DARE UK (Data and Analytics Research Environments UK)

DARE UK Interest Group (IG) Charter Template

Name of proposed Interest Group: *Synthetic Data Adoption Group*

The WHY

Introduction: [A brief articulation of what broad-based challenge(s) or issue(s) the IG will address, how this IG is aligned with the DARE UK mission, and how this IG would be a value-added contribution to the sensitive data research community]

This Interest Group (IG) will address the challenge of advancing the real-world adoption of synthetic data across sensitive data research settings. Despite growing interest, the application of synthetic data remains hampered by governance uncertainty, low technical familiarity, and limited trust from both practitioners and the public.

This IG builds on the foundations laid by the previously DARE UK-funded UK Synthetic Data Community Group (UKSDCG), transitioning from identifying challenges to delivering practical solutions. It aligns with the DARE UK mission by promoting trustworthy, privacy-enhancing technologies that support a more coordinated, federated and secure data research infrastructure across the UK. The IG will serve as a national platform to support collaboration, align governance, embed public involvement, and produce practical outputs that directly improve the use and acceptability of synthetic data.

This IG adds value by convening stakeholders across sectors: TREs, regulators, researchers, data owners, the public, to co-develop interoperable approaches and best practices for synthetic data. It addresses urgent needs around trust, capacity, and operational alignment.

User scenario(s) or use case(s) the IG wishes to address: [what triggered the desire for this IG in the first place]

There is a growing need for synthetic data to improve the research experience across the UK's sensitive data research ecosystem. Synthetic data can provide earlier, easier access to datasets for researchers during the often-lengthy approval and data access process. By enabling pre-analysis data discovery, code development, and research planning, synthetic data significantly reduces friction in the research journey, allowing researchers to work efficiently and within safe governance boundaries.

Synthetic data also provides low-risk datasets for training and education, enabling new analysts and students to develop skills and familiarity with data environments without access to real-world sensitive data. Furthermore, it facilitates technological innovation, by offering test environments for developing and implementing new technology, such as federated capabilities, while governance and trust is still being established.

Synthetic data can also play a key role as a privacy-preserving method for AI model development, ensuring that models can be safely released from TREs.

However, despite this clear utility, adoption remains low. Through the work of the UKSDCG, we identified several key barriers that are preventing synthetic data from being widely and responsibly implemented:

- **Limited understanding among data owners, governance leads, and the public:** Many stakeholders are unfamiliar with how synthetic data is generated, how it protects privacy, and where it should (and should not) be used. This lack of familiarity undermines trust and wider social acceptability.
- **Lack of expertise:** Many TREs and institutions do not currently have the skills, tools, or confidence to develop, evaluate, or deploy synthetic data.
- **Inconsistent governance and policy frameworks:** There is a lack of shared guidance, standards, and policies across TREs for synthetic data. This inconsistency creates confusion and hinders cross-institutional interoperability and wider trust in synthetic data.
- **Regulatory uncertainty:** While synthetic data has strong privacy-enhancing potential, there is no clear, nationally recognised regulatory position on its use. This lack of clarity creates hesitation among data custodians, especially in risk-averse settings such as health, government, and commercial sectors.

These challenges collectively represent the trigger for the formation of this IG. By addressing these gaps, through consensus-building, shared guidance, capacity building, and public dialogue, this IG will help unlock the value of synthetic data as a key enabler of a trustworthy, innovative, and federated data research ecosystem.

The WHAT

Objectives: [Set of focus areas for discussion linked to the driving user scenario(s)/use case(s) above, articulate how this group is different from other such groups inside or outside of the DARE UK programme, which specific strategic theme within the DARE UK programmes recommendations (see page 19) is this IG primarily aligned with. *]

This IG will advance the safe, trusted, and effective adoption of synthetic data across the UK's sensitive data research ecosystem. Building on the foundations laid by the UKSDCG, this IG will address key barriers and support the development of best practices, practical tools, and community capacity. It aligns with two strategic themes from the DARE UK recommendations: Capability and Capacity and Data and Metadata Standards.

This IG is uniquely positioned as a unifying national forum for synthetic data, bringing together leaders from existing working groups, infrastructure providers, and government-aligned projects across the UK. Many of the IG's co-chairs and members already lead synthetic data initiatives in their own domains: including national working groups in England, Wales and Scotland, and are involved in the development and governance of synthetic data within their respective TREs and public sector organisations.

While these localised and organisation-specific efforts are vital, they often operate independently. This IG fills a critical gap by convening these disparate efforts into a cohesive, UK-wide community. It serves as a cross-sector coordination mechanism to foster alignment, reduce duplication, and build shared guidance, standards, and capacity.

Focus 1 : Refinement & Adoption of Community Best Practices and Guidelines

The responsible and effective use of synthetic data depends on the development and adoption of clear, practical, and widely accepted standards. In our previous work, we identified a lack of consistent approaches to synthetic data generation, evaluation, and reporting as a major barrier to adoption. In response, we developed a preliminary set of governance recommendations and tools. This IG will build on that foundation by increasing uptake, engaging new sectors, and refining those recommendations through broader consultation. Our aim is to establish community-endorsed best practices that are technically robust, publicly acceptable, and aligned with real-world operational needs.

- *Workshop on synthetic data in NHS Secure Data Environments:* To achieve this, we will deliver a series of stakeholder-led activities, including a workshop with the NHS and SDE partners such as NHS England, SAIL Databank, and Public Health Scotland. This workshop will focus on refining draft governance frameworks, sharing lessons learned from implementation, and identifying practical requirements for scaling adoption.
- *Workshop on synthetic data for federated research:* A dedicated workshop will address the specific challenges and opportunities of using synthetic data across distributed TREs. This session will bring together federation-focused initiatives (e.g. DARE UK projects and infrastructure leads) to examine how synthetic data can support safe collaboration across environments, and how consistent governance frameworks can be designed to operate effectively within a federated ecosystem.
- *Delphi study on synthetic data use:* Building on these two workshops and our previous work, a Delphi study will be launched to formally gather consensus on best practices and reporting standards across different use cases and fidelity levels.

These activities will result in the publication of a practical and consensus-driven set of guidelines for the generation, evaluation, and responsible deployment of synthetic data. Public involvement will be integrated throughout, ensuring that emerging guidelines reflect societal values, particularly around transparency, privacy, and accountability. Adoption of these guidelines will lead to increased standardisation and interoperability of practices across TREs and sectors, supporting both research efficiency and public trust. It will also lay the foundation for long-term governance maturity in the UK synthetic data landscape.

Focus 2 : Training & Capacity Building

Despite growing interest in synthetic data, our previous workshops and stakeholder consultations revealed a consistent lack of technical understanding among analysts, data owners, and governance teams, particularly around what synthetic data is, how it is created, and how its privacy and utility are assessed. This knowledge gap limits confident adoption and contributes to fragmented, inconsistent practices. At the same time, public trust and acceptance remain constrained by low awareness and uncertainty about how synthetic data is used safely and ethically. This focus area aims to address both technical and public-facing gaps through targeted training, co-developed educational resources, and practical tools. Our goal is to equip the community with the knowledge and skills needed to adopt synthetic data responsibly and effectively.

- *Workshop on synthetic data in government departments:* A central activity will be a hands-on workshop with government departments and public sector data owners, offering practical experience in generating and evaluating synthetic data using the open-source tools developed through the UK SDCG. This session will be tailored to the relatively structured datasets commonly found in public sector TREs and will enable participants to apply tools to their own scenarios.
- *Resources for the public and data providers:* To build public understanding of synthetic data, we will co-develop online learning resources with members of the public that explain key concepts, safeguards, and benefits in accessible terms. In parallel, we will produce training materials, templates, and guides that data providers and analysts can use to support responsible synthetic data implementation in their own organisations.

These activities will result in increased technical capacity within public sector and research teams, reducing barriers to adoption and encouraging more consistent and confident use of synthetic data tools. The public-facing materials will help foster greater transparency and trust, while enabling informed dialogue about data use. Overall, this work will support broader, safer uptake of synthetic data across the UK.

Focus 3 : Supporting Adoption & Trust in Synthetic Data across Sectors

One of the strongest and most consistent themes from our previous expert and data owner workshops was the lack of regulatory clarity as a key barrier to deploying synthetic data in practice. Without clear regulatory or policy direction, many organisations, despite being interested, remain risk-averse and hesitant to act. This focus area addresses that challenge by building bridges between technical innovation and the governance landscape. Our objective is to foster meaningful engagement with regulators and policymakers, develop clearer expectations and frameworks, and support more confident adoption of synthetic data across sectors, underpinned by public trust and regulatory alignment.

- Privacy Regulations: A dedicated engagement session will be held to explore how synthetic data aligns with UK data protection and privacy regulations. This session will bring together privacy experts, legal experts, and synthetic data practitioners to examine key questions around identifiability, data minimisation, and legal status under UK GDPR. The focus will be on developing a principles-based approach that can flexibly accommodate different types of synthetic data and their associated risks, while maintaining high standards of privacy protection and public trust. Outcomes from this session will inform guidance and recommendations for privacy-compliant deployment of synthetic data in Trusted Research Environments (TREs).
- Medical Device Regulations: A separate session will explore the role of synthetic data in AI-driven medical research, clinical device development, and real-world decision making. This workshop will focus on how synthetic data can be used to support algorithm training, testing, and validation within regulated environments. Participants will include stakeholders from the NHS, academia and industry. This was a key area of concern identified by stakeholders, so by addressing concerns with the use of synthetic data in real world decision making, we hope to improve trust and acceptability in these settings.

Through this work, we aim to achieve clearer, more actionable guidance for the deployment of synthetic data, aligned with existing data protection and governance requirements. Regulators and policymakers will be better informed to support innovation while maintaining public trust. The outcome will be a stronger foundation for safe, compliant, and trusted synthetic data use across diverse public and research settings.

Focus 4 : Demonstrating Value/Impact of Synthetic Data

To drive adoption, synthetic data must demonstrate not only safety but also real-world value. While synthetic data is gaining traction, its adoption remains at an early stage, and many data owners and decision-makers have asked for tangible evidence of its benefit and value before committing to wider implementation. Our previous work revealed a shared desire to "see it in action" to understand how synthetic data improves research efficiency, access workflows, or supports technical development. In this phase, we aim to generate and communicate that evidence. Our objective is to show real-world value through studies, accessible case examples, and partnerships that demonstrate how synthetic data can meaningfully support research while maintaining privacy and security. This was identified as a particular priority for the public.

- *Synthetic data to support data access:* We will conduct a study evaluating how synthetic data can be used to streamline the data access process. This includes supporting researchers to refine hypotheses, prepare analysis code, and test methodologies while waiting for access to real data. In parallel, we will collaborate with SeRP TREs to deploy the open source tools developed from the UKSDCG, enabling the generation and release of high-quality synthetic datasets for use in training and development scenarios. These activities will be complemented by collecting real-world case studies and user stories that illustrate the benefits and challenges of applying synthetic data in practice.

The expected outcomes include a growing evidence base demonstrating the effectiveness of synthetic data in accelerating research and improving data governance workflows. TREs will benefit from wider deployment of SynthOpt, improving capacity to offer synthetic data as part of standard services. Collectively, these efforts will support public and institutional trust by showing not only that synthetic data is safe, but that it is valuable, impactful, and essential for modern, privacy-conscious research.

<u>Outcomes:</u> [What does this IG intend to achieve including outputs that may lead to WG topics, any initial ideas on outputs that could be delivered through WGs spawned under this IG]

Community-endorsed best practice guidelines for the generation, evaluation, and responsible deployment of synthetic data, developed through multi-stakeholder engagement, including public input and a Delphi study.

Training resources and educational materials for public audiences, data providers, researchers, and analysts to increase understanding and confidence in the use of synthetic data.

Open-source tools and practical templates, including example End User Licence Agreements (EULAs), policy templates, code notebooks, and a curated repository of synthetic data generation/evaluation tools.

Real-world case studies and evaluation reports that evidence the impact of synthetic data on research efficiency, data access workflows, and innovation in secure environments.

These outputs will serve as a foundation for creating one or more focused Working Groups (WGs) under the IG. Potential WGs may include:

- A Governance & Standards WG to continue refining guidelines, develop frameworks, and ensure interoperability across TREs.
- A Technical Tools WG focused on developing and maintaining open-source tools for synthetic data generation and evaluation.

The HOW

<u>**Participation/Collaboration:**</u> [Which communities will be involved, and what relevant skills/knowledge/experience should they have, are there other adjacent or relevant IGs/WGs this group will coordinate and/or collaborate with]

The group will be led by co-chairs who already play a central role in the UK's synthetic data landscape. These include leaders of the ADR UK Synthetic Data Working Group, the Scottish Synthetic Data Working Group, and representatives from organisations that have driven governance development and implementation of synthetic data, such as Research Data Scotland, Dementias Platform UK, and the UK Data Service. Public engagement will be a core pillar, drawing on existing work such as the DELIMT study, which has explored public values around synthetic data and data minimisation.

By convening this IG, we aim to unite and build on these existing initiatives to reduce fragmentation, align standards, and share practical experience. The group will ensure that efforts across different regions and sectors are coordinated, mutually reinforcing, and informed by both technical expertise and public values.

In addition, the IG will work closely with other DARE UK-supported projects and groups, including those focused on federated data infrastructure, AI training, and disclosure control. These collaborations will ensure that synthetic data needs and challenges are considered across key areas of emerging infrastructure and policy, and that our outputs support broader programme-wide goals.

Participation will be open to a wide range of stakeholders. We will engage researchers and analysts with technical expertise in synthetic data generation, evaluation, and privacy assurance; governance and policy

professionals working in TREs and data custodianship roles; regulators and legal experts interested in data protection, AI safety, and ethical standards; and public contributors, particularly those with prior experience engaging in data governance conversations.

<u>Mechanism:</u> [Describe how often the group will meet and sustain progress between meetings]

The IG will meet monthly through virtual meetings to maintain momentum, foster collaboration, and provide a regular forum for sharing updates, showcasing ongoing work, and identifying emerging challenges or opportunities across the synthetic data landscape. These monthly meetings will include presentations from group members, updates from related projects, and focused discussions to guide the development of outputs.

To sustain progress between meetings, we will use Slack where members can communicate, share drafts, review resources, and coordinate group activity. Sub-groups may also form around specific deliverables such as guidance documents, public resources, or technical tools, and will be encouraged to meet independently as needed.

In addition to online engagement, we will hold in-person workshops and engagement activities during the funding period. These events will serve as key milestones to test ideas, consolidate outputs, and drive wider adoption through stakeholder-led co-creation.

<u>Potential members:</u> [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
Emma	Squires	emma@chi.swan.ac.uk	Co-chair
Simon	Thompson	simon@chi.swan.ac.uk	Co-chair
Emily	Oliver	emily.oliver@esrc.ukri.org	Co-chair
Sophie	McCall	sophie.mccall@researchdata.scot	Co-chair
Puja	Myles	puja.myles@mhra.gov.uk	Co-chair
Cristina	Magder	dcmagd@essex.ac.uk	Co-chair
Timothy	Rittman	tr332@medschl.cam.ac.uk	Co-chair
Anmol	Arora	aa957@cam.ac.uk	Co-chair
John	Gallacher	john.gallacher@psych.ox.ac.uk	Co-chair
Fiona	Lugg-Widger	luggfv@cardiff.ac.uk	Co-chair
Robert	Trubey	trubeyrj@cardiff.ac.uk	Co-chair
Steve	Harris	steve.harris@ucl.ac.uk	Co-chair
Steve	Moore	irishsteve38@hotmail.com	Co-chair

We will also continue to engage with members of our community built up from the UKSDCG which contains over 200 members.

Note, please do not hesitate to point out gaps in the current DARE UK set of strategic themes and/or recommendations that the programme should consider as it continues to evolve these. Community feedback and input is welcomed.