DARE UK (Data and Analytics Research Environments UK)

DARE UK Working Group (WG) Charter Template

Name of Proposed WG: *Synthetic Data Community Group*

Synthetic Data
Community Group

Dementias Platform UK
DPUK/SeRP Synthetic Data Working Group
ADR UK
ADR UK Synthetic Data Working Group
Research Data Scotland
Scottish Synthetic Data Working Group
UK Data Service
Data providers' perspectives of low-fidelity synthetic data Project
Cardiff University
Public attitudes to the use of low-fidelity synthetic data Project
Swansea University Prifysgol Abertawe
University of Cambridge
University of Oxford
SeRP
UCL

The WHY

Introduction:

Trusted Research Environments (TREs) play a pivotal role in enabling access to personal data for research, while protecting privacy through strict governance processes. Whilst these strictly governed environments have paved the way for research, the very nature of the governance can result in data being shut behind doors which presents new emerging issues. These challenges include researcher training where access to realistic data is often a challenge due to lengthy timelines of approval and access for many TREs. Additionally, data discovery is limited by just being able to view metadata which means on the surface it may look like there is enough data to carry out analysis, but once access is granted, key variables are missing, or don't have enough data, rendering the project infeasible. Some federated data analysis solutions rely purely on metadata as the real data cannot be seen. This creates challenges for developing algorithms as it is hard to know what the data actually looks like. Finally, the development and release of AI models on pseudonymised data carry a risk of data disclosure outside of the TRE. This has resulted in many data providers refusing the release of AI models which has created a barrier to AI research.

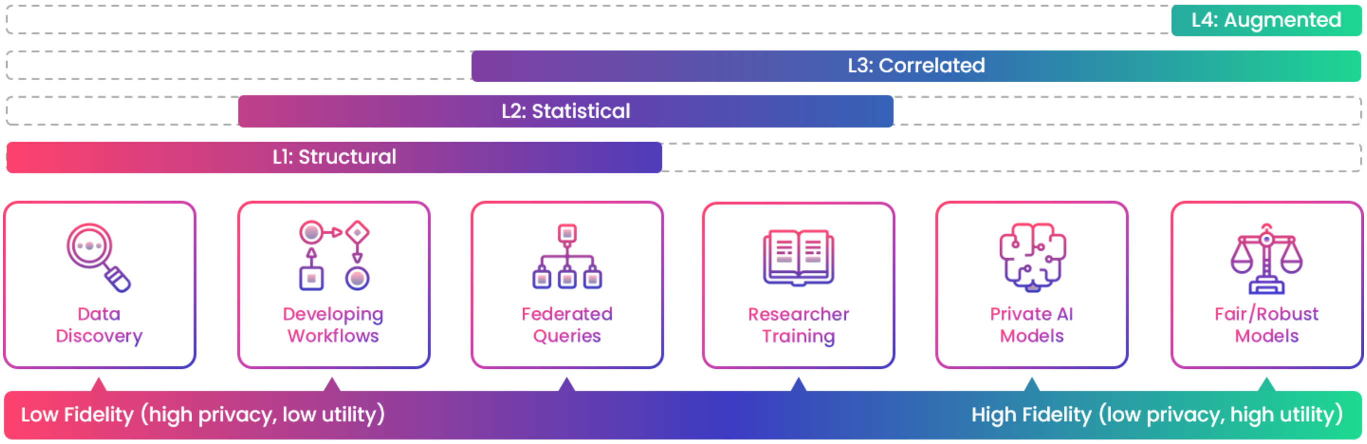


Figure 1 (Levels of synthetic data and example use cases)

A solution to many of these challenges is the development of synthetic data. Synthetic data can be defined as data generated by statistical or machine learning methods that resemble real data to varying levels (Figure 1). Synthetic data has gained mainstream attention from academia and industry in recent years because of its potential to democratise access to personal data for researchers, whilst protecting privacy.

Allowing researchers access to synthetic data with more straightforward and quicker governance processes, researchers and TREs would be empowered to be able to:

- Discover the data at project conception to ensure that the real data will have sufficient coverage to ensure the project is feasible.
- Develop and test algorithms on the synthetic data versions to ensure that they will run efficiently against the real data on access approval. This is also of benefit to some federated solutions where the real data cannot be seen.
- Train the next generation of data experts using synthetic data of varying fidelity to ensure they learn a variety of analysis techniques without the lengthy approval process for access to real data, with reduced administrative burden on TRE staff.
- Train AI models on synthetic data to ensure privacy is protected.

However, whilst there has generally been an optimistic view towards synthetic data expressed by the Information Commissioner’s Office, Food and Drug Administration and the EU AI Act, the biggest limitation to its use is the lack of consensus standards to evaluate privacy. Though there have been major advancements in recent years proving that synthetic data can provide utility for researchers, there remains a dearth of governance frameworks for its use, especially in higher levels of fidelity. Most importantly, there is no uniformly accepted method of assessing whether synthetic data is private.

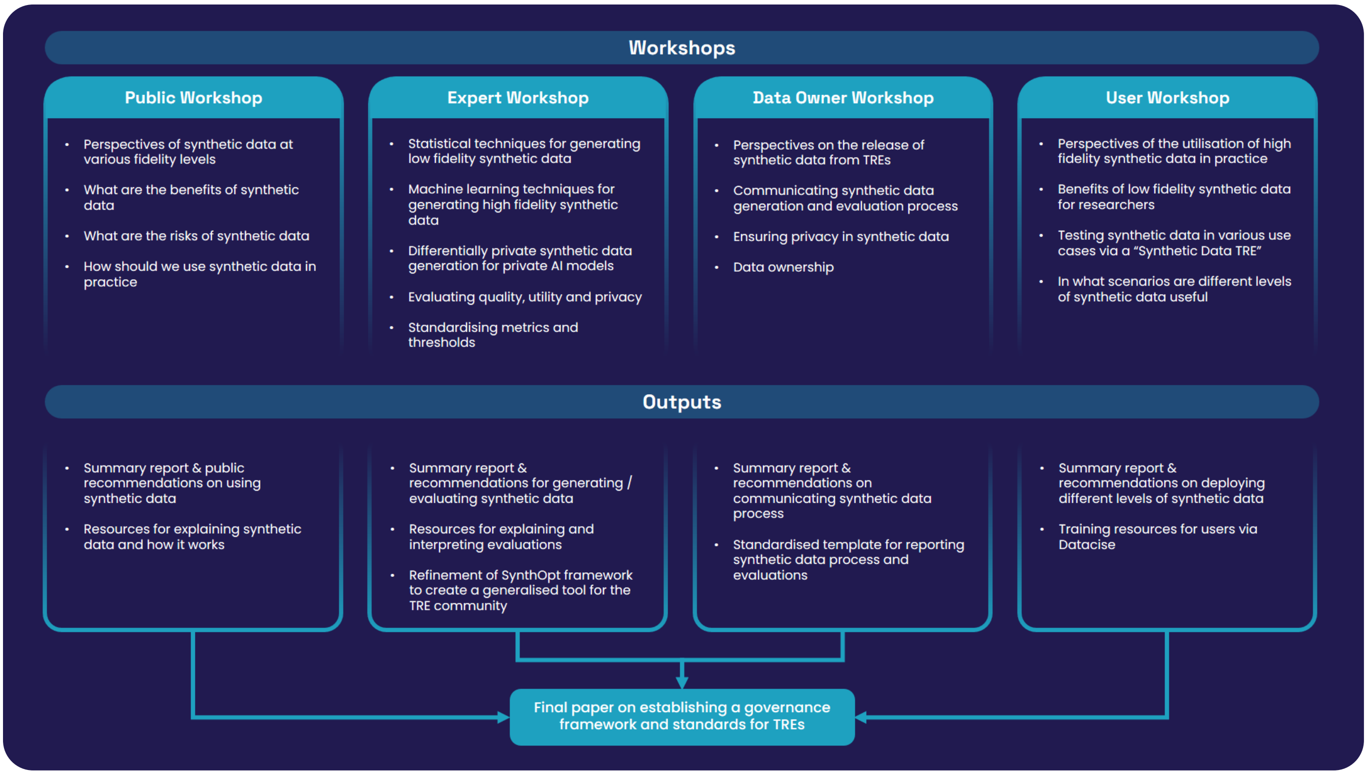
The Synthetic Data Community Group will address the critical need for robust, standardised approaches to generating, evaluating, and deploying synthetic data within TREs. Despite the potential benefits of synthetic data, ranging from low-fidelity for facilitating algorithm development to high-fidelity enabling privacy-preserving machine learning, there are currently significant gaps in standards, governance, and best practices. These gaps create uncertainty and inconsistency across the community, hindering the effective and responsible use of synthetic data.

Currently, there is no clear framework that standardises how synthetic data should be generated or evaluated for privacy and utility, which leads to fragmented solutions that may not meet the diverse needs of data users, owners, and regulatory bodies. The WG will explore the diverse use cases of synthetic data including data discovery, researcher training, running federated analysis and training private AI models to develop a comprehensive framework to enable the deployment of synthetic data for various scenarios. This will be achieved by building upon existing work in this area and hosting workshops to develop standards and governance recommendations for the TRE community.

This WG is aligned with the DARE UK mission to promote safe, secure, and efficient use of sensitive data in research, and will work towards establishing recommendations and tools that ensure robust privacy protection, utility, and deployment of synthetic data across different use cases. Through a series of workshops engaging diverse stakeholders - including researchers, data owners, experts, and the public, the WG will draw on current studies and projects from those within its network and beyond. It aims to create recommendations for standards, governance frameworks, and open-source tools that will support the broader adoption and integration of synthetic data within TREs. The WG will also provide support to TREs implementing synthetic data to ensure that the governance processes are clearly defined and communicated to stakeholders.

The WHAT

Following on from the successful format of the AI Risk Evaluation Group, we will host a series of workshops with different stakeholders to gather perspectives and recommendations on the use of synthetic data for TREs. Each workshop will generate specific resources, tools, and reports that will address the needs and concerns of different stakeholders and contribute to the overall goal of establishing robust governance and standard practices for synthetic data in TREs. Each workshop will aim to recruit 30 attendees. The key workshops and their intended outputs are outlined below:



Workshops

Public Workshop

- Perspectives of synthetic data at various fidelity levels
- What are the benefits of synthetic data
- What are the risks of synthetic data
- How should we use synthetic data in practice

Expert Workshop

- Statistical techniques for generating low fidelity synthetic data
- Machine learning techniques for generating high fidelity synthetic data
- Differentially private synthetic data generation for private AI models
- Evaluating quality, utility and privacy
- Standardising metrics and thresholds

Data Owner Workshop

- Perspectives on the release of synthetic data from TREs
- Communicating synthetic data generation and evaluation process
- Ensuring privacy in synthetic data
- Data ownership

User Workshop

- Perspectives of the utilisation of high fidelity synthetic data in practice
- Benefits of low fidelity synthetic data for researchers
- Testing synthetic data in various use cases via a "Synthetic Data TRE"
- In what scenarios are different levels of synthetic data useful

Outputs

- Summary report & public recommendations on using synthetic data
- Resources for explaining synthetic data and how it works

- Summary report & recommendations for generating / evaluating synthetic data
- Resources for explaining and interpreting evaluations
- Refinement of SynthOpt framework to create a generalised tool for the TRE community

- Summary report & recommendations on communicating synthetic data process
- Standardised template for reporting synthetic data process and evaluations

- Summary report & recommendations on deploying different levels of synthetic data
- Training resources for users via Datacise

Final paper on establishing a governance framework and standards for TREs

Figure 2 (Workshops and planned outputs)

Public/Patient Workshop: Public Perceptions, Benefits, and Concerns of Synthetic Data

This workshop will engage members of the public to explore their perceptions, benefits, and concerns surrounding the use of synthetic data generated from sensitive datasets in TREs. The aim is to understand how the public views the concept of synthetic data at various fidelity levels. Due to existing work on the low fidelity synthetic public workshops carried out by Fiona Lugg-Widger (co-chair), we will build upon this work and focus more on high fidelity use cases. We will gather perceived risks and advantages, their expectations for access, and requirements for privacy protection of synthetic data.

Outputs:

- A summary report detailing public perspectives and recommendations on the use and deployment of synthetic data which protects their privacy.
- Resources that explain the concept, benefits, and safeguards of synthetic data, developed for a non-technical audience. These resources will be made publicly accessible to foster better understanding and transparency around the use of synthetic data.

Expert Workshop: Methods for Generating & Evaluating Synthetic Data

This workshop will bring together experts in synthetic data, statistics, machine learning, and privacy to discuss and refine methodologies for the generation and evaluation of synthetic data. It will focus on addressing technical challenges related to different fidelity levels of synthetic data and the assessment of privacy/utility trade-offs.

Outputs:

- Refinement and development of the Python package "SynthOpt", an open-source tool for the TRE community. SynthOpt will be enhanced to include comprehensive capabilities for generating and evaluating synthetic data, with built-in metrics for assessing both privacy protection and data utility.
- A report summarising expert insights on best practices, methodologies, and technical recommendations for synthetic data generation and evaluation. This will also include standards for evaluating aspects such as privacy in synthetic data.

Researcher/User Workshop: The Implementation and Use Cases of Synthetic Data in Practice

This workshop will focus on the practical implementation of synthetic data across various applications. Researchers will be invited to discuss the challenges and benefits of the use of synthetic data in various scenarios and whether it meets their needs. As part of this, they will be given access to synthetic data within a TRE, to explore the use of synthetic data in practice.

Outputs:

- A summary report that captures researcher perspectives, use cases, and recommendations on the effective use of synthetic data in various scenarios.
- Training and learning materials that will be developed and integrated into the learning platform Datacise. These materials will cover key topics such as how to use synthetic data, interpret synthetic data outputs, and integrate synthetic data solutions into existing research workflows.

Data Owner Workshop: Deployment and Release of Synthetic Data Collections

This workshop will bring together data owners to discuss practical considerations in releasing synthetic data collections. It will address communicating privacy evaluations, managing data ownership, and making informed deployment decisions, building upon work by the UK Data Service on low-fidelity data.

Outputs:

- A report summarising data owners' perspectives and recommendations on deploying synthetic data collections while ensuring their data's privacy is protected.
- Standardised reporting templates for demonstrating evaluations and communicating the quality and privacy safeguards of synthetic datasets helping data owners clearly understand their generation and privacy.
- Incorporation of feedback into the "SynthOpt" tool to create automated interpretable PDF reports that communicate data quality and privacy for informed decision-making.

Final Report & Governance Framework:

At the end of all workshops, the working group will bring together the perspectives, recommendations, and outputs from each stakeholder group into a comprehensive final report. This report will outline a governance framework for the responsible deployment of synthetic data from TREs, providing clear guidelines on generation, evaluation, privacy protection, and communication practices.

Outputs:

- A paper that brings together the perspectives and recommendations of all stakeholder groups, proposing a governance framework for synthetic data in TREs and building on previous work in this area. This paper will be shared with the wider community to encourage the adoption of standardised practices.
- Publication of all workshop summary reports, resources, and open-source tools developed through the working group, ensuring that these outputs are accessible and reusable by the broader TRE and research community.

By delivering these outputs, the Synthetic Data Working Group aims to address existing gaps in the field, promote best practices, and establish a foundation for the safe, effective, and transparent use of synthetic data across various use cases.

The WHO

The proposed Synthetic Data Working Group brings together a diverse team of experts with proven experience in the area of synthetic data. Each member of this group has a unique background that aligns with the key challenges this WG aims to address, ensuring a comprehensive approach to the development of standards, tools, and governance frameworks for synthetic data in TREs. The multidisciplinary team of Co-Chairs involves data owners, academics, clinicians, TRE managers and public representation.

Simon Thompson (co-director of SeRP UK, SAIL Databank and Dementias Platform UK) and **Lewis Hotchkiss** (Research Officer at Dementias Platform UK) both led the AI Risk Evaluation Group which was previously funded by the DARE UK community group initiative. Following on from this group, synthetic data was identified as an important area of exploration which is why we established a Synthetic Data Working Group to address the governance challenges of synthetic data in TREs. We have already built up a community of researchers and data owners who are interested in developing governance in this area. Additionally, we have been actively working on addressing the practical challenges of synthetic data through the development of a synthetic data generation/evaluation framework called "SynthOpt". **Anmol Arora** (University College London) is an Academic Clinical Fellow and has also actively worked in the practical development of synthetic data, as well as governance as an advisor for the Utah AI Bill, which was the first legislation to specifically define and provision for synthetic data.

The working group is further strengthened by co-chairs from other existing synthetic data communities. Bringing together existing synthetic data groups is crucial for developing best practices and governance models, building on existing work and ensuring consensus.

Emily Oliver from Administrative Data Research UK (ADR UK), leads an existing synthetic data working group which is currently undertaking exploratory work to understand the perspectives of different stakeholders. Furthermore, **Sophie McCall** from Research Data Scotland leads the Scottish Synthetic Data Working Group and **Cristina Magder** (UK Data Service), leads a project looking at the data owner's perspective of benefits, costs and utility of low-fidelity synthetic data in TREs. The proposed community group will be able to expand on the findings from these current groups and incorporate findings to support the evolution of a governance framework for synthetic data.

Additionally, both **Fiona Lugg-Widger** and **Robert Trubey** from the Centre for Trials Research at Cardiff University bring expertise in public engagement, having co-led a project that gathered public perceptions on low fidelity synthetic data. Their previous work in organising workshops and capturing public feedback ensures that the WG will effectively engage the public and incorporate their views into the governance framework. **Steve Moore** represents the public and attended all of the public consultation workshops run by Cardiff University and contributed to the final recommendations for low fidelity synthetic data.

Additionally, **Timothy Rittman** (University of Cambridge) and **John Gallacher** (University of Oxford, Director of DPUK) bring the perspectives of data owners which will be important to ensuring the effective implementation of synthetic data in practice. **Emma Squires** brings the perspective of a TRE manager and will be important in ensuring that the governance, processes and practicalities of implementing synthetic data access into a mature TRE are surfaced and addressed by the WG.

With several members who have previously established synthetic data communities, the WG is well-positioned to consolidate existing networks and foster collaboration. This synergy will help drive forward robust and unified solutions, leveraging the expertise of each member to produce practical outputs that address the challenges facing the broader TRE community.

The HOW and WHEN

The Synthetic Data Community Group will actively engage with several related initiatives and groups to ensure a collaborative and integrated approach to developing standards and best practices. This WG will collaborate with existing synthetic data groups, including the DPUK Synthetic Data Working Group, the ADR UK Synthetic Data Working Group, the RDS Scottish Synthetic Data Working Group, the CTR Public Perspectives of Synthetic Data Group and the UK Data Service group.

Beyond these groups, the WG will coordinate with the International Synthetic Data Working Group established from the International Population Data Linkage Network (IPDLN), which includes members from Canada, the United States, and Australia. This collaboration with an international community will help to align standards and practices across different countries, facilitating the broader applicability and adoption of the outputs developed by the WG.

To ensure consistent progress, the co-chairs of the WG will meet monthly to review ongoing work, address any challenges, and plan upcoming activities. These regular meetings will help maintain momentum and coordination, particularly between the organisation of workshops and the development of outputs. The WG will hold workshops every 2 months, each dedicated to a specific stakeholder group (public/patients, experts, researchers/users, data providers). These workshops will be structured to achieve clear objectives, and the outcomes will feed into the next stage of the WG's activities. Between workshops, smaller focus groups and task teams will work on refining tools, drafting reports, and developing materials. Regular communication through virtual meetings, slack, and shared documents will ensure that members remain engaged and aligned on deliverables.

Potential members: [Including a minimum of two proposed chairs and all members who have expressed interest]

FIRST NAME	LAST NAME	EMAIL	(Co-)Chair / Member
Lewis	Hotchkiss	l.f.hotchkiss@swansea.ac.uk	Co-chair
Simon	Thompson	simon@chi.swan.ac.uk	Co-chair
Emma	Squires	emma@chi.swan.ac.uk	Co-chair
Emily	Oliver	emily.oliver@esrc.ukri.org	Co-chair
Sophie	McCall	sophie.mccall@researchdata.scot	Co-chair
Cristina	Magder	dcmagd@essex.ac.uk	Co-chair
Timothy	Rittman	tr332@medschl.cam.ac.uk	Co-chair
John	Gallacher	john.gallacher@psych.ox.ac.uk	Co-chair
Anmol	Arora	aa957@cam.ac.uk	Co-chair
Fiona	Lugg-Widger	luggfv@cardiff.ac.uk	Co-chair
Robert	Trubey	trubeyrj@cardiff.ac.uk	Co-chair
Steve	Moore	irishsteve38@hotmail.com	Co-chair
Note	We will recruit members from existing groups established by the co-chairs.		

The DPUK Synthetic Data Working Group contains 38 members, the ADR UK Synthetic Data Working Group contains around 42 members and the DELIMIT study contained around 40 public members. Members from the DPUK WG have already expressed interest and we will recruit members from the other existing groups, as well as recruit new members.