on these sensitive datasets through a virtual machine. Stata, R, or Python can be used. There is a concept of a digital airlock where the outputs or reports/findings may be released to the analysts but not the actual data itself. Workflows and customisable notifications may be designed and implemented here on request. Data is encrypted at rest and in transit. Bitfount will sit within this zone and enable cross-TRE federation.

Dedicated cloud solution

The TRE will be hosted as a single tenant, isolated platform hosted in AIMES in-house private cloud servers in the United Kingdom. The data centre is located on the Liverpool Innovation Park Campus.

The operational delivery model for the hosted environment has been contracted under an Infrastructure as a Service agreement (IaaS) to provide the University of Cambridge with flexibility to scale up the environment as required. The services include managed firewall, managed active directory, managed 2FA remote access, managed HSCN connectivity and managed shared user space.

Data availability

AIMES guarantees 99.9996% uptime.

System access to the hosted environment

AIMES will deliver a fully managed 24/7 265 manned UK service desk. All details and contact information is documented in the relevant agreements between AIMES and the University of Cambridge.

In order to carry out general system maintenance and any technical fixes, AIMES will need to access the hosted environment at times. For all maintenance and most fixes, this will not require access to any patient information. Any approach would seek to minimise access and subsequent exposure. All default accounts will be removed.

Incident management and error reporting

AIMES owns and operates a UK-based service helpdesk and is staffed during normal business hours. Outside of these hours, the online logging service may be used to raise issues, or a message may be left on the customer service helpdesk recording facility. Faults which require immediate attention may be reported by designated users, by calling the out of hours fault line.

The logging tool, provides:

- Log and assign a unique identifier within the AIMES Autotask Service Management Tool
- Mutually agree with the user and assign a level of priority for all Incidents, based on an assessment of 'Business Impact' and 'Urgency'
- Make reasonable efforts to diagnose and resolve all Incidents at first point of contact using tools including, Knowledge Base, remote assistance and diagnostic capabilities
- The Service Desk has responsibility for end-to-end ownership for logged Incidents
- Keep users updated on progress and status of Incidents, Service Request or Change requests
- Receive and analyse customer feedback via customer satisfaction surveying and other Client perception sensing methods
- Response times commence at the point of receipt of incident or request into AIMES service management tool and are dependent on the specific client contracted service hours
- Incident resolution will apply when confirmation has been received or identified by AIMES that the requirement has been met or the appropriate fix has been applied and successfully tested and/or confirmed by the customer
- Where an incident or request is pending feedback from a customer/User and where contact is unable to be made, the incident or request record will be placed in a hold state, pending a response from the customer.

- Any incident or request that will potentially breach the target SLA will be escalated according to the AIMES hierarchic escalation procedure.
- Where a customer does not have support outside of office hours in their contract, incidents can still be logged out of these hours, either via email or over the Service Desk Portal. However, no work will start, and the timing Clock will not run outside of the contracted support period.
- Where the resolution to an incident requires an escalation to a designated third party the work will continue with regular communication with the third party, but the SLA timing will be on hold until the third party have successfully resolved or completed their element of the resolution.
- Where the Service Desk is unable to establish a fix or workaround for a desktop or infrastructure issue remotely, and an on-site support option has been taken by the customer, the service record will be allocated to a field-based resource for attendance to customer locations. Alternatively, even when a formal contracted service line for desk-side support does not exist, then this can still be purchased separately and on an ad-hoc basis, but at a 'premium' rate.

Migration

The proposed solution and the datasets are newly brought in the TRE, therefore there is no migration to consider. Nevertheless, to ensure the continuity of the solution, the partners considered if/how it would be possible to potentially migrate the system/data from one environment to another.

In the context of migrating data from one environment to another, this could be facilitated with AIMES's Digital Airlock function, which is a secure SFTP server connected to the TRE. This could also be achieved by the technical team at AIMES in the background.

Data backups and disaster recovery

AIMES is providing secure, policy-based backup and recovery for virtual machines hosted on the TRE. The standard back is incremental with synthetic full conducted once daily.

- **Recovery Point Objective (RPO) = 23 Hours.** This means that AIMES will perform a backup of the VM once daily during your backup window which runs from 18:00 to 06:00. AIMES can provide a time during this window for the backup set to run or if client requires a specific time this can be accommodated.
- **Recovery Time Objective (RTO) = 2 hours.** This is the time frame that AIMES requires to acknowledge the restore ticket and perform the necessary restore of the server as agreed in conjunction with the client.

AIMES has also been contracted for Backup as a Service for the TRE. Because the TRE will gather copies of the source electronic records, there is no need for a formal disaster recovery to be in place. Should such an event occur, the backup of the critical components will be used.

The data contributors will have to review their business continuity processes to ensure that the procedures in place will allow for the solution's restoration and a re-uploading of the databases.

Source Data Retention

To build the TRE, health, education, and social care records will be de-identified before being transferred to the hosted environment and processed. The length of time data is retained at source will depend on the different data contributors' positions and these will be reflected in their transparency materials.

Secure Disposal

Where information needs to be erased (whether at the end of the retention period, or on request of a data contributor/data subject), including a virtual machine, physical machine or files or folders, AIMES will use a 'Secure Erase' software that will securely delete data from the TRE. It can also be used on any operating system. AIMES will also produce a certificate attesting the data was erased and erased successfully.

Federated Analytics

Bitfount¹ will be the software enabling federated access to data across the TREs and integrated with InterMine within the AIMES TRE. The AIMES TRE will connect to Bitfount via outgoing HTTP connections and will sit behind a firewall.

The basic component of the Bitfount network running within the AIMES TRE is the Pod (Processor of Data). Pods are co-located with data, checked users are authorized to do given operations on the data and then do any approved computation.

The data can be configured to never leave the Pod and is not accessible to Bitfount or any other parties unless access is specifically granted. The only Pod information shared with Bitfount is metadata. More information on the metadata Bitfount has access to can be found in their [privacy policy](#).

When appropriate researcher access has been granted, Bitfount allows for federated data analyses to be run across the agreed-upon TREs and supports various privacy protection techniques for protecting data (see below). In order to do such an analysis, the researcher runs a service from outside of the TREs, specifying which algorithm to run, with which parameters and privacy controls and on which Pods. The Pods check whether the researcher has the requisite permissions and if approved the analysis is run.

All data entering or leaving Pods, InterMine templates or Bitfount infrastructure use TLS/HTTPS. All communication and federated-analysis messages are end-to-end encrypted.

Bitfount's systems are ISO27001 certified and have undergone extensive security checks, including a full external security review by Blacksmiths Group (headed by former GCHQ Head of Security). Automated security tests are run on all changes to code. Regular penetration tests are run on infrastructure. Monitoring tools run continuously to try to catch intrusions and incidents. Cloud infrastructure is segregated into completely separate production, staging and sandbox environments with limited human access.

1

Full technical documentation at <https://docs.bitfount.com/>

Bitfount operates various process-level security policies, including a secure development policy, supplier management policy, incident management policy and regular security training for all staff.

Secure Aggregation

One of the privacy enhancing techniques supported by Bitfount is secure aggregation. This technique enables the researcher to obtain a sum of statistics across the TREs without being able to calculate any individual contribution. This is used for calculating totals across any analysis, as well as for training machine learning models across federated data.

Differential privacy

A second privacy enhancing technique supported in the Bitfount platform is differential privacy. Differential privacy is a data-perturbation-based privacy approach, which can reduce information about the single individual while retaining the capability of statistical reasoning about the dataset. The parameters corresponding to a chosen level of privacy that is considered acceptable for disclosure are set within the platform by the PI.

5. Data De-identification

Synthetic data will be used at first. At the implementation stage, when real data is ingested, the TREs will be designed to remove all personal identifiable information before processing any people records. The solution will use CRATE in conjunction with the Data Dictionary to define at a field level the identifiers to be omitted, masked, or truncated.

Clinical Records Anonymisation and Text Extraction (CRATE)

CRATE will be used to de-identify the electronic patient record data. CRATE is free and open-source software that uses methods such as scrubbing, hashing and truncation to de-identify/pseudonymise both structured and unstructured data in a source database, thereby producing a new, de-identified destination database.

CRATE is a validated software system for removing identifiers from structured and unstructured data to create anonymous or pseudonymised databases.² It will transform one relational database to another, via the Data Dictionary that describes the source database (including the location of identifiers such as names, dates of birth, local/national identity numbers, addresses, and so forth). The Data Dictionary also defines transformations (e.g. "blur date of birth to the first of the month", "remove all known identifiers from this free-text field"); it governs which data are translated through to the destination database, and how.³

If required, CRATE can extract free text from external files (e.g. Word documents, PDFs) referenced in the source database, for de-identification and incorporation into the destination database. The Data Dictionary can be automatically drafted by CRATE but is then edited and verified by the operator.

For pseudonymised databases, source identifiers can be replaced by an irreversibly encrypted version (e.g. via HMAC-SHA256). As well as the removal of identifiers known in the source database, and handling of typographical errors (to a configurable threshold), CRATE can remove a range of identifiers not recorded in the source database—for example, it can be configured to remove all n -digit numbers from free text (e.g. 10-digit UK NHS numbers, 6/11-

² <https://pubmed.ncbi.nlm.nih.gov/28441940/>

³ <https://crateanon.readthedocs.io/en/latest/>

digit phone numbers), all UK postcodes, and arbitrary names (e.g. from a large public list of names, minus medical eponyms), and it can blur all dates found in free text. It supports recursive de-identification where inter-patient relationships are known.

CRATE also supports de-identified exact and fuzzy linkage (in development, validation paper pending), dynamic de-identification via an application programming interface, the implementation of opt-outs, and other functions not relevant to the current project.⁴

6. Role-Based Access Controls

Access to the TREs requires formal authentication methods. The user will need network credentials to access the TRE's system, a unique username and password. Only authorised individuals will be authorised and granted access to the environment. User management systems are in place at each step as described below.

Access to InterMine interface

InterMine⁵ is a system to integrate data from various sources and access those both through a web application and web service API⁶.

It provides a sophisticated query builder allowing construction of advanced custom searches across the integrated data, and a mechanism to construct and save predefined searches called templates. The templates usually have one or more data types returned as output and one or more editable or not-editable constraints (or filters) to restrict the subset of data provided.

InterMine templates are made available for direct analysis by the researchers approved within the TRE in accordance with the IG approvals detailed in Section 8. InterMine templates are a mechanism to enforce data access control, in particular the non-editable constraints enforce ethics permissions, and editable constraints allow filtering of the allowed data.

In the TRE, via SSH protocol, the authorized people can access the InterMine features through the web application.

The authorizer will create an InterMine account for every approved researcher and generate an API access token which uniquely identifies the approved researcher's account. The authorizer creates, in the approved researcher's account, private templates, which allow the approved researcher to access those data permitted.

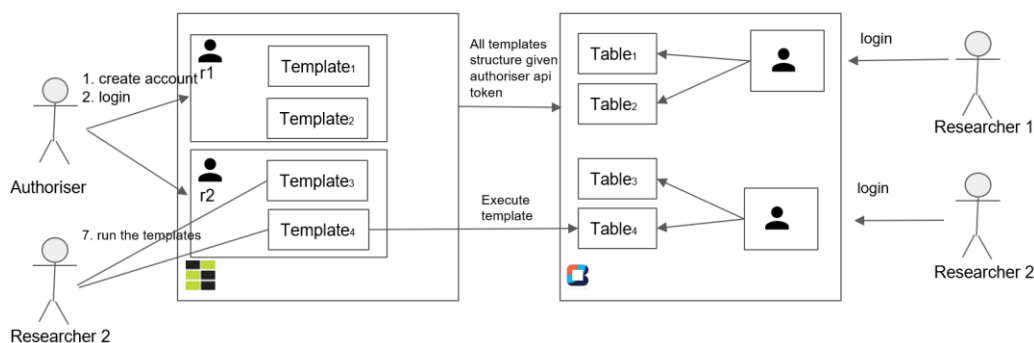
In the TRE, the approved researchers can only access the templates (via the web service API) that have been assigned to them. To execute the template, they need their API access token. Network controls will not allow the approved researchers to access InterMine features such as query builder or template editing.

⁴ <https://pubmed.ncbi.nlm.nih.gov/28441940/>

⁵ <https://doi.org/10.1093/bioinformatics/bts577>

⁶ <https://doi.org/10.1093/nar/gku301>

One InterMine user for each approved researcher and full access to Bitfount Pod



1. The authoriser creates a new account using the email of the approved researcher
2. The authoriser logs into the new approved researcher account
3. The authoriser creates a private template, or a set of private templates (non-editable constraints enforce ethics permissions; editable constraints allow filtering of the allowed data)
4. The authoriser generates an API access token linked to the new account and send it to the researcher (other options?)
5. The researcher changes his API access token via web service
6. The researcher can get the templates belonging to him (metadata) via web service
7. The researcher can execute the templates via web service (curl or client library e.g. [InterMine Python Client WS](#))
8. The research can access to the model description via web service

Access to Bitfount environment

Usage-based permissions for federated analysis of InterMine templates hosted across multiple TREs is governed within the Bitfount platform.

Access rights for the Pods are managed by Bitfount's access manager service. Access to the hub and access managers is protected by strong authentication and authorization controls, with user passwords not being held by Bitfount.

There are three key user types of the Bitfount platform:

- Data Scientist: someone who wants to build a model on some data. Typically, a data scientist, machine learning engineer, analyst or researcher.
- Data Custodian: someone who makes data and computation available via a Bitfount Pod.
- Authoriser: a person who decides whether a Data Scientist can access data and compute that have been made available by a Data Custodian, and exactly which algorithms the Data Scientist is allowed to run. An authoriser manages Pod access via the access manager service.

All user authentication uses OAuth, SAML or OIDC. Every attempt to use data is checked against the permissions that the requesting user has been authorized for before being allowed. Users can request permissions and be granted permissions through the platform.

Role-Based Access Control (RBAC) systems are integrated to enable easy, centralised permission management. These access permissions allow Authorisers to manage which users are allowed access to which data (i.e. which Pods), as well as how they can access that data (e.g. must have differential privacy, or may only evaluate machine learning models, but may not train them).

Audit

All activity in the Interoperability and Analytics Zones of AIMES TRE(s) is fully audited and recorded. In the Data Provisioning Zone, services also include audit.

Moreover, Bitfount maintains and makes available a fully auditable activity history. All queries that have been made per date, Pods, or activity are logged.

New user registration

All new researchers will need to complete a specific procedure and provide evidence of their supporting documents such as their training certificates and contracts. The data controllers will determine how they will delegate such a process.

Data export

By default, all data must reside at all times within the TREs' network infrastructures to ensure that these data are subject to the same security standards. The VDI must be used to access and analyse the data. The Extraction Policy will define exactly the conditions under which researchers may, or may not, export the data outside of the TRE's environment.

7. Procedural Controls

Access to the controlled environment is only given under a series of procedural controls to ensure only those who are authorised and trained use the system. These controls provide a solid governance system to manage and mitigate risks associated with accessing de-identified patient data. Such controls are detailed in this section.

Steering group and Data Access Committee

The data sharing framework requires two governing bodies that separately operate in the context of the Cambridge TRE and of the data federation:

- **At a strategic level:** a steering group comprised of representatives from all the data controllers shall have the right and power to decide, among other things, upon:
 - requests made by third-party organisations to join the framework as members;
 - data curation projects that involve bringing additional into the database;
 - any amendment proposed to the framework's DSP or DPA;
 - other requests of a strategic nature.
- **At operational level:** a data access committee comprised of representatives from all the data controllers and other stakeholders (e.g., lay persons) is responsible for reviewing and approving data access requests made by researchers that satisfy the criteria established in the data sharing framework.

The governing bodies are responsible for overseeing and monitoring the use of the Cambridge TRE. Terms of Reference (ToR) regulate each of the mentioned governing bodies, establishing rules, among other things, about membership, appointment of Chair, frequency of meetings, quorum for deliberation and approval, powers and responsibility and accountability (e.g., reporting obligations).

Access Form

All projects' leads proposing to use data from the Cambridge TRE are required to submit a project application to the data access committee. These are assessed based on a certain number of criteria such as the potential benefits of the project, whether the appropriate supervision and governance is in place, the confidentiality safeguards and the risk of re-identification.

If there are any questions or concerns feedback on an application, the applicant will be given an opportunity to resubmit an updated/consolidated form. Once the project is approved, the authorised users will then be able to access the TRE and carry out their search.

Federated Queries

Researchers can run queries across multiple TREs where the appropriate third parties have agreed to collaborate and join the framework.

In the case of the data federation, a separate governing body comprised of representatives from the different TREs integrating the federation operates. Working at an **operational level only**, this body is responsible for reviewing and approving requests for access to data across the federation, made by researchers that satisfy the criteria established in the Data Federation Framework (DFF).

Strategic level decisions, including any amendment proposed to the DFF, will need to be taken to each TRE's own strategic level governing body for deliberation and approval.

To run a federated search across multiple TREs, the project application process follows a similar path to the local project application process. New project applications are completed, specifying that the user(s) wishes to run a federated query and the TREs they want to collaborate with. The governing body reviews applications on the same risk benefits basis. If approved, the project is then submitted to each TRE's own body for them to review the application and make their decision.

Terms of Use

All individual researchers named in the access forms will have to adhere to Terms of Use clarifying their roles and responsibilities in terms of liabilities for the individuals concerned. These ensure that users are contractually obliged to comply with certain standards and with the TRE(s)'s applicable policies, especially regarding confidentiality and data protection.

The signed Terms of Use will be attached to each application to be included in the new user registration process.

Training

All TRE administrators are trained in all aspects of administering the system. All end users will also require demonstrating they have up to date Information Governance training. Evidence of such training completion will be uploaded as part of the new user registration process.

Data Quality and Risk

As a basic principle the key set of criteria that make up good data quality are:

- Complete
- Accurate
- Relevant
- Accessible
- Timely
- Valid
- Defined.

Under Article 5(1)(d) of the UK GDPR, controllers have a duty to keep personal data accurate and up to date. Data quality and the risk of personal data being disclosed are directly correlated. Lower data quality, incomplete information, spelling mistakes in the data source lead to a higher level of risk. This risk can be managed, and overtime, mitigated through applying a process to surface errors and missing data so they can be fed into the controllers' data quality improvement plans to correct and/or complete these data. This process should form part of a continuous quality improvement cycle to help improve data quality and therefore reduce the risk of an IG breach.

8. Operational Processes

Opt Out

The National Data Opt-Out allows individuals to opt-out of their personal confidential information being used for secondary purposes, such as research or planning. It does not apply if the information is being used for primary (i.e. direct care) purposes, or if the information is not identifiable.

Even though the data will be de-identifiable, it will still be possible for data subjects (or their parents) to opt out. The transparency materials addressed to the public include the contact details and detail the process they can use to opt-out, either online or by post.

Data Subjects' Rights

Subject Access Requests

Under Article 15 of the UK GDPR, data subjects are entitled to access any personal data that controllers hold concerning them, and related information, such as the purposes of the processing, the recipients of the personal data, and the existence of automated decision-making. The data processed in Cam-CHILD is extracted from the existing records of the Data Contributors. In the absence of a legal basis to refuse a request, the Data Contributors must be able to provide a copy of the records they hold about each data subject when requested to do so.

Right to rectification

Under Article 16 of the UK GDPR, data subjects have the right to have inaccurate personal data concerning them rectified, and incomplete data completed. Again, it will be the responsibility of the Data Contributors to rectify and/or complete their records as necessary, to ensure compliance.

Right to be forgotten

Under Article 17 of the UK GDPR, data subjects have the right to have personal data concerning them erased in certain circumstances. This right is sometimes referred to as 'the right to be forgotten'. If the deletion of a specific record from Cam-CHILD is required, it must be ensured that it is possible to proceed with such a request. The specific record could be deleted by the administrators, in order to remove it from use by any researchers (if materially possible considering that the data will be pseudonymized before transfer into the TRE). Data Contributors are responsible for deleting the data if a data subject requests so.

Right to restrict processing

Under Article 18 of the UK GDPR, data subjects have the right to restrict the processing of personal data concerning them in certain circumstances. Again, it must be ensured that it is technically possible to give effect to this right upon request.

Right to object

Under Article 21 of the UK GDPR, data subjects have the right to object at any time to the processing of personal data concerning them. This would be similar to the 'opt-out' process, which has been explained above.

Automated Decision Making and Profiling

Article 22 of the UK GDPR restricts organisations from making solely automated decisions which have legal or similarly significant effects on individuals, except in certain limited circumstances. However, this will not be applicable to the FAIR TREATMENT project, as all data held and made available for research purposes in the TRE is de-identified, and no automated decisions are made.

9. Incident Management

Whilst all efforts are being made in the design and implementation of the Cambridge TRE to minimise risks as far as possible, processing personal data always has an element of risk attached to it. Managing such risks is a crucial aspect of a complete governance model.

The Data Protection Impact Assessment (DPIA) outlines some of the potential risks that may arise in the use of the TRE. The DPIA covers risks that have been identified and therefore are included in the security model, however if further risks are discovered later than the DPIA and the security model should be updated to address those risks.

Unauthorised Disclosure, Loss or Destruction of Personal Data

In the event of any unauthorised disclosure, destruction, access, or loss of patient identifiable information it should be managed and reported according to TRE(s)' policies and procedures. All administrators, users and researchers should understand what constitutes a data breach and know what action to take in such circumstances.

System security testing, audit and reporting

The TREs will require ongoing review. This will include continuing to update the project's Data Protection Impact Assessment (DPIA) on a scheduled basis to ensure all relevant changes in law and policy are being applied, and any new risks are being identified and appropriate action taken. The current DPIA has recommendations on this basis. As part of this review process

system security should be routinely reviewed and tested. The governing bodies should determine how often these audits and security tests take place, and how they will review the reports and take any required/recommended actions as a group.

Appendix

- Committees TOR - to be drafted once a model and specifics are determined
- Data Dictionary - to be finalised

7.2. Appendix 2: InterMine screenshots demonstrating research query process

Diagnosis → Sex and Age

Given the diagnosis, show the sex and the age of children

Patient > Age In 22

Patient > Age In 22

Diagnosis > Name

Results Preview

Patient > Gender	Patient > Date Of Birth	Patient > Age In 22	Diagnoses > Name
F	2005-10-01	16	moderate depressive episode
F	2005-10-01	16	moderate depressive episode
F	2005-10-01	16	moderate depressive episode
F	2007-10-01	14	moderate depressive episode
F	2007-12-01	14	moderate depressive episode

+ 12 more results

VIEW 17 ROWS

RESET

EDIT QUERY

EDIT TEMPLATE



Web service URL

Close <<

Patient Gender		Patient Age In 22	Diagnoses Name
F	2 Patient Genders		
F		16	moderate depressive episode
F		16	moderate depressive episode
F		16	moderate depressive episode
F		14	moderate depressive episode
F		14	moderate depressive episode
F		12	moderate depressive episode
F		10	moderate depressive episode
F		8	moderate depressive episode
F	2015-06-01	6	moderate depressive episode
F	2015-06-01	6	moderate depressive episode

2 Patient Genders

Search for a value... LINEAR

<input checked="" type="checkbox"/> Item	Count
<input type="checkbox"/> F	13
<input type="checkbox"/> M	4

▶ FILTER

⚙️ ✖️ ⏴ 📊	⚙️ ✖️ ⏴ 📊	⚙️ ✖️ ⏴ 📊	⚙️ ✖️ ⏴ 📊								
Patient Gender	Patient Date Of Birth	Patient Age In									
F	2005-10-01	16	<div data-bbox="826 120 1477 687"> <p>Showing numerical distribution for 9 Patient Age In 22s</p> <table border="1"> <thead> <tr> <th>Min</th> <th>Max</th> <th>Average</th> <th>Std Deviation</th> </tr> </thead> <tbody> <tr> <td>4</td> <td>16</td> <td>10.059</td> <td>4.575</td> </tr> </tbody> </table> <p>Trim from <input type="text" value="4"/></p> <p>Trim to <input type="text" value="16"/></p> <p>FILTER</p> </div>	Min	Max	Average	Std Deviation	4	16	10.059	4.575
Min	Max	Average		Std Deviation							
4	16	10.059		4.575							
F	2005-10-01	16									
F	2005-10-01	16									
F	2007-10-01	14									
F	2007-12-01	14									
F	2009-12-01	12									
F	2011-02-01	10									
F	2013-04-01	8									
F	2015-06-01	6									
F	2015-06-01	6									

Oppositional Defiant Disorder ➔ Social Services Risk Flags

Show the association between oppositional defiant disorder and the social services risk flags

Diagnosis > Name

Results Preview

Patient > Housing Alert	Patient > Safeguarding Alert	Diagnoses > Name
0	0	oppositional defiant disorder
0	0	oppositional defiant disorder
0	1	oppositional defiant disorder
0	1	oppositional defiant disorder
0	1	oppositional defiant disorder

+ 1 more results

VIEW 6 ROWS

RESET

EDIT QUERY

EDIT TEMPLATE



Web service URL

Close <<

ADHD ➔ Sex and Social Services Risk Flags

Show the association between ADHD disorder and sex, and social services risk flags

Diagnosis > Name

Results Preview

Patient > Gender	Patient > Housing Alert	Patient > Safeguarding Alert	Diagnoses > Name
F	0	0	Hyperkinetic disorder, unspecified
M	0	0	Hyperkinetic disorder, unspecified
M	0	0	Hyperkinetic disorder, unspecified
M	0	0	Hyperkinetic disorder, unspecified
M	0	0	Hyperkinetic disorder, unspecified

+ 2 more results

VIEW 7 ROWS

RESET

EDIT QUERY

EDIT TEMPLATE



Web service URL

Close <<

ADD COLUMNS

MANAGE FILTERS

MANAGE RELATIONSHIPS

UNDO

SAVE LIST

PYTHON

EXPORT

Showing 1 to 7 of 7 rows

Rows per page: All (7) Page 1

Patient Gender	Patient Safeguarding Alert	Diagnoses Name
F	0	Hyperkinetic disorder, unspecified
M	0	Hyperkinetic disorder, unspecified
M	0	Hyperkinetic disorder, unspecified
M	0	Hyperkinetic disorder, unspecified
M	0	Hyperkinetic disorder, unspecified
M	0	Hyperkinetic disorder, unspecified
M	1	Hyperkinetic disorder, unspecified

2 Patient Genders

Search for a value... LINEAR

Item	Count
<input type="checkbox"/> M	6
<input type="checkbox"/> F	1

FILTER

Dashboard > Manage Pods > Intermine Aimes Pod > Manage Access

Access Request

Status

Pod

Assigned Roles

Username

Grant Access



Pod name

Intermine Aimes Pod

User

matt

Assign role

DP Restricted Modeller

Description

This modeller can perform any modelling task on the selected pods, apart from running custom models, and get back model weights and metrics with the restriction that the Differential Privacy parameter epsilon expended in the modelling task is less than 3.

Permissions

Protocol

✓ Federated averaging

✓ Results only

✓ Any installed protocol

Algorithm

✓ Train

✓ Evaluate

✓ Train and evaluate

CANCEL

GRANT ACCESS

GRANT ACCESS

1-1 of 1

Search assigned roles

0-0 of 0

7.3. Appendix 3: Proposed Information Governance Model



INFORMATION GOVERNANCE SERVICES

Helping you make the most of your data



Furlong House, 10A Chandos Street,
London, United Kingdom, W1G 9DQ



info@informationgovernanceservices.com
www.informationgovernanceservices.com

Cambridge Trusted Research Environment (TRE)

Information Governance Model Overview

Contents

1. Introduction	2
2. Purpose	2
3. Case studies	2
4. Proposed Information Governance Model	5
4.1. Data Sharing Framework	6
4.2. Terms of Reference (ToR)	8
4.3. Data Access Request Form and Terms of Use	9
4.4. Data Pseudonymisation, Anonymisation and Extraction Policy	10
4.5. Information Security Model – Standard Operating Procedures	10
4.6. Privacy / Fair Processing / Transparency material	11
4.7. Data Protection Impact Assessment (DPIA)	12
5. Compliance of the Proposed Model with the Six Safes	13

1. Introduction

The Department of Health and Social Care's data strategy, outlined in "*Data saves lives: reshaping health and social care with data*", sets out the Government's overarching vision for the use of data in the health and care sector. In its chapter dedicated to "*Empowering researchers with the data they need to develop life-changing treatments, models of care and insights*", the policy paper highlights the important role that Trusted Research Environments ("TRE") will continue to have as a means of providing reassurance to the wider public that the individuals and organisations entrusted with their data are keeping it safe.

In consonance with this policy, HDR UK (Health Data Research Alliance UK) has developed a series of papers setting out the principles and best practices applicable when building TREs. These principles and best practices build upon the concept of "6 Safes" earlier adopted by the Office of National Statistics (ONS) and requires the adoption of effective controls to achieve "Safe people", "Safe projects", "Safe setting", "Safe data", "Safe outputs" and "Safe returns".

The ambition of the University of Cambridge is to implement the FAIR TREATMENT project (Federated analytics and AI Research across TREs for Adolescent MENTAL health), which involves collaborating with health and care organisations to create a research database and build a federated TRE that will enable research in the field of child and adolescent mental health conditions.

2. Purpose

The purpose of this document is to provide a **high-level view** of the documentation that we aim to put in place to support the implementation of the FAIR TREATMENT project and ensure that a lawful, transparent, robust, consistent and safe information governance model is implemented. This document aims to provide all stakeholders with a **preliminary** idea of what the information governance model may look like and how data would be handled and shared in compliance with data protection and other relevant legislation, as well any national guidance from organisations such as, but not limited to, the Information Commissioner's Office (ICO) and HDR UK.

3. Case studies

We will present three similar examples of de-identified databases of patients' data enabling research and other secondary purposes to take place within safe trusted research environments. We hope they will provide real-life comparisons of existing processes and initiatives that currently allow for data sharing across different organisations.

3.1. Discover-NOW

[Discover-NOW](#) is a database which is utilised by the NHS, analysts, researchers, commissioners, local authorities and primary care networks for secondary purposes such as research, service evaluation, improvement, planning and population health management. It is the de-identified version of the Whole Systems Integrated Care ("WSIC") platform comprised

of electronic health records which are used by health and care professionals to save key information about their patients. WSIC is made up of various patient healthcare records, including information from both health and social care.

The WSIC programme began in 2013 and the implementation of local plans started from April 2015. The de-identified database has been operated as Discover-NOW since 2020.

The personal data which resides in the identifiable WSIC database goes through a pseudonymization process, removing various aspects of the identifiable data to ensure that the data within Discover-NOW is de-identified. Any user who accesses the Discover-NOW database is not able to re-identify the patient. Discover-NOW is hosted in an entirely different instance and server to WSIC in order to allow for separation of the datasets and for specific security measures and access controls.

Any user who wishes to gain access to the Discover-NOW database must go through a strict process, governed by the North-West London Sub-Data Research Access Group ("SDRAG"), including an application on why they need access to the data and how the public will benefit from their project reviewed by a Data Access Committee. All data they will have access to is completely anonymized.

All data within Discover-NOW is held within a Trusted Research Environment ('TRE'), held on a secure Microsoft Azure SQL server, which is based in the UK. The access controls are benchmarked against the 'Five Safes' framework. The infrastructure has numerous controls, such as:

- o Virtual Private Network and Virtual Desktop Infrastructure;
- o Role-based access controls, which ensure that authorised individuals obtain granular access to the data, on a need-to-know basis, in accordance with their role;
- o An Active Directory to manage user accounts, which includes identity management and password functions;
- o A comprehensive suite of audit and security control tools, such as dashboards and resource-monitoring.

Users are unable to copy or extract any data from Discover-NOW outside of the TRE.

The Discover-NOW Citizens Advisory Group (CAG) comprises members who are reflective of the North-West London population. Their role is to deliberate with the Discover-NOW Board on issues pertaining to Discover-NOW dataset (such as access criteria) and provide recommendations which are used to shape the operation and development of Discover-NOW. To ensure that the deliberation process, content and direction is authentic and balanced, Discover-NOW set up a virtual CAG Steering Group to underpin this work and act as an advisory body and a critical friend.

3.2 SAIL Databank

The [SAIL Databank](#) was established in 2007 by the Population Data Science group at Swansea university. SAIL stands for Secure Anonymised Information Linkage. Its purpose is to make individual data, collected in the course of health and other public service delivery, accessible safely in order to answer important questions that could not otherwise be addressed without prohibitive effort and cost. The scope of SAIL data has expanded to include administrative data that were not previously accessible (such as education, housing and employment) and emerging health data types (such as genomic, free-text and imaging).

Anonymised, person-based records are held in the SAIL Databank and can be linked together to address research questions. The data linkage solution, which allows the identification of

patterns across entire populations to give a broad picture, is provided by Secure e-Research Platform (SeRP).

The SAIL Databank does not receive or handle identifiable data. Commonly recognised identifying details are removed before datasets go to SAIL Databank and once anonymised they cannot be re-identified. Because SAIL holds only anonymised data, researchers carry out their work without knowing the identities of the individuals represented in the data. Digital Health and Care Wales operate as a trusted third party (TTP) to anonymise records and then match these records with non-personal event information that's used for research.

The security and protection of the data is protected through their 'Privacy by Design' methodology, regulated by a team of specialists and overseen by an independent Information Governance Review Panel (IGRP). The role of the IGRP is to provide independent guidance and advice on Information Governance policies, procedures and processes for SAIL Databank. The Panel comprises representatives from various organisations and sectors, including the Welsh Government, Public Health Wales and the public. All proposals to use SAIL Databank are subject to review by the IGRP to ensure that they are appropriate and in the public interest.

When access has been granted, the requested data can be viewed using the SAIL Gateway, a privacy-protecting safe haven and remote access system. This enables research to be carried out in a secure and protected environment and it safeguards the data from external linkage attacks that may risk individual privacy. SAIL's remote access system provides time-limited access to the datasets and is subject to researcher verification, a data access agreement, and physical and procedural controls. SAIL Gateway has a number of levels of security:

- o Fire-walled Virtual Private Network (VPN);
- o Enhanced user authentication;
- o Auditing of all SQL commands;
- o Configuration controls to ensure that data cannot be removed or transferred unless authorised.

SAIL Databank also has a long-standing Consumer Panel which was established in 2011. It currently has 12 members with on-going recruitment. Panel members are involved in all elements of the SAIL Databank process, from developing ideas, advising on bids through approval processes (via the independent Information Governance Review Panel), to disseminating research findings.

3.3 Clinical Records Interactive Search (CRIS)

The [CRIS](#) system was developed at South London and Maudsley NHS Foundation Trust (SLaM) and provides a means of analysing de-identified patient data from the Oxford Health NHS Foundation Trust (OHFT) Electronic Patient Records (EPRs). CRIS transforms information from the Trust's EPR system into a pseudonymised database appropriate for research, service evaluation and clinical audit use.

CRIS extracts data from the medical record in an identifiable state, which is processed to remove the patient identifiers, and a new pseudonymous database provisioned. The transformation process uses the patient's EPR system identifier to derive a unique ID for each patient in the database. This ID does not allow CRIS users to identify patients. However, where patients have given appropriate consent, the ID can be used by authorised personnel to

contact patients who have been identified as potential recruits to an ethics approved research project.

CRIS information is held securely with strict arrangements about who can access the information. This will include Trust staff, clinicians, and approved researchers. CRIS was developed with extensive service user involvement and adheres to strict governance frameworks managed by service users. It has passed a robust ethics approval process acutely attentive to the use of patient data. The data is used in an entirely anonymised and data-secure format and all patients have the choice to opt out of their anonymised data being used.

CRIS can only be accessed from the Trust network. Data from CRIS must be kept within the Trust's firewall and can only be saved on the CRIS shared drive on a Virtual Desktop Infrastructure ("VDI"). Additional permission is required for derived data to be analysed outside of the environment, from the CRIS Oversight Group.

The security model also includes regular audits of searches carried out using CRIS (all searches by all users are recorded and can be audited). A monthly audit report is provided to the CRIS Oversight Group.

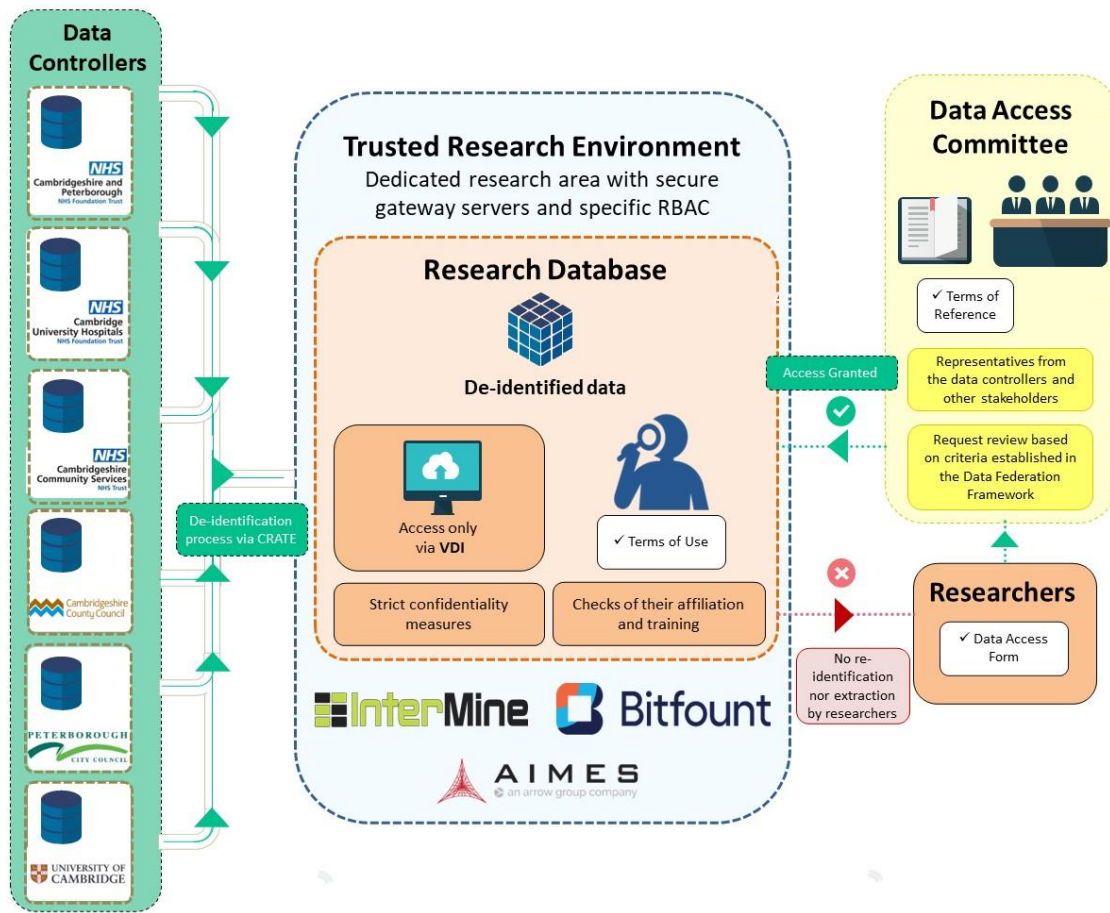
The CRIS Oversight Group, led by the Trust Caldicott Guardian, reviews all requests to use CRIS as a de-identified database. In order to grant access, the Trust must demonstrate that OHFT clinical data are used responsibly and for projects with demonstrable research and clinical importance. Access to CRIS requires either an NHS contract or a Research Passport. A Research Passport is an application form that a researcher, not employed by an NHS organisation, completes to inform an NHS Trust of the research activity that is intended to be conducted within an individual Trust.

The UK-CRIS system was developed at the University of Oxford through NIHR funding, and has the ability to federate over 14 Mental Health Trusts. The UK-CRIS Network are a group of NHS Mental Health Trusts who work together to accelerate research work in dementia and mental health. Authorised researchers from the UK-CRIS network may be granted access to the CRIS pseudonymised dataset. Access to UK-CRIS is via a secure private network on a research platform with Amazon Web Service (AWS). UK-CRIS was originally hosted in an "on-premises" datacentre and is now hosted on SeRP UK.

4. Proposed Information Governance Model

We propose an information governance model that will ensure the obligations placed on data controllers by the law are robustly complied with. The model will require the adoption of the information governance documentation explored in this section.

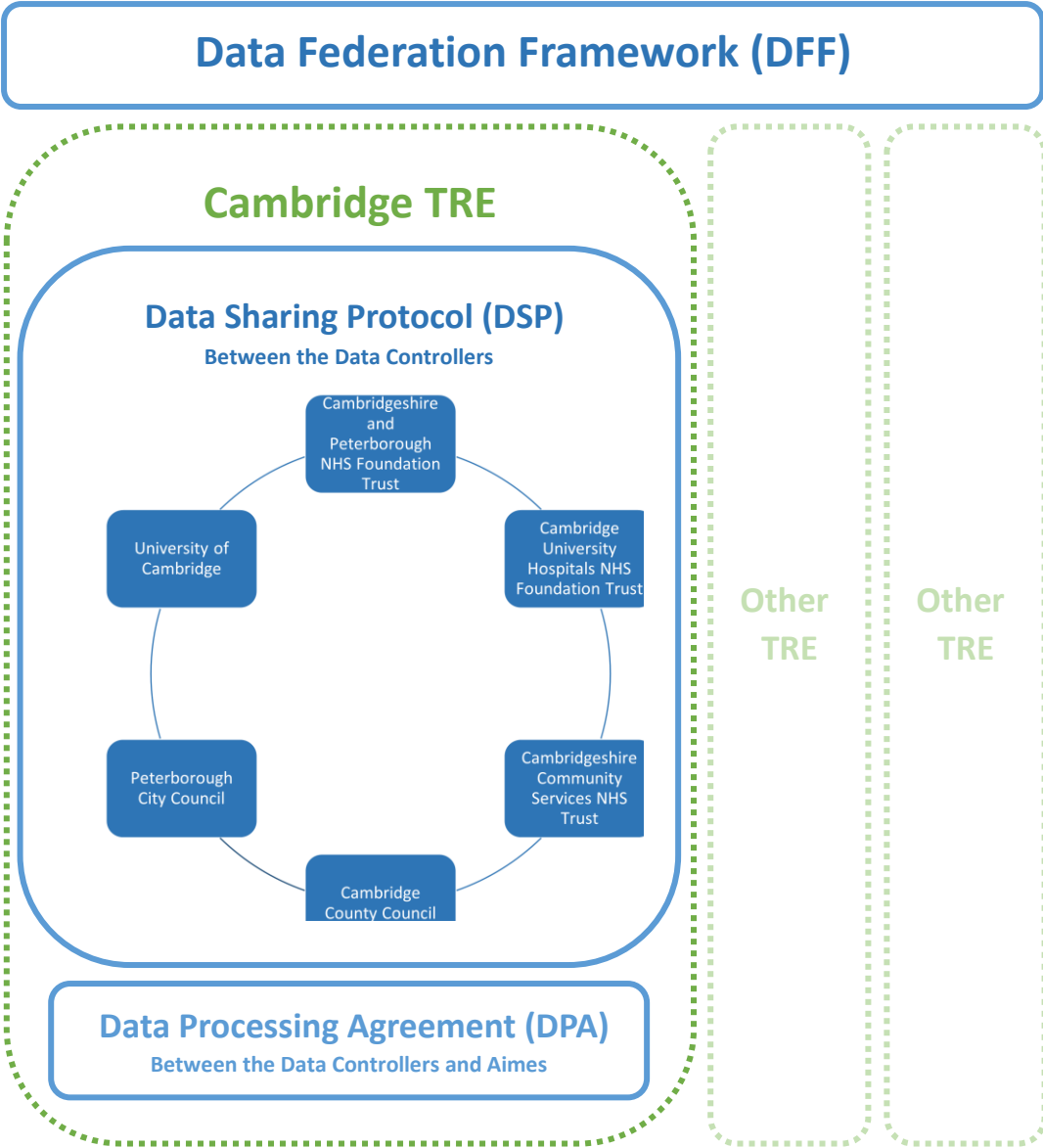
The diagram below illustrates the end to end process to extract, share and process the data in the Trusted Research Environment and how that model would be deployed among different stakeholders.



4.1. Data Sharing Framework

The UK GDPR, under Article 26, requires joint controllers (i.e., controllers that jointly determine the purposes and means of processing) to have an arrangement between them that transparently determines their respective responsibilities for compliance with the obligations under data protection legislation, in particular as regards the exercising of the rights of the patients and their respective duties to provide the fair processing information. Importantly, data sharing agreements, even when not mandated by law, are considered by the Information Commissioner’s Office (ICO) as good practice for the purpose of demonstrating their compliance with the accountability obligations under the UK GDPR. Under Article 28, the UK GDPR also requires that the relationship between one or more data controllers and a data processor is governed by a binding contract that includes a number of mandatory provisions.

We propose that a two-level data sharing framework is established to underpin the complex relationship between the different stakeholders both within and beyond the boundaries of the Cambridge Trusted Research Environment (TRE), as follows:



In accordance with the visual representation above, the data sharing framework will be comprised of the following:

a) Top Level: the Data Federation Framework (“DFF”) will be conceived to regulate the proposed data federation model envisaged between the Cambridge TRE and other TREs under development. In general, the federation model will be operationalised using a third-party system that will allow approved individuals to query the databases contained in each of the participating TREs for the purposes of conducting their research analysis.

From a data protection perspective, due to the fact that the Cambridge TRE will not, and other TREs will very unlikely, have the status of a legal entity, the DFF will constitute an agreement between all the Data Controllers contributing to the different research databases. Among other things, the DFF will establish the Data Federation, determine the criteria for other TREs to join the framework, regulate the governing bodies, specify the criteria and process for approval of requests to access federated data, and define the general principles and standards that each data controller shall adhere to and the reciprocal responsibilities that they owe in respect of the shared data.

b) Lower Level: the lower level will specifically on the relationship between the different stakeholders acting within the Cambridge TRE. From a data protection standpoint, this will require the following:

- Data Sharing Protocol (“DSP”): the DSP will regulate the sharing of data between the different Data Controllers (Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge University Hospitals NHS Foundation Trust, Cambridgeshire Community Services NHS Trust, Cambridge County Council, Peterborough City Council, University of Cambridge) for the purposes of establishing a research database.

The DSP will determine the criteria for other organisations to join the framework as members, set out the process for the curation of other datasets, regulate the governing bodies, specify the criteria and process for approval of requests to access the shared data within the secure environment. It will also determine the data which will be processed, the purposes for processing, the legal basis under the UK GDPR and the common law duty of confidentiality, the nature of the relationship between the parties (i.e., data protection roles). Finally, it will establish the general principles and standards that each data controller shall adhere to and the reciprocal responsibilities that they owe in respect of the shared data.

- Data Processing Agreement (“DPA”): a DPA will be put in place between the data controllers part of the DSP and each third-party organisation (including but not limited to Aimes and Bitfount) that, by processing data on their behalf and under their strict instructions, performs the role of a Data Processor under data protection legislation. The DPA, which will be binding on the Data Processor, will satisfy all requirements under Article 28 of the UK GDPR and, among other things, set out the subject-matter and duration of the processing, the nature and purpose of the processing, the type of personal data and categories of data subjects and the obligations and rights of the controller.

The DFF, DSP and DPA will be drafted in compliance with data protection legislation (including the UK GDPR and the Data Protection Act 2018) and with guidance from the Information Commissioner’s Office (e.g., *Data sharing: a code of practice, Guide to the UK General Data Protection Regulation (UK GDPR)*).

4.2. Terms of Reference (ToR)

As explained above, the proposed data sharing framework will require the creation of governing bodies that separately operate in the context of the Cambridge TRE and of the data federation.

In the specific case of the Cambridge TRE, we envisage the creation of two governing bodies:

- **Strategic Level:** a steering group comprised of representatives from all the data controllers shall have the right and power to decide, among other things, upon:
 - requests made by third-party organisations to join the framework as members;
 - data curation projects that involve bringing additional into the database;
 - any amendment proposed to the framework’s DSP or DPA;

- other requests of strategic nature.
- **Operational Level:** a data access committee comprised of representatives from all the data controllers and other stakeholders (e.g., lay persons) will be responsible for reviewing and approving data access requests made by researchers that satisfy the criteria established in the DSP.

In the case of the data federation, we envisage the creation of a separate governing body comprised of representatives from the different TREs integrating the federation. Working at an **operational level only**, this body will be responsible for reviewing and approving requests for access to data across the federation, made by researchers that satisfy the criteria established in the DFF. **Strategic level** decisions, including any amendment proposed to the DFF, will need to be taken to each TRE's own strategic level governing body for deliberation and approval.

Terms of Reference (ToR) will be drafted to regulate each of the mentioned governing bodies, establishing rules, among other things, about membership, appointment of Chair, frequency of meetings, quorum for deliberation and approval, powers and responsibility and accountability (e.g., reporting obligations).

4.3. Data Access Request Form and Terms of Use

Data Access Request forms will be drafted and implemented to capture all relevant information about each research project that would allow the governing bodies to consider the request and decide whether it satisfies the criteria set out in the framework. Among other things, the Data Access Request forms will require researchers to input:

- indication of the participating data controller sponsoring the application;
- information about the project leads;
- information about the project (e.g., title, purpose, public interest, whether it is commercially funded, whether ethics approval is required);
- information about the individuals who will need access to the data (e.g., their respective affiliations, confirmation that they have completed the mandatory Information Governance training, confirmation that they have signed the Terms of Use document, confirmation that all staff within their organisation work under appropriate confidentiality clauses);
- information about potential conflicts of interest.

The Terms of Use document, as listed above, will consist in a separate document setting out the specific terms in accordance with which researchers will be able to access the research database via the TRE for the purposes of their approved project. This will require the researchers to undertake to a series of conducts, including to:

- only access the data via the VDI solution;
- use the data under strict confidentiality;
- not attempt to re-identify patients without prior approval;
- not attempt to extract data from the secure environment in any way;
- not attempt to link the data with other data without permission;
- keep password credentials in absolute confidence and never share them with other individuals in any circumstances;
- prevent 'shoulder surfing' (i.e., allowing the data to be viewed by other individuals standing behind the User);

- comply with all relevant policies and laws;
- notify cases of known or suspected data security incidents, and/or violations of terms set out in the Agreement.

4.4. Data Pseudonymisation, Anonymisation and Extraction Policy

The UK GDPR, under Article 89, subjects the processing of data for scientific research purposes to appropriate safeguards protecting the rights and freedoms of the patients. The law explicitly requires the data to be anonymised where the purposes for processing can be fulfilled in that manner and, where it cannot, the law encourages the use of pseudonymisation as a technical security measure to ensure respect for the principle of data minimisation.

The research database which will be created through the collaboration of the six data controllers will inevitably require the processing of personal identifiable information, even if it is just to enable the de-identification process to take place. The construction of de-identified longitudinal, integrated health and social care record, through the linkage of data coming in from the different data controllers will without any doubt fall under the scope of current data protection legislation. In order to ensure compliance with the law, technical security measures shall be implemented to ensure that the data contained in the research database is de-identified. The TRE will ensure that the data made available to approved researchers for consultation, manipulation and analysis does not allow the identification of data subjects.

The proposed Data Pseudonymisation, Anonymisation and Extraction Policy will determine the rules that must be complied with to ensure that pseudonymisation is effectively applied as a security measure to protect the data and ensure confidentiality. The Policy will be drafted to ensure that, with the exception of a very circumscribed number of senior managers that will have access to the additional information which will allow re-identification, the dataset contained in the locked environment is regarded as anonymised in the hands everyone one else, including all researchers that obtain the necessary approvals to access the environment and conduct their studies. Recognising the factual reality that research often requires the extraction of aggregated data resulting from the analysis of the dataset, the Policy will further establish the process required for the extraction of data and the anonymisation techniques that must be strictly adhered to as a means of ensuring that the data leaving the environment can survive the “motivated intruder test” and effectively hinder attempts to re-identify data subjects.

The Policy will be drafted in line and in accordance with any existing policies on this subject-matter from all organisations that are part of the project, of course it will also be in compliance with all relevant data protection legislation (including the UK GDPR and the Data Protection Act 2018) and with guidance from the Information Commissioner’s Office (e.g., *Anonymisation: managing data protection risk code of practice*, *Guide to the UK General Data Protection Regulation (UK GDPR)*).

4.5. Information Security Model – Standard Operating Procedures

The UK GDPR, under Article 32, requires data controllers to implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk posed by the corresponding processing activities. For this, controllers must take into account the state of technological development, the costs of implementation, the nature, scope, context and

purposes of processing, the risks that are presented by processing and the risk of varying likelihood and severity for the rights and freedoms of natural persons.

The Information Security Model document will include the technical and organisational measures implemented with the aim of preventing unauthorised or unlawful processing, accidental loss, destruction or damage. It will specify the encryption measures which will be put in place, as well as those designed to ensure the ongoing confidentiality, integrity, availability and resilience of the both the research database and the secure environment. It will also highlight the back-up measures conceived for restore the availability and access to data in a timely manner in the event of a physical or technical incident. Finally, it will indicate how the effectiveness of the measures implemented will be monitored, assessed and improved regularly to mitigate any risks.

This Information Security Model will be included in the Standard Operating Procedures it underpins. Such a document will generally outline the procedural controls for informational purposes. It will function as a comprehensive handbook, including guidance and specifications not only about the data sources, the data flows, the technical infrastructure and security measures, but also about how the environment is hosted, controlled and accessed, what is the information governance framework in place, how to handle incidents and subjects access requests, and what are the backups and disaster recovery procedures.

4.6. Privacy / Fair Processing / Transparency material

The processing of personal data does not only have to be lawful, but it must also be fundamentally fair and transparent. Articles 13 and 14 of the UK GDPR specify what individuals have the right to be informed about. Essentially, people should know from the start who the controllers are, as well as how and why they are collecting and processing their data, all in a way that is easily accessible and easy to understand with clear and plain language. This information is even more crucial when child data is being processed.

The privacy notices will provide the public with all the relevant information, such as:

- the details of the organisation;
- the purposes of the processing;
- the lawful bases;
- the types of personal data;
- how it is obtained;
- if/how it is shared or transferred;
- for how long;
- data subjects rights and how to exercise them.

In order to ensure compliance with legislation, these fair processing materials will be drafted in concise, transparent, intelligible and easily accessible manner, using clear and plain language.

4.7. Data Protection Impact Assessment (DPIA)

The UK GDPR, under Article 35, requires data controllers to carry out, prior to commencing the project, an assessment of the impact of the corresponding processing activities on the protection of personal data whenever it is likely that the processing will result in a high risk to the rights and freedoms of natural persons.

The Data Protection Impact Assessment (DPIA) will systematically review all processing activities relating to the FAIR TREATMENT project, contrasting their necessity and proportionality against the envisaged purposes, assessing the risks to the rights and freedoms of data subjects and the measures conceived to address the risks.

The level of risk attributed to both the impact on the rights and freedoms of the individuals and the likelihood of those rights and freedoms being compromised will be determined after an analysis conducted through the following sections:

- project purpose and necessity;
- data requirements;
- legal basis;
- compliance with the Caldicott Principles;
- data storage;
- external data transfer;
- data accuracy and retention;
- security, integrity and confidentiality;
- consultation; and
- data subjects rights.

In particular, we will conduct a thorough legal analysis of the justification for processing and sharing data for this specific project to ensure that such processing is compatible with the purpose for which the personal data was collected in the first place and that the data is being shared in a lawful, harmonised, safe and secure manner by the organisations involved. The data sharing process will be built with privacy in mind and according to the above-mentioned bespoke sharing framework.

Proposed solutions and actions will be included in such assessment to result in the risks being accepted, reduced or eliminated.

5. Compliance of the Proposed Model with the Six Safes

The proposed Information Governance Model complies not only with data protection legislation, as highlighted in the previous section, but also with the “Six Safes” applicable to Trusted Research Environments, as follows:

Safe people	<ul style="list-style-type: none">✓ Public engagement throughout all steps of the project will take place so that different opinions are heard to ensure a transparent process.✓ Fair processing materials will be published so that individuals are aware of their rights and how to exercise them.✓ Only specific individuals with the necessary credentials as approved by the relevant data access committee(s) will be granted access to the Trusted Research Environment. Their responsibilities will be clearly determined under a Terms of Use document, and they will be bound by such terms. They will receive appropriate training and all accesses will be kept under regular monitoring.
Safe projects	<ul style="list-style-type: none">✓ Data Access Request form will be adopted to capture all relevant information about each project, including its purpose, funder/sponsor information, ethics approvals and time period of access.✓ Extensive guidance will be made available online about the data access request process, including requirements and decision-making process.✓ Data access committees will have meaningful involvement of lay representatives.✓ A public data use register will be platformed online and regularly with newly approved projects.
Safe setting	<ul style="list-style-type: none">✓ The Information Security model will include all technical and organisational measures implemented to ensure that the data is held and managed securely. The document will further detail how the secure environment will allow individuals to perform their analysis of the data, using in-built tools, and will not allow data to be transferred, copied or otherwise extracted.
Safe data	<ul style="list-style-type: none">✓ All data in the TRE will be de-identified. Guidance on pseudonymization and anonymization will also be provided and considered as part of the project.

	<ul style="list-style-type: none"> ✓ Data controllers will agree to be part of a sharing framework to commit to only sharing data under appropriate safeguards which complies with the relevant data protection, privacy and confidentiality regulations.
Safe outputs	<ul style="list-style-type: none"> ✓ The Data Extraction Policy will establish the process required for the extraction of data and the anonymisation techniques that must be strictly adhered to as a means of ensuring that only fully anonymised data leaves the secure environment.
Safe return	<ul style="list-style-type: none"> ✓ A separate “Consent to Contact” is currently under discussion. If implemented, it will sit parallel to the research database, with robust processes and systems being put in place to support it.

7.4. Appendix 4: Information Governance model comparison



INFORMATION GOVERNANCE SERVICES

Helping you make the most of your data



Furlong House, 10A Chandos Street,
London, United Kingdom, W1G 9DQ



info@informationgovernanceservices.com
www.informationgovernanceservices.com

Cambridge Trusted Research Environment (TRE)

Information Governance Model Comparison

Table of Contents

1. Introduction	2
2. Definition of data controller, joint-controller and independent controller	2
2.1. Criteria for determination of data protection roles.....	2
2.2. The concept of data controller	3
2.3. Relationship between data controllers.....	3
a) Joint-controllership	3
b) Independent controllership	5
3. Cambridge TRE – Information Governance Model Comparison.....	5
3.1. The FAIR TREATMENT project and the organisations participating therein.....	5
3.2. Review of Information Governance Model 1.....	6
3.3. Review of Information Governance Model 2.....	8
3. Recommendation.....	10
4. Conclusion.....	10

1. Introduction

FAIR TREATMENT (“Federated analytics and AI Research across TREs for Adolescent MENTAL health”) is a project sponsored/led by the University of Cambridge which aims to: 1) combine two new technologies to demonstrate it is possible to analyse data across trusted research environments in different places and preserve the privacy of individuals; 2) consult with patients, the public, organisations contributing data, and legal/ethics experts to agree the best way to oversee data use, ensuring it’s managed safely and fairly.

Ethical permission has been granted to construct a linked whole-population, de-identified, database of electronic patient record data in Cambridge and Peterborough (NHS REC ID: 20/EM/0299) called Cam-CHILD. This includes data from five other organisations mentioned below. Cam-CHILD will be replicated in Essex and Birmingham, with equivalent ethics applications submitted for both.

Different health and care organisations have been asked to participate in the project, by contributing with data for the creation of the research database and potentially taking part in the decision-making process in regard to the information governance elements applicable thereto. This includes the following organisations: Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge University Hospitals NHS Foundation Trust, Cambridgeshire Community Services, Cambridge County Council and Peterborough City Council.

We, Information Governance Services Ltd (“**IGS**”), have been commissioned to assist in the development of an information governance model to support the data controllers in the implementation of the research database.

A high-level document was initially drafted with the aim of providing an overview of the envisaged model and presented to IG representatives from each of the participating organisations. We have been asked by the group to draft a legal report assessing, in light of the requirements under data protection legislation, two alternative information governance models which could potentially be adopted to support the project.

2. Definition of data controller, joint-controller and independent controller

2.1. Criteria for determination of data protection roles

In accordance with guidance from the European Data Protection Board (“**EDPB**”), the data protection roles are essentially functional concepts¹. Ascertaining whether a party performs a specified data protection role (e.g., data controller) requires an in-depth analysis of the factual elements of each case in light of the definitions provided by the law, which will include any relevant case law.

As opposed to many other contractual relationships, the allocation of these roles is not negotiable. This means that, whilst a contract may specify the rights and obligations of each party and, upon doing so, govern the relationship between them and assist in the determination of the roles allocated to each, the formal designation of data controller given to

¹ [Guidelines 07/2020 on the concepts of controller and processor in the GDPR](#), Version 2.0, Adopted 07/07/2021.

a party in a contract, in itself, will not suffice if it is in any way contrary to the legal definition of that specific role.

2.2. The concept of data controller

According with Article 4(7) of the UK GDPR, 'controller' means "*the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data*".

As we can see from the definition, a controller is responsible for determining both the purposes (i.e., the why) and means (i.e., the how) of the processing. It is the organisations that exercise decision-making power over certain key elements of such processing.

According to guidance from the Information Commissioner's Office ("ICO"), data controllers are responsible for deciding on matters such as²:

- whether to collect personal data in the first place;
- what the lawful basis for doing so is;
- what types of personal data to collect;
- what is the purpose for using the data;
- which individuals to collect data about;
- whether to disclose the data, and if so, to whom;
- what to tell individuals about the processing;
- how to respond to requests made in line with individuals' rights; and
- how long to retain the data or whether to make non-routine amendments to the data.

The cited guidance from the ICO makes it clear that these are all decisions that can only be taken by the controller as part of its overall control of the data processing operation, which consequently means that an organisation will likely be regarded as a data controller if it makes any of these decisions determining the purposes and means of the processing.

2.3. Relationship between data controllers

The UK GDPR recognises that more than one organisation may act as a data controller. In some cases, different organisations may jointly determine the purposes and means of the processing of personal data, being characterised as "joint-controllers". In other cases, different organisations may process the same personal data for different purposes and potentially through different means, acting as "independent controllers". The two concepts will be further analysed below:

a) Joint-controllership: the concept of joint-controllership reflects the factual circumstance whereby, in accordance with Article 26 of the UK GDPR, two or more controllers jointly determine the purposes and means of the processing of personal data.

According to the EDPB, joint controllership "*can take the form of a common decision taken by two or more entities or result from converging decisions by two or more entities regarding the*

² Guide to the General Data Protection Regulation (GDPR)/Controllers and processors/[How do you determine whether you are a controller or processor?](#)

*purposes and essential means*³. The adoption of a common decision reflects a situation in which the parties essentially decide together, whereas the adoption of converging decisions results from a process where each parties' decisions "*complement each other and are necessary for the processing to take place in such manner that they have a tangible impact on the determination of the purposes and means of the processing*"⁴.

Whilst the adoption of common or converging decisions in regard to the determination of the purposes and means of processing may be the ordinary practice, there are extraordinary circumstances in which joint-controllership may otherwise be recognised in the absence of these types of decision.

In this sense, the EDPB explicitly recognises that a joint determination of the means of processing may be characterised even in the absence of common or converging decisions, where a single organisation provides the means and makes it available to other organisations, who then decide to make use of those means for a converging purpose:

*'It may also be the case that one of the entities involved provides the means of the processing and makes it available for personal data processing activities by other entities. The entity who decides to make use of those means so that personal data can be processed for a particular purpose also participates in the determination of the means of the processing*⁵.

Further to this, case law from the Court of Justice of the European Union ("**CJEU**") appears to indicate that joint-controllership may be characterised even in the complete absence of common or converging decisions in regard to the purposes and means of processing, where there is a mutual benefit arising from the same processing operation, as long as each of the organisations involved participates in the determination of the purposes and means of the relevant processing operation.

In the *Fashion ID* case⁶, for example, the CJEU held that a website operator participates in the determination of the purposes and means of the processing by embedding a social media plug-in on their website (in order to optimize the publicity of its goods by making them more visible on the social network) which causes the visitor's browser to capture and transmit personal data to the provider of the plug-in. The CJEU considered that the processing operations at issue were performed in the economic interest of both the website operator and the provider of the plug-in, regarding them as joint-controllers. The Court, however, made it clear that the website operator's liability is limited to the operation or set of operations involving the processing of personal data in respect of which it actually determines the purposes and means, that is to say, the collection and disclosure by transmission of the data at issue.

Likewise, in the *Facebook Fan Page* case⁷, the Court regarded Facebook and the administrator of a fan page hosted on the platform as joint-controllers. According to the Court, whilst the data processing activity at issue was essentially carried out by Facebook placing

³ [Guidelines 07/2020 on the concepts of controller and processor in the GDPR](#), Section 3.2.2, Version 2.0, Adopted 07/07/2021.

⁴ *Idem*, para 55.

⁵ *Idem*, para 64.

⁶ Judgment of 29 July 2019, *Fashion ID GmbH & Co. KG*, Case C-40/17, EU:C:2019:629, paragraph 85.

⁷ Judgment of 5 June 2018, *Wirtschaftsakademie Schleswig-Holstein GmbH*, Case C-210/16, EU:C:2018:388, paragraph 85.

cookies on the computer or other device of persons visiting the fan page, the creation of the fan page required the definition of parameters by the administrator, depending inter alia on the target audience and the objectives of managing and promoting its activities, which had an influence on the processing of personal data for the purpose of producing statistics based on visits to the fan page. With the help of filters, the administrator defined the criteria in accordance with which the statistics are to be drawn up and even designate the categories of persons whose personal data is to be made use of by Facebook. Therefore, each entity in this case pursued its own interest but both parties participated in the determination of the purposes and means of the processing of personal data as regards the visitors to the fan page.

b) Independent controllership: the concept of independent controllership reflects the factual circumstance whereby two or more controllers separately determine the purposes and means of the processing of personal data. As with all other data protection roles, such qualification requires a case-by-case analysis of each processing activity and the exact role performed by each organisation with respect to each processing.

It is perfectly possible, and commonly seen in practice, that various organisations successively process the same personal data in a **chain of operations**, with each organisation processing data for an independent purpose and through independent means in their part of the chain. Due to the absence of a joint determination of purposes and means of the same processing operation or set of operations, the two or more organisations could be regarded as successive independent controllers.

3. Cambridge TRE – Information Governance Model Comparison

3.1. The FAIR TREATMENT project and the organisations participating therein

Several different health, care and educational organisations will participate in the creation of the Cambridge research database for the FAIR TREATMENT project, as follows:

- Cambridgeshire and Peterborough NHS Foundation Trust
- Cambridge University Hospitals NHS Foundation Trust
- Cambridgeshire Community Services
- Cambridge County Council
- Peterborough City Council
- University of Cambridge

The first five organisations are all health and social care providers which, in the exercise of their statutory duties, collect and further process personal data from data subjects for the purpose of providing them with health and social care. Whenever discharging such functions, these organisations determine the purpose and means of processing of personal data for the specified purposes and thus adopt the role of data controller under the UK GDPR.

The sixth organisation is a higher education provider which, despite not playing a role in the provision of direct health and care to data subjects, may, alone or jointly with other organisations, sponsor and set up studies and databases for scientific research purposes. When discharging these functions, the University will, alone or jointly with others, determine the purpose and means of processing of personal data for the specified purposes and thus adopt the role of data controller under the UK GDPR.

The relationship between the six organisations in the context of the FAIR TREATMENT project may vary depending on which information governance model is ultimately adopted. We will conduct a legal review of two alternative models, scrutinising each of them against data protection legislation and ascertaining whether they would constitute legally viable options.

3.2. Review of Information Governance Model 1

Information Governance Model 1 involves regarding all six organisations as controllers of the data contained in the database for research purposes.

As further described below, we commence by ascertaining whether it would be legally sound to consider the six organisations as independent data controllers of the research database and, after concluding that the factual circumstances of the case would very likely negate such legal characterisation, we proceed to consider whether the six organisations could be regarded as joint-controllers of the research database.

a) Independent data controllers: in the first instance, we explored whether the six organisations could be regarded as independent data controllers.

From the outset, it is important to note that a model based on independent controllership does not seem to satisfactorily accommodate the envisaged role performed by the University of Cambridge. By not constituting a health and care provider and not originally being a controller of any data being made available for research, the University would not technically control any data and would therefore not be deemed an independent data controller. Whilst it could be theoretically possible to regard the University as a data processor, were this organisation to be responsible for managing the research database on behalf of and under strict instructions from each of the independent data controllers, such characterisation does not appear to coincide with the factual circumstances of the case and the desire of the parties to the project.

Further to this, the characterisation of the five health and care providers as independent data controllers would arguably be equally challenging. For starters, we would need to ensure that each of the five are solely responsible for making certain key decisions about the processing of their data.

Firstly, these organisations would need to generally decide whether to make data under their controllership (i.e., personal data collected from data subjects and further processed for the purpose of providing them with health and social care) available for research purposes. This would be practically achieved by extracting the data from their own internal systems (e.g., Electronic Patient Record, in the case of healthcare providers) and having it hosted in a dedicated secure environment. In this case, this would be a common environment, where all the data from the different health and care providers would be stored under appropriate access controls capable of ensuring segregation.

Secondly, these organisations would need to concretely decide whether to make such “repurposed” data (i.e., data originally collected for health and care purposes and now repurposed for research) available for specific researchers and projects. In this case, even if we were to conceive a Data Access Committee responsible for reviewing data access requests made by researchers, a model predicated on independent controllership would require each organisation to retain the power to independently decide whether, or not, to participate in each project referred to the Data Access Committee. This means that, rather

than having the Committee decide on the basis of majority, we would have a model in which each organisation participating in the Committee would have the power to allow or veto the use of their data in any given case.

At first sight, it would appear that this factual setting could be compatible with the envisaged “independent controllership” model. In this sense, guidance from the EDPB indicates that “*Joint controllership may also be excluded in a situation where several entities use a shared database or a common infrastructure, if each entity independently determines its own purposes*”⁸. Even if different organisations use a shared database, the EDPB makes it clear that they may still be regarded as independent data controllers if they each enter the data of their own data subjects and process such data for their own purposes only. In such case, each organisation would need to decide independently on the access, the retention periods, the correction or deletion of their data subjects’ data. And they would not be able to access or use each other’s data.

A closer review of the model, however, casts doubt as to whether it could survive scrutiny against the “mutual benefit” assessment established under case-law from the CJEU. As discussed in section 2.3(a) of this document, joint-controllership may be characterised even in the absence of common or converging decisions in regard to the purposes and means of processing, where each of the organisations participates in the determination of the purposes and means of the relevant processing operation and there is a mutual benefit arising from this processing operation.

Even if each data controller retained approving and vetoing powers when reviewing data access applications to the database, it is conceivable a mutual benefit could be derived from each organisation hosting their data in a shared platform and making it available for research purposes. This is due to the fact that having a shared platform is not a prerequisite for the participating organisations to make data available for research. Even without such platform, each individual organisation already holds the decision-making power and the means to allow data under their controllership to be made available for research purposes (for example, a number of sponsor organisations regularly enter into agreements with NHS organisations requiring access to data for scientific analysis). By having their data platformed in a shared environment, more data is potentially made available to researchers through a single application mechanism, which inevitably makes the database attractive to researchers. The organisations mutually benefit from this, as a minimum, by receiving credits in all published scientific papers that involved the use and analysis of data from the database.

In addition to this, the characterisation of the five organisations as independent data controllers of the research database could be called into question where other decision-making powers are held by these organisations. The more the organisations jointly decide upon strategic matters pertaining to the research database (e.g., curating additional datasets, bringing in additional partner organisations as controllers, changing the supplier of the data hosting platform), the more likely that the factual circumstances will point to a characterisation of joint, rather than independent, data controllers.

Therefore, for all the above reasons, our view is that an information governance model based on the independent controllership of the parties should not be adopted for the desired purpose.

⁸ [Guidelines 07/2020 on the concepts of controller and processor in the GDPR](#), para 71, Version 2.0, Adopted 07/07/2021.

b) Joint data controllers: we then proceed to consider whether the six organisations could be regarded as joint-controllers of the research database.

In our view, joint determination of the purposes of processing could be established where all organisations will, via either common or converging decisions, abstractly stipulate what criteria should projects satisfy in order for the respective researchers to be able to apply for access to data and, on the basis of majority, concretely decide which applications should be approved. In furtherance to this, any and all other strategic decisions, including the curation of additional datasets, would be jointly made by the participating parties.

It is also our view that joint determination of the means of processing could also be established in the present case. As discussed in section 2.3(a) of this document, the EDPB explicitly recognises that a joint determination of the means of processing may be characterised even in the absence of common or converging decisions, where a single organisation provides the means and makes it available to other organisations, who then decide to make use of those means for a converging purpose. Consequently, the fact that, in the present case, the University of Cambridge took the initiative of engaging with all the technology partners and thus provided the means should not pose as an obstacle to the characterisation of joint-controllership, considering that the adoption of such means by all the organisations for a converging purpose can adequately satisfy the legal requirement.

Therefore, provided that the above coincides with the practical/factual management and operationalisation of the research database, our view is that a model based on joint-controllership between the six organisations could be legally viable.

In order to ensure its compliance with the law, however, a Data Sharing Protocol between the six organisations would be necessary to ensure compliance with Article 26(1) of the UK GDPR. Such provision requires joint-controllers to determine, in a transparent manner, their respective responsibilities for compliance with their legal obligations, in particular as regards the exercising of the rights of the data subject and their respective duties to provide the transparency / fair processing information.

Whilst Article 26(3) of the UK GDPR does make it clear that “*Irrespective of the terms of the arrangement referred to in paragraph 1, the data subject may exercise his or her rights under this Regulation in respect of and against each of the controllers*”, it is important to note that EU case-law firmly establishes that the existence of joint liability does not necessarily imply equal responsibility of the various operators engaged in the processing of personal data. According to the CJEU, operators may be involved to different degrees, with the result that the level of liability of each of them must be assessed taking into account all circumstances of the particular case⁹.

3.3. Review of Information Governance Model 2

Information Governance Model 2 involves regarding the University of Cambridge (sponsor/ lead organisation) and Cambridgeshire and Peterborough NHS Foundation Trust as the data controllers of the database for research purposes. Applying the same rationale as that from section 3.2(a) and (b) of this document, these two organisations would be regarded as joint-

⁹ Judgment of 10 July 2018, *Jehovan todistajat*, C-25/17, EU:C:2018:551, paragraph 66.

controllers, by jointly determining the purposes and means of processing of data in the context of the research database.

All the other four health and care organisations, which unquestionably act as data controllers when collecting and processing data for the purpose of discharging their statutory duties in the area of health and social care (see section 3.1 of this document), would be asked to share data to support the creation of the research database. In this sense, Cambridge University Hospitals NHS Foundation Trust, Cambridgeshire Community Services, Cambridge County Council and Peterborough City Council (as data controllers for health and care purposes) would be asked to separately engage with the University of Cambridge and Cambridgeshire and Peterborough NHS Foundation Trust (as joint-controllers for research purposes) in a relationship whereby the personal data will be successively processed in a chain of operations, with each processing data for an independent purpose and through independent means in their part of the chain.

In this model, the health and care organisations would have decision making powers as to whether to share data to the research database in accordance with the assurances provided by the joint-controllers. Once the data is shared, however, the University of Cambridge and Cambridgeshire and Peterborough NHS Foundation Trust, as joint-controllers of the data for research purposes, would be solely responsible for making key decisions about the processing of the data. Among other things, this would include the unilateral power to decide upon which data processors to engage with and whether to curate and bring additional datasets to the database. Importantly, it would also include the power to establish and change the criteria applied by the Data Access Committee when reviewing and approving applications made by researchers.

Provided that Cambridge University Hospitals NHS Foundation Trust, Cambridgeshire Community Services, Cambridge County Council and Peterborough City Council agree to share data with the University of Cambridge and Cambridgeshire and Peterborough NHS Foundation Trust as described above, thus relinquishing the power to determine the purposes and means of all data processing for research purposes, we believe that this would also constitute a legally sound model that could stand scrutiny against the requirements under data protection legislation.

In our view, it could be legally acceptable, in specific circumstances, for the joint-controllers to invite members from the contributing health and care organisations to comprise the Data Access Committee without these organisations being regarded as joint-controllers, where the terms of reference under which the Committee would act is strictly defined by the controlling organisations beforehand. However, it is important to note that this would always remain at the discretion of the University of Cambridge and Cambridgeshire and Peterborough NHS Foundation Trust as joint-controllers, which would retain the right and power to unilaterally change the composition of the Committee and the terms under which it operates.

However, the more powers these other health and care providers claim over the processing of data for research purposes, the more likely they will be regarded as joint-controllers, which could compromise the robustness of this information governance model and generate risks to all parties involved.

3. Recommendation

Comparing the two information governance models which we have explored in this report, we strongly recommend the adoption of Information Governance Model 1, as this model gives all contributing health and care organisations more power with respect to the use of their data.

Additionally, the model has a proven track record in large projects. The model is currently being used by over 400 data controllers, which include GPs, Acutes, Mental Health Trusts, Community Services and Local Authorities in North West London, to govern an integrated care record for the delivery of direct patient care. That integrated care record is called Whole Systems Integrated Care (“**WSIC**”). Whilst we acknowledge that an integrated care record is substantially distinct from the research database being created for the FAIR TREATMENT project, the de-identified version of WSIC, which is called Discover-NOW, allows researchers to undertake research projects on de-identified data from circa 2.3 million people.

In the context of Discover-NOW, data access requests begin with an application being made by the relevant researcher through the completion of a detailed form, which is then referred to a data access committee for consideration and approval. Whilst the Committee includes representatives from all data controllers, a minimum number of which being required for each meeting to be quorate, it also includes services users/lay members, researchers, clinicians, information governance professionals, technology experts and data scientists. Applicants then attend the meeting in which their application is being considered, where they will be required to justify the merits of their application, explain why their research is important and answer any questions the committee members might have. The committee would then make a decision based on all available information and, if happy, grant access to the data for a set period.

Other than very small updates due to legislation, WSIC has been operating under this model for over 6 years and Discover-NOW for 3 years.

Finally, our view is that Model 1 allows for greater scalability, when compared to Model 2. This is due to a general reluctance of health and care organisations to simply hand off data, as required in Model 2, which could jeopardise the longer-term growth of the database (via curation of additional datasets - e.g., GP Practices have historically been reluctant to contribute to research databases without having decision-making power) and its consequential attractiveness for research purposes.

4. Conclusion

It is our view is that the two information governance models explored in this paper could be viable from a legal standpoint, provided that the conditions outlined for each are adequately satisfied. Although we have expressed our recommendation, it is only our place to provide an opinion which is based on our interpretation of the law, and others may interpret it differently. Since the decision, and ultimate responsibility, rests with the data controllers, we can proceed with any of the two models which the partners in this project decide. However, we felt it was necessary to highlight the risks to ensure that the partners have all the information required to make an informed decision.

7.5. Appendix 5: Overview of supporting documents for Information Governance Framework



INFORMATION GOVERNANCE SERVICES

Helping you make the most of your data



Furlong House, 10A Chandos Street,
London, United Kingdom, W1G 9DQ



info@informationgovernanceservices.com
www.informationgovernanceservices.com

Cambridge Trusted Research Environment (TRE)

Proposed Information Governance Documentation

In order to support the governance framework developed as part of the HDR UK DARE Sprint for the Cambridge TRE, IGS recommend that the following documentation is used to support compliance with data protection legislation.

Data Sharing Framework: A two-level data sharing framework to underpin the relationship between the stakeholders both within and beyond the boundaries of the Cambridge TRE. First, the Data Federation Framework (“DFF”) will regulate the proposed data federation model between the Cambridge TRE and other TREs. Then, the Data Sharing Protocol (“DSP”) will regulate the sharing of data between the different data controllers. Finally, a Data Processing Agreement template will be put in place between the data controllers’ part of the DSP and each third-party organisation (including, but not limited to, AIMEs and Bitfount) which, by processing data on their behalf and under their strict instructions, performs the role of a data processor under data protection legislation.

Terms of Reference: The framework will require the creation of governing bodies that separately operate in the context of the Cambridge TRE and of the data federation. The Terms of Reference regulate each of the governing bodies, establishing rules about, among other things, membership, appointment of Chair, frequency of meetings, quorum for deliberation and approval, powers and responsibility, and accountability (e.g., reporting obligations).

Data Access Request Form and Terms of Use: The Forms capture all relevant information about each research project, to allow the governing bodies to consider the request and decide whether it satisfies the criteria set out in the framework. The Terms of Use consist in a separate document setting out the specific terms in accordance with which researchers will be able to access the research database via the TRE for the purposes of their approved project.

Data Pseudonymisation, Anonymisation and Extraction Policy: This Policy ensures that, with the exception of a very circumscribed number of senior managers with access to the additional information to allow re-identification, the dataset contained in the TRE is regarded as anonymised in the hands of everyone else, including all researchers that obtain the necessary approvals to access the TRE and conduct their studies

Information Security Model – Standard Operating Procedures: Includes the technical and organisational measures implemented with the aim of preventing unauthorised or unlawful processing, accidental loss, destruction, or damage. It specifies the encryption measures put in place, as well as those designed to ensure the ongoing confidentiality, integrity, availability and resilience of both the research database and the TRE. It also highlights the back-up measures conceived for restoring the availability and access to data in a timely manner in the event of a physical or technical incident. This document will be completed and updated over the course of the implementation of the project (Attachment X - SOP and security model).

Privacy / Fair Processing / Transparency material: The processing of personal data does not only have to be lawful, but it must also be fundamentally fair and transparent. In essence, people should know from the start who the data controllers are, as well as how and why they are collecting and processing their data. This information is even more crucial when children's data is being processed. Consequently, materials are being drafted in a concise, transparent, intelligible and easily accessible manner, using clear and plain language for the public to understand the data processing and how they can exercise their rights.

Data Protection Impact Assessment (DPIA): We systematically reviewed all processing activities relating to the FAIR TREATMENT project, contrasting their necessity and proportionality against the envisaged purposes, assessing the risks to the rights and freedoms of data subjects and the measures conceived to address the risks (Attachment X - initial DPIA).

7.6. Appendix 6: Patient and Public Involvement workshop slides



Timely Workshop

Today's aims

- 1 Get to know each other
- 2 What is the problem we are trying to solve and how data can help with this?

(Break)

- 3 Understand your views on the Timely project

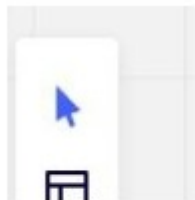
Three things to remember

1. **There are no wrong answers**
2. **You can share your opinions however you want:** out loud, in the chat to the group or using post it notes on this board
3. **There are no silly questions** so please feel free to ask questions anytime, you can also send them in the chat

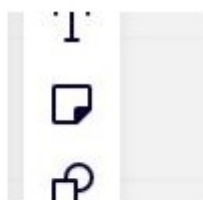
Using Miro

Using Miro is **optional**, but if you'd like to give it a try the link is in the **Zoom** chat.

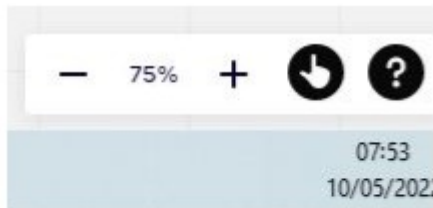
You'll only need to use 3 buttons:



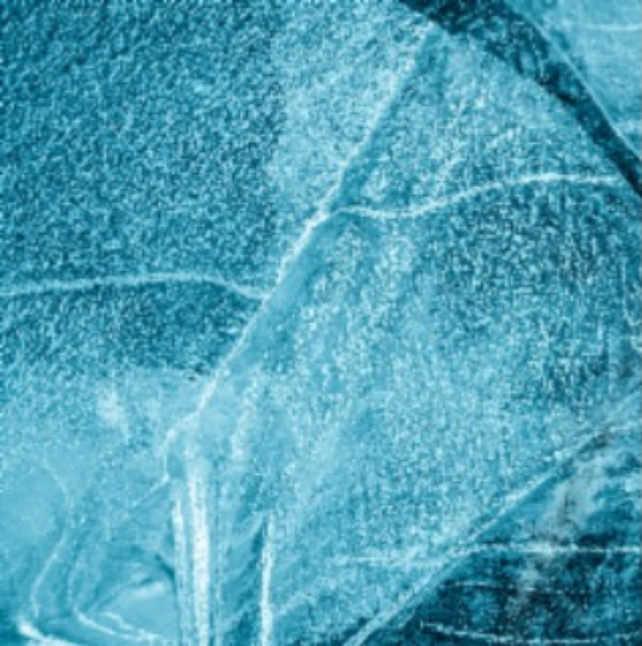
Mouse
Top of the tool
bar



Post-it
4th down on the
tool bar



Zoom
Bottom right of the board



Ice Breaker

You have **1 minute** each to tell the group:

1. Your name
2. Your pronouns
3. The town or city you live in
4. Something interesting that happened to you recently



Who are we?



UNIVERSITY OF
CAMBRIDGE
Department of Psychiatry



Anna Freud
National Centre for
Children and Families



UNIVERSITY OF
BIRMINGHAM



University
of Essex

NHS
Cambridgeshire
Community Services
NHS Trust

NHS

Cambridge
University Hospitals
NHS Foundation Trust

NHS

Cambridgeshire and
Peterborough
NHS Foundation Trust



Cambridgeshire
County Council

PETERBOROUGH
CITY COUNCIL

Who are we?



Emily



Rachel



Anna



Alisa



Tia

What is the problem we are trying to solve?



Studies show that 1 in 6 young people aged 5-16 years had a probable mental health problem in 2021. That's 5 young people in every classroom of 30.

When a young person is struggling with their mental health it can be hard for the adults around them to realise that they need help.

Their doctor or teacher might notice that something is off, but not have the "full picture" of what is happening in their life.

If can we link data together in a database, we might be able to find ways to get young people **the right kind of help** and **at the right time**.



How could
using data
help?

In the UK we have many services that help the public.

Places like...



Schools



Your GP
(doctor)



Accident & Emergency



Hospitals



Social Services



Mental Health
Services

When you go to see someone at one of these services, they keep information about you. For example...



Schools might know if students have been absent a lot

GPs might know if their patients have been having headaches



A&E might know how many young people in the area have broken bones

Social Services might know about families who have been having problems at home



Mental health services in the community might know if more young people in the area have been struggling with depression



Patient data saves lives: Asthma

for a preventative appointment.

YouTube

BUT WHEN RESEARCHERS GET ACCESS TO THIS TYPE OF DATA THEY DO NOT SEE ANY IDENTIFIABLE INFORMATION



Miss Max Jones

Miss o2mKr8a 9Wbas6B

Researchers can't see this

Nonsense numbers & letters

Max Jones	→	73e2c83 2bd927v
NHS# 1234567	→	NHS# 923ensdv28p
CB2 1TN	→	msa72e0zg173

Miss Max Jones, NHS# 1234567, living at CB2 1TN attended A&E with a sprained ankle on the 6th of May, 2017

Miss 73e2c83 2bd927v, NHS# 923ensdv28p, living at msa72e0zg173 attended A&E with a sprained ankle on the 6th of May, 2022

No way for researchers to decode this

That means things like:

- Names
- Addresses
- Phone numbers
- NHS numbers
- Names or addresses of family members

Zoom poll!



Do you have any questions or
comments?

Does anything worry you?

DISCUSS!



Take a break



So...

What is the Timely
project trying to do?



Let's look at
some examples
of what using
linked data
could look like in
the real world...



MAX'S JOURNEY



Max does okay in school, even though they get in trouble sometimes. Their favourite subject is Maths. They like playing football.



One day Max sprains their ankle at a football game. Their grandad takes them to A&F.



A social worker has come to speak to Max about possibly needing a SEND (special education needs & disabilities) plan, but it was decided Max didn't need one.



Max misses a few days of school because their parents take them on a camping trip.



Max's grades in Science start to slip. Their parents get them a tutor and their grades improve.



Max's mum takes them to the GP because they have been complaining about a stomach ache. They are given a prescription.

KEY



What the school know



What the NHS know



What social services know



WHAT THE "FULL PICTURE" LOOKS LIKE WITH ALL THE INFORMATION PUT TOGETHER

Max does okay
in school...

...but gets in
trouble
sometimes

A&E visit...

...but just a
sprained ankle

Visit from social
worker...

...but everything
is fine at home

Routine GP visit.

Missing school
but rarely.

Grades got
worse but
improved.

Computer program looking
at all the information
together



Max is
probably doing
okay...



ALEX'S JOURNEY



Alex is smart, but they gets into fights a lot with other children at school.



Alex's dad has taken them to A&E 5 times in the last year complaining of headaches. Doctors aren't sure what's wrong.



Alex misses school often because they say they can't be bothered and would rather stay home and play video games.



Alex's mum is worried about them. She takes them to a GP appointment. The GP puts in a referral to CAMHS. The next available appointment is in 8 months.



Alex's grades have been slipping for the past year. They don't do much homework and don't do well on tests.



A social worker has been to the family home because the school is concerned. Alex seems quiet and sad, but their parents insist they're just being moody.

KEY



What the school knows



What the NHS knows



What social services know



WHAT THE "FULL PICTURE" LOOKS LIKE WITH ALL THE INFORMATION PUT TOGETHER

Alex gets into fights at school.

Alex has missed a lot of school days this year.

Social worker has been to the house...

Alex is on a CAMHS waiting list, but won't get help for some time.

Alex's grades have slipped...

5 A&E visits but no diagnosis.

...but isn't too concerned.

...but teachers say they're smart

Looks like Alex is struggling... someone needs to check in on them.

Computer program looking at all the information together



SO CAN WE DO THIS?

...not yet!



First we would need to look at lots and lots of data to see if we can **learn to tell apart** “these young people are probably doing okay” and “looks like these young people are struggling.”

The first thing we would need to do is **link** all the different types of data together in a single digital database.

Some of the information we need to include could be quite **sensitive** in nature.

Let's look at some examples...

- 1 **GP:** the young person has a substance misuse problem



- 2 **Mental health:** a child has disclosed abuse; a child is a young carer



- School:** whether a child receives free school meals, or if there are welfare concerns about a child

- 4 **Community services:** vaccination records, dental records, 2 year old health visitor checks.



- 5 **Social care:** if a child is a 'looked after child', if there are safeguarding concerns, if they are involved in county lines

Zoom poll!

What do you think about
the Timely project so far?

DISCUSS!

Today we...

- 1 Got to know each other
- 2 Understood your views on how data can be used
- 3 Understood your views on the Timely project

How are you feeling after today's session?

Vote!



Relaxed

Nervous

Excited

Confused



Next
Steps

In the next couple of months we will:

Workshop 2 (31st May)

Work together to make a plan for how we will share information and keep it safe on the Timely project.

Workshop 3 (21st June)

Exploring ways in which we can explain this to other people clearly

In **July** we'll get a new group of people who know nothing about the project and show them what we've been working on. We'll do a survey to see what they think and we'll analyse the results.

We hope to continue working on this project with you beyond July, although we don't know exactly what that will look like yet!



Timely Workshop 2

Today's aims

- 1 Recap - what is the timely project about?
- 2 Understand your views on keeping data safe
- 3 Understand your views on what kind of people, organisations and projects should be able to use the data.
- 4 Understand your views on who should make decisions about the Timely project data.

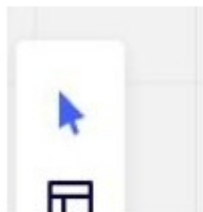
Three things to remember

1. **There are no wrong answers**
2. **You can share your opinions however you want:** out loud, in the chat to the group or using post it notes on this board
3. **There are no silly questions** so please feel free to ask questions anytime, you can also send them in the chat

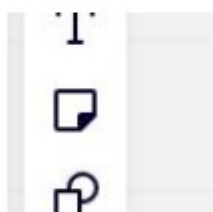
Using Miro

Using Miro is **optional**, but if you'd like to give it a try the link is in the **Zoom** chat.

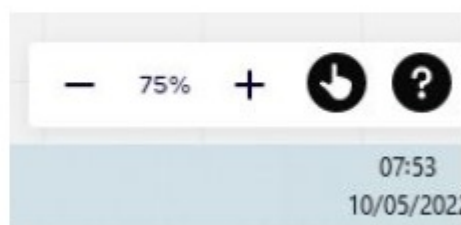
You'll only need to use 3 buttons:



Mouse
Top of the tool
bar



Post-it
4th down on the
tool bar



Zoom
Bottom right of the board

You have **1 minute** each to tell us :

Your name and pronouns

The town or city you live in

If you were marooned on an island, what 3 things would you bring with you?



Ice Breaker



Recap

What is the problem we are trying to solve?

Studies show that 1 in 6 young people aged 5-16 years likely had a mental health problem in 2021.

That's 5 young people in every classroom of 30.

When a young person is struggling with their mental health it can be hard for the adults around them to realise that they need help.

Services (like the NHS, schools, and social services) already collect data which could indicate that something is wrong, but this data is not stored in one place so it's hard to see "the full picture"



What is the Timely project trying to do?



This means that things like your name and date of birth have been removed

The Timely project would like to link **de-identified data** from different services in one database.

We would then like to use this de-identified data to train computer programs ("algorithms") to learn the difference between "this young person is probably doing okay" and "it looks like this young person is struggling".

Right now we are just in the **research** phase of trying to understand **if** we can do this - this is why there is no identifying information in the database



Any questions
so far?



Today's
workshop

It's also important for researchers to see data from different **areas** of the UK



CAMBRIDGESHIRE



ESSEX



BIRMINGHAM

...but, individual data stays **local** & researchers can only get **summaries** from each database



How many girls aged 12-17y have been diagnosed with anxiety in these 3 locations?



This is important because:

1. We can understand when patterns might be different in different areas of the UK
2. Some symptoms or diagnoses are **rare** so we need to "find" more people to be able to train the algorithms to help them

What do you think about researchers being able to see data from all 3 locations?

DISCUSS!





What do you
think we should
be doing to keep
data safe?

DISCUSS!



There is **no identifying information** (for example, name and date of birth) in the data. Researchers only being able to access **relevant information**.

For example:

For a project about young people aged 13 - 15 with anxiety, researchers wouldn't be able to see the data of 17 year olds with bipolar disorder.



Researchers have to **explain their project** and how the data will help their project.

For example:

Only allowing researchers working on projects which will actually help young people in some way.

We'll come back to this one and give you some more examples!



Researchers have to **get permission** to use the data.

For example:

- Going on a training course on how to use the database
- Asking researchers to sign a document, which tells them how they are and are not allowed to use the data



The data is physically and electronically **locked away**.

For example:

- Researchers have to be given unique login information so that what they do in the database can be tracked
- The computers that the data is stored on are locked away in a secure building.



Researchers can only download **permitted information**.

For example:

The researcher puts the information they want to download in a file. The data protection expert reviews what the project needs and what is in the file, to make sure it is OK.



Do you have any questions or comments?

Does anything worry you?

DISCUSS!



Take a break

Zoom poll!



Deciding which
projects are
'safe projects'



We talked about using the database to help young people with mental health problems...

...but linking data from the NHS (GPs, hospitals, A&E etc.) could be useful for projects looking to improve **physical health** as well.

If you were a physical health researcher what might you want to find out?



People from the NHS?

For example, a your local NHS Trust wants to know how to make waiting times shorter.



People from local councils?

For example, a social worker wants to know what kind of issues young people in their area are having.



People from charities?

For example, the charity Cancer Research wants to understand how inequalities can affect access to cancer treatments.



People from universities?

For example, a researcher wants to know how we can better help parents going through divorce to support their children.



People from private care homes?

For example, a children's care home wants to know how to better support the mental health needs of young people in their care.



AN IMAGINARY START-UP

Technology start-ups?

For example, a start-up company wants to use the database to create a new mental health support app for young people.



Big medical companies?

For example, AstraZeneca want to use the data to develop and sell a better drug for depression.



Big technology companies?

For example, Google wants to create an "early detection" system to help young people with diabetes preserve their eyesight.

Should these projects have access to your de-identified data?

Definitely!

I'm not sure...

No way!



CANCER
RESEARCH
UK

AstraZeneca 



ParkviewCare
Changing lives. Shaping futures.

Google

NHS

Google



maybe not
as helpful
as the
others?



AN IMAGINARY START-UP



What are the **pros** and **cons** of for-profit corporations accessing the data?

Their project might benefit the public e.g. Google helping with diabetic eye disease

If for-profit organisation pay to access the data it can help us to pay for costs which keep data safe

They might not be bound by the same values as health care providers or charities

It will cost money to run the database. If for-profit corporations pay to access the data we can provide access to the NHS and charities for less.

If for-profit organisations pay to access the data it can help us to pay for groups like this

They might make money from accessing the data

There could be unintended negative outcomes

What can we do to make it **safer** for for-profit corporations to access the data?

Only allow for-profit organisations to use the data if they are using it for a project for social good

Ensure that only 'safe people' for not-for profit organisations can access the de-identified relevant information.

Ensure that for-profit organisations are only allowed to access de-identified information that is relevant to a specific project which has been approved by a committee

Only allow for-profit corporations to access the data if all payments are used to run the database and PPI

Who should decide which projects are safe?

The committee

Therapists/doctors

Mental health researchers

school leaders

data security experts
(experts in data technology)

The founder
(Anna)

Organisations who are providing the data e.g. mental health trust or local council

young people with lived-experience of mental health

CAMHS representation

Ethics experts

Parents with children with lived experience of mental health

Information governance experts
(experts in data law)

Not on the committee:

- commercial organisations
- politicians
- journalists/news agencies

Any final thoughts?

About linking data together?

About the Timely project?

About keeping data safe?

Do you have any concerns?



Today we...

- 1 Recapped what the timely project is about
- 2 Understood your views on keeping data safe
- 3 Understood your views on what kind of people, organisations and projects should be able to use the data.
- 4 Understood your views on who should make decisions about the Timely project data.

How do you feel after today's workshop?

Vote!

1



2



3



4



5





Next
Steps

Workshop 3 (21st June)

Exploring ways in which we can clearly explain the project and our plan for keeping data safe to other people.

Focus group

In July we'll get a new group of people who know nothing about the project and show them what we've been working on. We'll do a survey to see what they think and we'll analyse the results.

Future work

We hope to continue working on this project with you beyond July, although we don't know exactly what that will look like yet!



Timely
Workshop 3

Today's aims

- 1 Recap - what is the timely project about? How are we keeping data safe?
- 2 Understand what information we need to tell the public about the database.
- 3 Hear your thoughts on how when and where we should share this information.

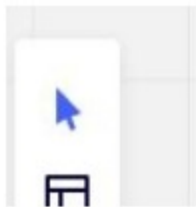
Three things to remember

1. **There are no wrong answers**
2. **You can share your opinions however you want:** out loud, in the chat to the group or using post it notes on this board
3. **There are no silly questions** so please feel free to ask questions any time, you can also send them in the chat

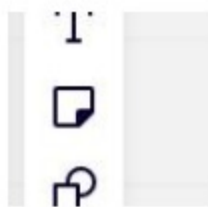
Using Miro

We'll be using Miro today, but if you have any problems we are to help!

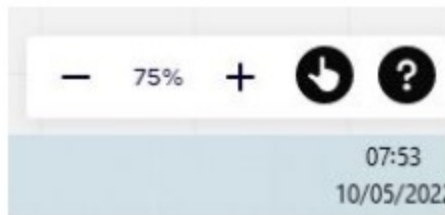
You'll only need to use 3 buttons:



Mouse
Top of the tool
bar



Post-it
4th down on the
tool bar



Zoom
Bottom right of the board

Where do you stand?

Coffee

Tea

Morning person

Night owl

Cats

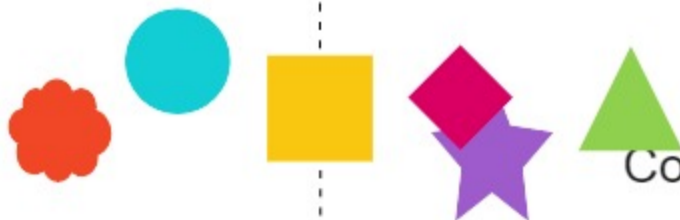
Dogs

Beach

Countryside

Scrunch

Fold





Recap

There are 2 parts to the project, but today we want to focus on just building the **database**

1

Building a linked database of de-identified data. This will enable researchers from lots of different projects to do research to help young people's physical and mental health

2

our project is just one example of how the database could benefit young people

Our project (Timely) will use the database to create computer algorithms to spot patterns in the data to tell apart young people who are likely to be struggling with their mental

What is the problem we are trying to solve?

There are many services in the UK which **routinely** collect data about young people. However this data is **stored separately**, making it very difficult for researchers to use this information in a helpful way.

For example, consider people with chronic health conditions. They might often see their GP, attend specialist services and will miss school due to their health.

It's difficult for researchers to **see the "full picture"** of what is going on with that person.



Your GP



A&E



Social Services



CAMHS



Hospitals



Schools

What can we do?

In order to help researchers understand the physical and mental health issues affecting young people (and find solutions!), we need to **link** all this data in one place.

By doing this in several **locations** we can also get results which are more helpful and diverse.



How will we keep data safe?

We will be sticking to the five safes

Safe data: There is no identifying information (for example, name and date of birth) in the data. Researchers can only access relevant information.

Safe projects: Researchers have to explain their project and how it will help the public.

Safe people: Researchers have to get permission from a committee to use the data.

Safe settings: The data is physically and electronically locked away and all activity in the database is recorded.

Safe outputs: Researchers can only download permitted information.





Any questions?

Communicating
with the public
about building a
database



Why? Transparency is an important part of ethical research

Lets people know what we'll be doing and why

Gives people the option to get more information

Gives people the opportunity to ask questions or voice concerns

Gives people the opportunity to opt out

What is it important to tell people about the **purpose** of the database?



What is it important to tell people about the **benefits** of the database?

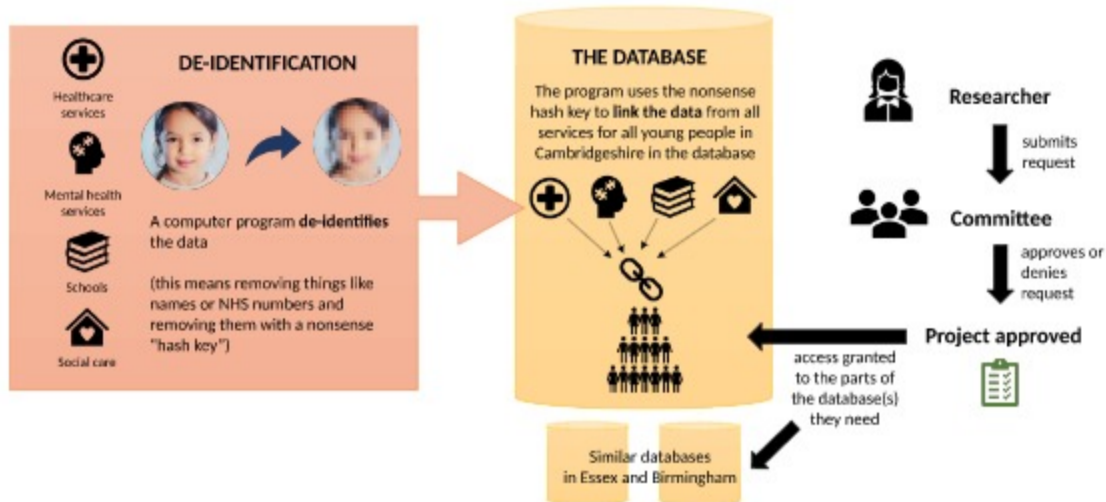


What might people **worry** about that we should clarify?



What **details** do we need to explain?

What happens inside the database?





Grab a post it! You have **5 mins** to write a Tweet about the database



Take a break

Who should we be telling
about the database?



How should we be telling them about the database?

Parents/ guardians

Young people

Professionals

Opting out

What do people need to know about opting out?

How can we let people opt out easily?

What is an **under-served group**?



The voices of some groups in society are less heard than others, particularly people from minority ethnic, religious or gender backgrounds. This is also true for research and can mean that findings are not relevant to all, or at worse are unhelpful or harmful.

It is essential that we are communicating about the project with people from under-served groups.

How do you think we should reach **under-served groups**?

Today we...

- 1 Recapped what the timely project is about.
- 2 Understood what information we need to tell the public about the database.
- 3 Heard your thoughts on how when and where we should share this information.

How are you feeling after today's session?

1



Relaxed

2



Nervous

3



Excited

4



Confused



Next
Steps

Focus group

In July we'll get a new group of people who know nothing about the project and show them what we've been working on. We'll do a survey to see what they think and we'll analyse the results.

Community of interest

We are working to keep this group active in the long term - to support the database, the Timely project, and other similar projects about young people's health and well-being. We will be in touch with more details as soon as we can!



Timely

Workshop 4

Today's aims

- 1 Introductions to the project and each other
- 2 Understanding how easy it is to find information in the communications tools.
- 3 Get your feedback on how we can improve the communications tools

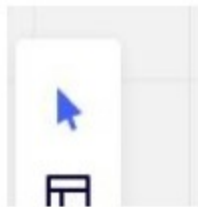
Three things to remember

1. **There are no wrong answers**
2. **You can share your opinions however you want:** out loud, in the chat to the group or using post it notes on this board
3. **There are no silly questions** so please feel free to ask questions any time, you can also send them in the chat

Using Miro

We'll be using Miro today, but if you have any problems we are to help!

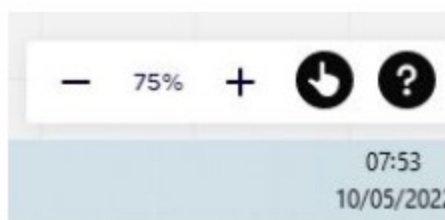
You'll only need to use 3 buttons:



Mouse
Top of the tool
bar



Post-it
4th down on the
tool bar



Zoom
Bottom right of the board

Icebreaker

If you could only eat one type of food for the rest of your life, what would it be?



When I pick your post-it note, tell us:

1. Your name
2. Your pronouns
3. What town or city you live in
4. Why you picked that food

What is the Community of Interest?



Public and Patient Involvement means working in partnership with patients and the public to plan, manage, design and carry out research.
It's really important!

So we have created a network of parents and young people (including you!) who can have a voice on a range of different research projects aiming to support young people's physical and mental health.

What is the project we're talking about today?



CamCHILD database

Linking together **routinely** collected information in a **de-identified** database to support **research**

Young people aged 0 to 17y



Any Questions?

Try and find the following information in the leaflet



1. Why is the Cam-CHILD database being created?

2. How will a researcher get access to the database?

3. Will it be possible to identify the people whose data is in the database?

You have 5 minutes!



**Discussion!**

How easy/difficult was it to find the information?
What could we do to make things easier?



Try and find the following information in the booklet

1. How do we de-identify the information in the database?


2. What kind of data will be included in the database?

3. What is the Data Access Committee?




You have 5 minutes!



**Discussion!**

How easy/difficult was it to find the information?
What could we do to make things easier?



Discussion!

Is there anything in the information that worries you?





Take a break

Discussion!

Is there anything in the information that confuses you?



Discussion!

What could we do to make things clearer?



Why aren't we asking for consent?

Not required under law -

GDPR says that data can be used without consent if it is for the benefit of public health

Consent has to be informed - this would mean everyone reading lots of information about the project

Excludes hard-to-reach groups - which makes research less effective at helping these groups

People can opt out - we will be communicating about what we're doing to ensure that people are aware and can opt out

What do you think of the language?

Works well



Needs improvement



What do you think of the images?

Works well



Needs improvement



What do you think of the layout?

Works well



**Needs
improvement**



Can you think of a time you've heard or seen information about health or research?



**What made you
notice it?**

**What made it
memorable?**

What images would make you think of a database for young people's mental and physical health?

You have **5 minutes!**

Activity!

1. Open **Google Images**
2. Right click on your chosen image and click **copy image**
3. Use **Ctrl V** to paste them below



Is there anything
we've missed?

Today we...

- 1 Introduced you to the project and each other
- 2 Understood how easy it is to find information in the communications tools.
- 3 Got your feedback on how we can improve the communications tools

How are you feeling after today's session?

1



Relaxed

2



Nervous

3



Excited

4



Confused



Next
Steps

Community of interest

We are working to keep this group active in the long term - to work on a range of projects about young people's health and well-being. We will be in touch with more details as soon as we can!

Timely
Workshop 4



Today's aims

- 1 Introductions to the project and each other
- 2 Understanding how easy it is to find information in the communications tools.
- 3 Get your feedback on how we can improve the communications tools

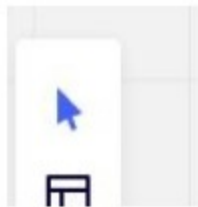
Three things to remember

1. **There are no wrong answers**
2. **You can share your opinions however you want:** out loud, in the chat to the group or using post it notes on this board
3. **There are no silly questions** so please feel free to ask questions any time, you can also send them in the chat

Using Miro

We'll be using Miro today, but if you have any problems we are to help!

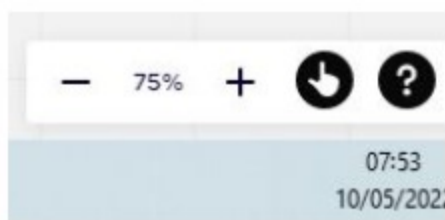
You'll only need to use 3 buttons:



Mouse
Top of the tool
bar



Post-it
4th down on the
tool bar



Zoom
Bottom right of the board

Icebreaker

If you could only eat one type of food for the rest of your life, what would it be?



When I pick your post-it note, tell us:

1. Your name
2. Your pronouns
3. What town or city you live in
4. Why you picked that food

What is the Community of Interest?



Public and Patient Involvement means working in partnership with patients and the public to plan, manage, design and carry out research.
It's really important!

So we have created a network of parents and young people (including you!) who can have a voice on a range of different research projects aiming to support young people's physical and mental health.

What is the project we're talking about today?



CamCHILD database

Linking together **routinely** collected information in a **de-identified** database to support **research**

Young people aged 0 to 17y



Any Questions?

Try and find the following information in the leaflet



1. Why is the Cam-CHILD database being created?

2. How will a researcher get access to the database?

3. Will it be possible to identify the people whose data is in the database?

You have 5 minutes!



Discussion!

How easy/difficult was it to find the information?
What could we do to make things easier?



Try and find the following information in the booklet

1. How do we de-identify the information in the database?


2. What kind of data will be included in the database?

3. What is the Data Access Committee?




You have 5 minutes!



**Discussion!**

How easy/difficult was it to find the information?
What could we do to make things easier?



Discussion!

Is there anything in the information that worries you?





Take a break

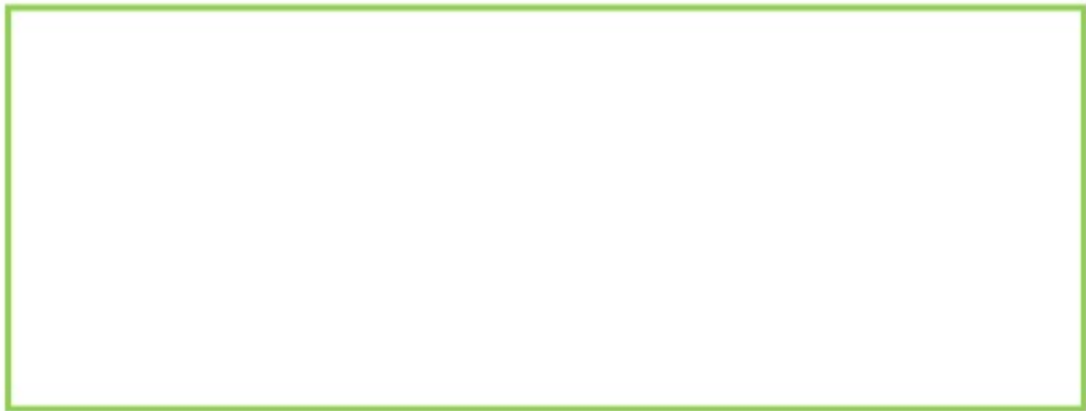
Discussion!

Is there anything in the information that confuses you?



Discussion!

What could we do to make things clearer?



Why aren't we asking for consent?

Not required under law -

GDPR says that data can be used without consent if it is for the benefit of public health

Consent has to be informed - this would mean everyone reading lots of information about the project

Excludes hard-to-reach groups - which makes research less effective at helping these groups

People can opt out - we will be communicating about what we're doing to ensure that people are aware and can opt out

What do you think of the language?

Works well



**Needs
improvement**



What do you think of the images?

Works well



Needs improvement



What do you think of the layout?

Works well



**Needs
improvement**



Can you think of a time you've heard or seen information about health or research?



**What made you
notice it?**

**What made it
memorable?**

What images would make you think of a database for young people's mental and physical health?

You have **5 minutes!**

Activity!

1. Open **Google Images**
2. Right click on your chosen image and click **copy image**
3. Use **Ctrl V** to paste them below



Is there anything
we've missed?

Today we...

- 1 Introduced you to the project and each other
- 2 Understood how easy it is to find information in the communications tools.
- 3 Got your feedback on how we can improve the communications tools

How are you feeling after today's session?

1



Relaxed

2



Nervous

3



Excited

4



Confused



Next
Steps

Community of interest

We are working to keep this group active in the long term - to work on a range of projects about young people's health and well-being. We will be in touch with more details as soon as we can!