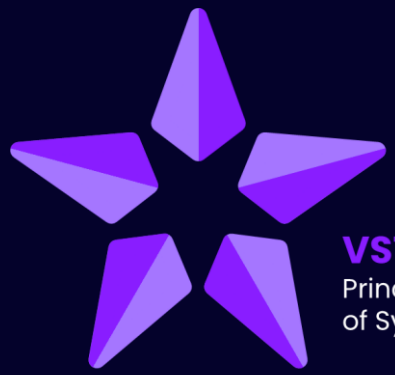


VSTAR Framework
Principles for the Development of Synthetic Data in TREs

Perspectives & Recommendations on the Development of Synthetic Datasets in Trusted Research Environments

Co-Chairs: Lewis Hotchkiss, Emma Squires, Emily Oliver, Sophie McCall, Cristina Magder, Anmol Arora, Fiona Lugg-Widger, Robert Trubey, Steve Moore, Timothy Rittman, John Gallacher, Simon Thompson

Acknowledgements: Lois Owumi, Pawel Darda, Kafayat Adeoye, Alieyeh Sarabandi

UK Synthetic Data Community Group

Funded By

UK Synthetic Data Community Group

This report was developed by the UK Synthetic Data Community Group, a group funded by DARE UK to advance the responsible use of synthetic data in sensitive data research. Through a series of stakeholder workshops bringing together researchers, data custodians, synthetic data experts, and members of the public, we explored the practical requirements, governance frameworks, and best practices needed to enable synthetic data to deliver value while maintaining privacy protections and public trust. This report presents the key findings from these workshops and provides recommendations to guide organisations in the generation, evaluation, governance, and deployment of synthetic data within Trusted Research Environments.

Funded by DARE UK

Table of Contents

This report is structured around the VSTAR framework principles: Valuable, Safe, Transparent, Accessible, and Responsible. Each section explores one of these core principles, presenting stakeholder perspectives and recommendations for ensuring synthetic data meets expectations around utility, privacy protection, appropriate access, and accountable governance. By organising the findings in this way, we provide a comprehensive yet practical guide that addresses the full lifecycle of synthetic data development and deployment, from initial value proposition through to ongoing responsible stewardship.

03 Introduction

The Introduction provides context for the UK Synthetic Data Community Group's work and outlines the purpose and scope of this report.

04 Background

The Background section introduces synthetic data concepts, defines fidelity levels, and explains the challenges and opportunities within Trusted Research Environments.

06 Executive Summary

The Executive Summary presents the VSTAR framework, summarising key findings and core recommendations from across all stakeholder workshops.

10 Valuable

The Valuable section explores the importance of defining intended use cases and requirements to deliver meaningful benefits.

20 Safe

The Safe section addresses disclosure risk assessment, privacy evaluation methods, and the balance between technical safety and reputational risk management.

32 Transparent

The Transparent section emphasises the important role of comprehensive documentation in building trust and enabling informed decision-making.

40 Accessible

The Accessible section examines proportionate access models for different fidelity levels, balancing ease of access with appropriate safeguards against misuse.

46 Representative

The Representative section explores ensuring fairness and quality in synthetic data through appropriate evaluation methods and mitigation strategies.

53 Future Work

The Future Work section outlines next steps of the community group for driving real-world synthetic data adoption and applications.

54 Acknowledgements

Acknowledgements of the people and organisations involved in delivering and supporting this report.

Lewis Hotchkiss
Dementias Platform UK

Emma Squires
Dementias Platform UK, SeRP

Simon Thompson
Dementias Platform UK, SAIL Databank, SeRP

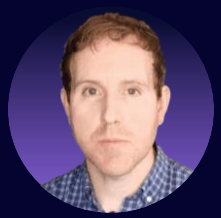
John Gallacher
Dementias Platform UK, University of Oxford

Tim Rittman
University of Cambridge

Emily Oliver
Administrative Data Research UK

Cristina Magder
UK Data Service

Fiona Lugg-Widger
Centre for Trials Research



Robert Trubey
Centre for Trials Research

Sophie McCall
Research Data Scotland

Anmol Arora
University College London

Steve Moore
Public Representative

Introduction to the UK SDCG

The UK Synthetic Data Community Group (UKSDCG) was established with funding from DARE UK to advance the responsible development and use of synthetic data within Trusted Research Environments (TREs). Recognising that synthetic data holds significant promise for accelerating research workflows, enabling training and education, supporting data discovery, innovation, and reducing operational burdens on TREs, while protecting privacy, the UKSDCG was formed to address how this can be achieved in a safe and responsible way that delivers real value.

To achieve this, the UKSDCG brought together diverse stakeholders to build consensus around the key principles and practical guidance needed to enable synthetic data to deliver genuine value while maintaining robust privacy protections and public trust.

Led by 12 co-chairs representing different data organisations, synthetic data initiatives, and areas of expertise from across the UK, the group delivered four comprehensive stakeholder workshops engaging researchers who use TREs, data custodians responsible for governance decisions, synthetic data experts with technical and methodological expertise, and members of the public whose data underpins research. This multi-stakeholder approach ensured that perspectives from across the synthetic data ecosystem were heard, understood, and integrated into the resulting framework and recommendations.

Through structured discussions and surveys across these workshops, we explored stakeholder perspectives on the requirements, concerns, and priorities surrounding tabular synthetic data in TRE contexts.

We examined what makes synthetic data valuable for different use cases, how safety and privacy can be assured, what documentation and transparency are needed to build trust, how access should be managed proportionately, and how quality and fairness can be maintained. From these rich insights, we developed the VSTAR framework, a set of five core principles: Valuable, Safe, Transparent, Accessible, and Representative, that emerged as essential for ensuring synthetic data is fit-for-purpose, trustworthy, and responsibly governed.

This report presents the findings and recommendations emerging from this collaborative effort, structured around the VSTAR framework principles. It provides guidance to support organisations in generating, evaluating, governing, and deploying synthetic data responsibly within the context of TREs and sensitive data research.

Background

The Challenge of Accessing Sensitive Data for Research

The sharing of healthcare and administrative records has created unprecedented opportunities for research that can improve public health, inform policy decisions, and advance scientific understanding. However, these datasets often contain highly sensitive personal information that must be rigorously protected. The fundamental tension in data research is how to enable legitimate, beneficial research while safeguarding individual privacy and maintaining public trust in data stewardship.

Current best practice for managing this tension involves sharing sensitive data within Trusted Research Environments (TREs) - highly secure computing platforms where accredited researchers can access data under strict governance controls. TREs provide essential privacy protections through technical safeguards and adhere to the Five Safes to ensure data is accessed and used in a responsible way. Access to data within TREs is typically governed by formal approval processes involving data access committees and user agreements that establish researcher responsibilities and organisational accountability.

While TREs play a crucial role in enabling trusted data access, they also present a number of barriers. Gaining access to real data within a TRE can involve lengthy approval processes, often taking months to be able to access. This can delay the start of research projects and create significant inefficiencies, particularly for time-sensitive work. Furthermore, training and education are hindered by these restrictions, as students and early-career researchers often struggle to access data for learning and skill development. Another limitation is the reliance on metadata to determine dataset suitability. Researchers are often asked to make decisions about whether a dataset will meet their needs without ever seeing the actual data. This can lead to wasted time and resources if, after going through the access process, the dataset proves unsuitable for their intended analyses.

These limitations could be overcome through the integration of synthetic data into TRE services, to accelerate the research process and broaden access to high-quality training resources, ultimately making TREs more effective.

What is Synthetic Data?

Synthetic data offers a potential approach to reducing these barriers while maintaining privacy protections. Rather than working directly with real patient records, synthetic data uses statistical or computational methods to generate artificial datasets that preserve key characteristics of the source data without containing actual individual records. The premise is that if synthetic data can capture relevant structural or statistical properties while eliminating the direct link to real individuals, it might be shared more openly and accessed more easily than the sensitive source data from which it was derived.

However, synthetic data is not a single, uniform concept. The term encompasses a wide spectrum of approaches that vary dramatically in their generation methods, the properties they preserve, and consequently in their appropriate use cases and governance requirements. Understanding these distinctions is essential for productive discussion about synthetic data's role in research.

Fidelity Levels

To support discussions throughout the workshops, we categorised different levels of synthetic data fidelity - the degree to which synthetic data preserves properties of the source data.

Fidelity exists on a spectrum: at the lowest end, synthetic data might preserve only basic structural properties like variable names and data types; at higher levels, it may preserve statistical distributions, and complex multivariate relationships found in the real data. Higher fidelity enables more sophisticated analytical uses but generally increases both the complexity of generation and the potential for privacy risks, as synthetic records that more faithfully reproduce real data patterns may be more similar to actual individuals. This distinction fundamentally shapes what synthetic data can be used for and what privacy considerations apply.

Fidelity Levels of Synthetic Data

For the purposes of the workshops, we pre-defined four levels of synthetic data fidelity along this spectrum, providing a framework to support discussion around use cases, privacy, and governance needs at each level. These are as follows:

Level 1 : Structural
Typically, this type of synthetic data is generated solely from metadata. The resulting dataset mimics the structural properties of the original data, maintaining variable types, the range of values, and overall schema, but the values themselves are entirely random. No statistical properties or relationships from the real dataset are retained.

Level 2 : Univariate
At this level, synthetic data is often generated by sampling from independent univariate distributions for each variable. This means that each variable's statistical properties, such as its distribution shape, mean, and standard deviation etc, are replicated individually. However, inter-variable relationships are not preserved.

Level 3 : Multivariate
This category includes synthetic datasets that preserve relationships between variables, capturing the joint statistical structure of the original data. These datasets are often produced by more advanced techniques such as generative models which aim to replicate the statistical properties and patterns of the original data.

Level 4 : Augmented
At the highest level, synthetic data is generated not to replace real data, but to enhance or extend it. This includes techniques that create synthetic examples for underrepresented cases in a dataset, such as class balancing in machine learning tasks.

Approach

The insights presented in this framework emerged from an extensive stakeholder engagement process comprising five in-person workshops held throughout 2025. The core engagement consisted of four stakeholder-specific workshops focusing on researchers, data custodians, synthetic data experts, and members of the public. Following these workshops, we additionally held a fifth focused workshop specifically exploring regulatory requirements and disclosure risk considerations in synthetic data. Across all five in-person workshops, we engaged a total of 143 participants representing 38 different organisations spanning the UK data research ecosystem.

5 **Stakeholder Workshops**

143 **Participants**

38 **Organisations Represented**

Researchers were recruited through TRE networks, and targeted dissemination through relevant mailing lists and professional networks. Experts were specifically selected based on demonstrated experience with synthetic data. Data custodians were recruited through established TRE networks and data governance communities, ensuring representation from different data organisations. Members of the public were recruited through established public involvement panels across the UK.

Researcher Workshop
Bristol – April 2025

The researcher workshop brought together individuals who regularly use TREs for data research, representing diverse career stages from PhD students to senior investigators across different research backgrounds. The primary aim was to understand researchers' practical requirements for synthetic data across different stages of the research lifecycle, and what is needed to deliver real benefit for them.

Synthetic Data Expert Workshop
Birmingham – June 2025

The expert workshop involved participants with technical expertise in synthetic data generation, privacy evaluation, and statistical disclosure control. The aims were to understand current best practices for generating, evaluating, and documenting synthetic data, and to hear from organisations across the UK about their current approaches and implementation experiences.

Data Custodian Workshop
London – July 2025

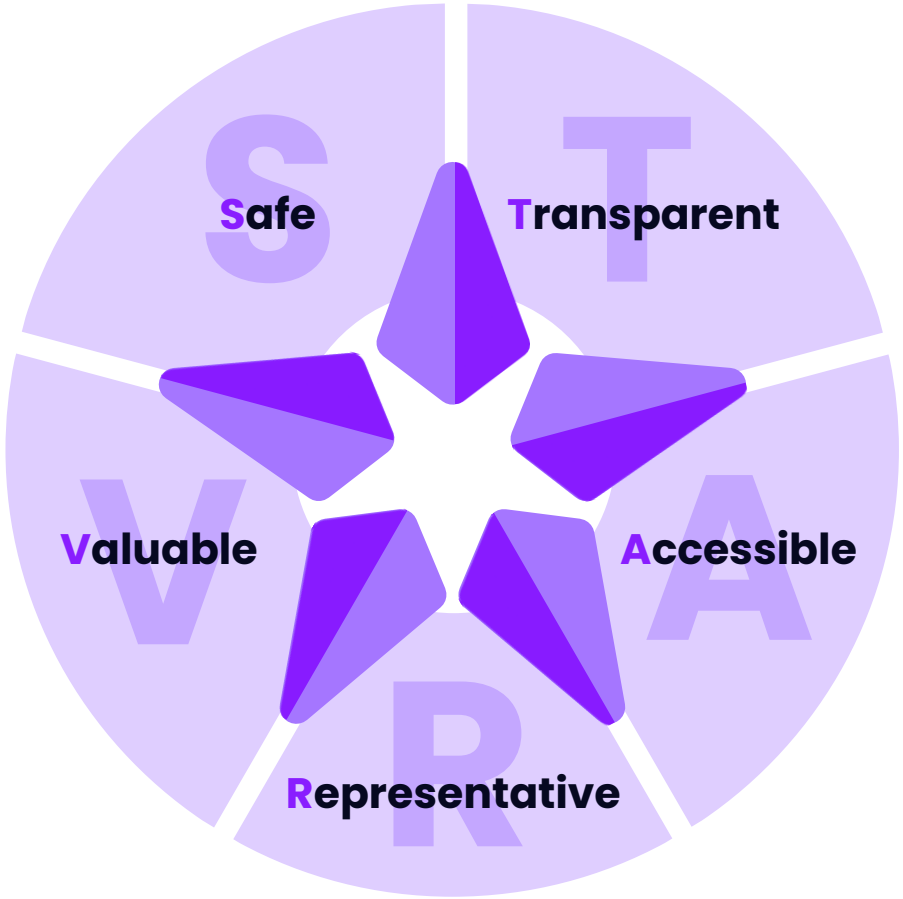
The data custodian workshop convened participants responsible for data governance decisions within TREs, NHS, research institutions, and national data services. The workshop explored what data custodians need to confidently approve synthetic data releases, and what is needed to increase trust, acceptability and ultimate adoption of synthetic datasets.

Public Workshop
Cardiff – October 2025

The public workshop engaged members of the public from patient involvement panels across the UK to understand perspectives on synthetic data. The workshop specifically focused on understanding how the public perceive synthetic data at different fidelity levels and to get their feedback into the VSTAR framework principles to ensure they meet their expectations for safe access and use.

Executive Summary

Through structured discussions and surveys across workshops, the UKSDCG explored stakeholder perspectives on the requirements, concerns, and priorities surrounding tabular synthetic data in TRE contexts. From these rich insights, we developed the VSTAR framework – a set of five core principles that emerged as essential for ensuring synthetic data is created and used responsibly in a way that delivers value. The framework addresses the full lifecycle of synthetic data: from initial generation guided by clear use cases, through rigorous privacy evaluation and transparent documentation, to proportionate access models and quality assurance. Each principle emerged directly from stakeholder discussions about their specific needs, concerns, and priorities. Together, these principles provide practical guidance for organisations seeking to generate, evaluate, govern, and deploy synthetic data within the context of Trusted Research Environments (TREs) and sensitive data research.



V **Valuable :** Synthetic data should meet the requirements and demonstrate clear utility for a defined use case. From the start, the intended application should guide data generation methods and evaluations, making sure that the synthetic data meets expectations and is genuinely useful.

S **Safe :** Synthetic data should minimise disclosure risk through robust, transparent generation methods and rigorous privacy evaluations to ensure GDPR-compliant anonymisation to enable safe release. Reputational risk should also be considered and managed through clear communication and appropriate safeguards.

T **Transparent :** The synthetic data generation process should be clearly documented and easy to understand, including the methods used, assumptions made, and any limitations. Evaluations should also be clearly communicated so that stakeholders can make informed decisions.

A **Accessible :** Synthetic data should generally often have lower governance barriers and simplified approval processes compared to real data. Safe sharing can be enabled through user agreements, making data more readily available while still maintaining responsible oversight.

R **Representative :** High-quality synthetic data should reflect the structure, diversity, missingness, irregularities, and patterns of the original data. It must be realistic, plausible, and fair, avoiding the amplification of biases and supporting trustworthy, equitable outcomes in downstream tasks.

Valuable

Synthetic data must be fit-for-purpose, with its generation and evaluation strictly guided by a clearly defined use case from the outset. Rather than viewing value as an inherent property of the data itself, it should be measured by its utility in supporting specific tasks, whether that be early-stage research planning, software testing, or complex model training. Decisions made throughout the data lifecycle, from the selection of generation methods to the design of access controls, must be driven by these requirements to ensure the data delivers tangible benefits rather than simply consuming resources.

For researchers, the primary value proposition is the acceleration of workflows by improving the significant delays often associated with accessing real sensitive data. Low-fidelity synthetic data (Levels 1 and 2) is particularly valuable for preparatory activities such as code development, data discovery, and some types of training, allowing researchers to debug pipelines and assess feasibility while waiting for governance approvals. While high-fidelity data (Levels 3 and 4) is often required for analytical tasks like privacy-preserving AI or improving analysis, researchers expressed caution about using it for definitive findings, often preferring to validate results against real data before publication to ensure reliability.

Data custodians see significant operational value from synthetic data through improved efficiency and risk management. By providing researchers with synthetic datasets for initial exploration, TREs can reduce the volume of unsuitable or incomplete access applications, thereby streamlining the approval pipeline and reducing the administrative burden on staff. Additionally, synthetic data serves as a safe "sandbox" for innovation, enabling organisations to rigorously test new technologies, such as federated analytics or collaborative tools, without exposing sensitive real data or navigating complex governance for unproven systems.

To fully realise this value, organisations must assess the specific problem they are trying to solve; simply using synthetic data to circumvent inefficient governance processes was viewed sceptically by public participants, who argued that resources might be better spent fixing the underlying access issues.

Therefore, a clear value proposition is essential: organisations should verify whether the desired outcome, such as faster access or improved data discovery, is best achieved through synthetic data or alternative improvements like improved metadata. This clarity helps prevent synthetic data from becoming a "solution in search of a problem" and ensures it is deployed only where it offers distinct advantages over other approaches.

Finally, evaluating value requires a broader understanding of "utility" that moves beyond just statistical accuracy. The framework distinguishes between analytical utility (statistical fidelity), functional utility (suitability for tasks like code testing), and operational utility (improvements to workflows and access speeds). For many high-value use cases like research planning or pipeline development, functional and operational utility are far more important than precise statistical replication. Aligning the generation method and evaluation metrics with these specific utility types ensures that the synthetic data is genuinely useful for its intended audience rather than being optimised for irrelevant metrics.

Safe

While synthetic data is often framed as a privacy-enhancing technology, it is not automatically classified as non-personal data under GDPR; rather, it must be evaluated to demonstrate that it meets effective anonymisation standards, specifically satisfying tests for singling out, linkability, and inference. An important distinction was made in the workshops between technical disclosure risk and reputational risk. While experts agree that well-generated synthetic datasets often carry minimal technical risk of re-identification, data custodians frequently cite reputational risk – such as the fear of public misunderstanding, mistakenly believing they have identified themselves or researchers misusing the data unintentionally, as the primary barrier to release. Therefore, ensuring safety requires not only minimising disclosure risk through robust technical means but also managing reputational risk through clear communication and stakeholder engagement.

To ensure technical safety, organisations must conduct comprehensive risk assessments, such as Data Protection Impact Assessments (DPIAs), before generation begins.

For low-fidelity data (Levels 1 and 2), basic identity disclosure checks are currently often used by organisations as sanity checks as these datasets preserve minimal information. However, higher-fidelity synthetic data, which retains complex relationships and distributions, requires far more extensive evaluation, including testing for membership inference and attribute disclosure risks, to ensuring compliance with anonymisation principles. The context of access is also paramount; risk assessments must consider the "means reasonably likely to be used" by an intruder, acknowledging that data accessed within a secure TRE faces different risks than data released publicly.

Transparency in generation methods is a crucial component of safety and trust, and this is heavily influenced by whether the synthetic data is publicly or privately sourced. Publicly-sourced synthetic data is generated using only open metadata or aggregate summary statistics that have already passed through statistical disclosure control. Custodians expressed significantly higher confidence in this approach, viewing it as carrying minimal disclosure risk since the inputs were already deemed safe for publication, and was something they were already familiar with. In contrast, privately-sourced synthetic data is generated directly from real, sensitive records to preserve complex relationships. While valuable for advanced analytics, this approach introduces greater privacy risks and requires more robust evaluation and transparent documentation to justify safety. Workshop participants also expressed a strong preference for transparent, interpretable generation methods (such as Bayesian networks) over "black box" AI techniques like Generative Adversarial Networks (GANs), as transparent methods allow custodians to scrutinise the process and understand exactly what information was used and is retained.

Communicating safety to non-technical stakeholders remains a significant challenge. Data owners often find quantitative privacy metrics difficult to interpret, which hinders confident decision-making. To address this, the framework recommends translating complex technical evaluations into accessible formats, such as "traffic light" risk categories (low, medium, high) accompanied by plain-language explanations, and contextualised risk.

Transparent

Transparency was seen as a crucial aspect for building trust in stakeholders. Without comprehensive documentation explaining how data was created, evaluated, and validated, stakeholders, from researchers to data owners, are reluctant to approve or use it. To address this, every synthetic dataset must be clearly labelled as synthetic and accompanied by well-structured documentation. This documentation serves as a fundamental requirement for enabling informed decision-making, allowing data owners to confidently approve releases and researchers to assess whether the data is fit for their specific needs.

A central component of transparency is explicitly defining the intended use cases and disclosing known limitations to prevent misuse. Documentation should provide positive guidance on appropriate applications, such as code development or training, while prominently warning against inappropriate uses, such as using low-fidelity data for statistical inference. Researchers specifically require clarity on data provenance, including details on the source data version and any preprocessing or cleaning applied prior to generation, as these factors significantly influence how code developed on synthetic data will transfer to the real environment. Furthermore, providing clear visual comparisons between synthetic and real data distributions enables users to quickly and intuitively assess the data's similarity and quality without needing deep technical expertise.

Transparency must also be tailored to different audiences through a multi-layered approach that combines technical rigour with accessible communication. Data owners and governance teams, who may not be technical experts, require plain-language explanations of generation methods and risk evaluations to defend release decisions and maintain public trust. Instead of relying solely on abstract metrics, risk evaluations should be presented using intuitive categories, such as traffic-light systems, alongside contextualised explanations of what those risks mean in practice. Additionally, experts emphasise the need to explicitly document all assumptions made during generation and evaluation, including privacy threats considered and thresholds used, to allow for critical review and prevent overconfidence in the results.

Accessible

To realise the true value of synthetic data for many early-stage research use cases such as pipeline development, enabling training, and supporting research planning, it must be governed with lower barriers and more simplified approval processes than real data. If accessing synthetic data involves the same lengthy delays and administrative hurdles as accessing sensitive real data, much of its utility is negated, particularly for preparatory tasks like code development. Accessibility must be proportionate to risk, moving away from a "one-size-fits-all" approach to a tiered model where governance stringency scales with data fidelity. Low-fidelity synthetic data which carries minimal disclosure risk, should be made readily accessible to maximise its functional utility, while higher-fidelity synthetic data may require additional safeguards to balance utility against higher potential risks.

The framework outlines a spectrum of access models ranging from open to secure. For low-fidelity structural data, data owners expressed comfort with "Registered Access," where users simply provide contact details without facing formal approval delays. For higher-fidelity datasets, "Safeguarded Access" was seen as suitable by most data custodians, enabling data to be shared outside of TREs but protected by robust End User Licence Agreements (EULAs). These agreements play an important role in managing misuse risk (rather than just disclosure risk) by requiring users to explicitly acknowledge the synthetic nature of the data, understand its limitations, and agree to prohibitions on re-identification attempts.

Researchers emphasised that for a lot of tasks, the ability to download data and work in their own local environments is paramount; restricting these low-risk datasets to secure TREs undermines their primary value. However, many data custodians felt more comfortable making some higher-fidelity synthetic data available via “Controlled Access” within a TRE, but with significantly lower governance and access controls.

Additionally, while automated approvals are efficient for low-risk data, many stakeholders, particularly the public, prefer a "human-in-the-loop" approach for higher-fidelity data to ensure accountability and responsible stewardship.

For high-fidelity synthetic data which is used for complex analysis or machine learning development, secure access within a TRE was seen as the most appropriate to maintain oversight. Ultimately, the goal is to apply the minimum level of restriction necessary to manage genuine risks, ensuring that synthetic data can flow freely enough to drive innovation and skill-building while maintaining the necessary safeguards to protect privacy and public trust.

Representative

High-quality synthetic data must authentically reflect the structure, patterns, and diversity of the original data to be considered fit-for-purpose. For low-fidelity synthetic data intended for functional tasks like code development and training, representativeness focuses on structural fidelity and "authentic messiness". Researchers emphasised that these datasets must preserve the imperfections, irregularities, and missingness patterns found in the real world, as exposure to realistic data quality issues is essential for building robust analysis pipelines and error-handling mechanisms. This includes maintaining accurate variable names, data types, and value ranges , while ensuring logical coherence through plausibility constraints, such as ensuring birth dates precede death dates or maintaining temporal sequences, so that the data behaves according to real-world rules.

For high-fidelity synthetic data designed for analytical purposes or AI model development, representativeness extends to statistical accuracy and the critical management of bias. Workshop participants raised significant concerns that synthetic data could inadvertently reinforce or amplify existing inequities present in source datasets, leading to discriminatory outcomes in downstream applications. To address this, organisations must assess source data biases before generation and rigorously evaluate the resulting synthetic data using fairness metrics. This evaluation must also navigate the tension between privacy and representation; while removing outliers is often used to reduce disclosure risk, stakeholders noted this can ethically fail by erasing minority populations or rare conditions from the data. Consequently, any analytical findings or AI models developed on synthetic data should be validated against independent real-world datasets before being considered reliable.

	Design	Generation	Evaluation	Release
Value	**V1** Clearly Define the Intended Use Case and Requirements **V2** Understand the Value Proposition	**V3** Align Generation Method with Use Case & Requirements	**V4** Evaluate Appropriate Type of Utility for the Use Case	
Safety	**S2** Carry out a Risk Assessment	**S3** Consider the Balance needed between Privacy and Utility **S5** Prioritise Transparent Generation Methods	**S4** Assess Disclosure Against Anonymisation Principles	**S1** Consider the Context the Synthetic Data will be Used **S6** Consider if Reputational Risk is the Main Barrier to Release
Transparency		**T4** Document Data Provenance and Preprocessing	**T3** Provide Clear Visual Comparisons **T5** Provide Accessible Summaries & Contextualised Risk **T6** Document any Assumptions Made in Evaluations	**T1** Synthetic Data should be Clearly Labelled & Documented **T2** Clearly Define Intended Use Cases & Limitations
Access				**A** Select Appropriate Access Model
Representative		**R1** Ensure Quality, Plausibility and Logical Coherence **R3** Authentically Preserve Data Messiness and Irregularities	**R2** Assess and Address Bias for Analytical Use Cases	

1

Valuable

Principle 1: Synthetic data must serve a clearly defined purpose and demonstrate genuine utility for that specific use case, with all decisions throughout the data lifecycle, from generation methods to governance arrangements, driven by explicitly defined requirements and intended applications. Value is achieved by aligning fidelity levels with actual use case needs, understanding distinct value propositions for different stakeholder groups, and ensuring utility evaluations capture what truly matters for the intended context rather than pursuing unnecessarily high fidelity that adds complexity, risk, and governance burden without delivering additional benefit.

01. Valuable

Synthetic data must be valuable - it should serve a clearly defined purpose and demonstrate utility for that purpose throughout the data lifecycle. The generation and evaluation of synthetic data should be guided by the specific use case it is intended to support, whether that be machine learning model training, exploratory data analysis, software testing, teaching, or another defined objective. Decisions throughout the synthetic data lifecycle, from generation methods to release decisions and access controls, should be driven by these clearly defined use cases and requirements to ensure the synthetic data delivers real, tangible benefits to all stakeholders involved. This ensures the synthetic data is fit-for-purpose and delivers benefits while avoiding wasted effort in generating unnecessarily high-fidelity synthetic data or implementing disproportionately restrictive governance that undermines practical utility.

Summary of Key Recommendations:

01. Clearly Define the Intended Use Case and Requirements

Before generating synthetic data, explicitly identify the specific use case it will support, the technical requirements this demands, and the primary value driver (e.g., faster access, reduced risk, external sharing). These factors should directly shape generation methods, required fidelity level, and governance approach, ensuring the synthetic data is fit-for-purpose and delivers genuine value to end users.

03. Identify a Suitable Generation Method

Select generation methods aligned with the intended use case and required fidelity: low-fidelity data can often use basic metadata and summary statistics for rapid, low-risk applications, while high-fidelity data may require more sophisticated techniques to capture complex relationships. Method selection should also account for interpretability, transparency, and scalability, as these factors impact governance, trust, and operational feasibility.

Higher fidelity is not always better and may introduce unnecessary complexity, cost, disclosure risk, and governance burden. Low-fidelity synthetic data, capturing structure, data types, and basic statistical properties, is often sufficient for many high-value use cases including code development, data discovery, and training. Reserve higher-fidelity generation for use cases that genuinely require preserved statistical relationships or analytical validity. This alignment ensures efficient use of resources while delivering appropriate utility for the intended application.

02. Understand the Value Proposition

Clearly identify why synthetic data is needed, what specific problem it solves, and for which stakeholder group, recognising that researchers, data custodians, and the public face fundamentally different challenges and therefore perceive value differently. For researchers, value may centre on eliminating access delays for preparatory work. For data custodians, value may lie in operational efficiency and reduced approval burden. For the public, value may focus on accelerating beneficial research while maintaining privacy. Before committing to synthetic data generation, consider whether alternative approaches, such as streamlining governance processes, improving metadata provision, or enhancing TRE infrastructure, might better address the underlying problem.

04. Evaluate Appropriate Type of Utility

Align utility evaluation with the intended use case by considering multiple dimensions of utility rather than relying solely on statistical similarity metrics. Analytical utility measures statistical fidelity and whether the data supports valid analytical conclusions. Functional utility assesses whether the data enables specific tasks such as code development, feasibility assessment, or training, which may not require high analytical utility. Operational utility considers whether the synthetic data improves access, reduces burden, or enables new workflows. For early-stage or exploratory tasks, functional and operational utility may matter more than precise statistical fidelity, ensuring evaluations capture what truly matters for the specific context rather than emphasising metrics irrelevant to the intended application.

01. Valuable

When deciding whether and how to create synthetic data, one of the earliest and most important questions is what value it is intended to deliver. Clarifying value from the outset makes it possible to define the requirements, level of fidelity, and governance conditions needed for synthetic data to be genuinely useful. For researchers, value centred on enabling early exploratory work, accelerating code development, and reducing delays caused by governance processes. For data owners and TREs, value was closely linked to operational efficiency, reducing the burden on secure environments, and supporting innovation while maintaining privacy protections.

By bringing together the perspectives of different stakeholders, we built a comprehensive understanding of the different ways synthetic data can create tangible benefits within the context of sensitive data research and TREs, and the conditions that must be met for it to be considered valuable.

1.1 What makes Synthetic Data Valuable?

1.1.1 Researcher Perspectives

Researchers identified the primary value of synthetic data in its ability to accelerate research workflows by overcoming the significant delays often associated with accessing real data within TREs. These delays typically arise from necessary but time-consuming processes including ethical approvals, information governance reviews, and technical access arrangements, which can extend timelines by several months.

While there is often a tendency to jump straight to generating high-fidelity synthetic data, participants felt this approach could introduce additional governance requirements and safeguards that recreate many of the same delays researchers hope to avoid. Workshop participants emphasised that the greatest value for early-stage research often comes from lower-fidelity synthetic data, where datasets are easier to access to support initial exploration, feasibility testing, and code development without triggering heavy governance processes.

Across the discussions, three use cases emerged as the most beneficial and immediately impactful: code development, data discovery, and researcher training. These use cases share a common characteristic in that they represent preparatory activities that do not require access to the full complexity of real data, yet they are frequently bottlenecked by the difficulty of obtaining data access.

Code Development and Data Discovery

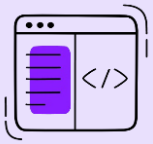
Synthetic data allows researchers to develop, test, and debug analysis scripts or pipelines without needing access to sensitive datasets. By providing a low-risk environment for technical work, it reduces delays caused by restricted access and enables iterative development in parallel with secure data approvals. Researchers described scenarios where months of waiting for data access could be productively used to build and refine analytical workflows, ensuring that once real data becomes available, substantive analysis can begin immediately rather than being delayed by technical setup and troubleshooting.

Similarly, synthetic datasets allow researchers to explore data structure, understand available variables, and perform preliminary investigations before committing to formal access applications. This facilitates more informed planning and design of analyses, reducing time wasted on proposals or queries that prove infeasible once real data characteristics are revealed. Several participants spoke about experiences of requesting access to datasets that, upon eventual receipt, turned out to be unsuitable for their research questions due to unexpected limitations in coverage, granularity, or variable definitions. Researchers felt that synthetic data can help to address this problem by providing a richer preview than metadata alone, enabling them to make better-informed decisions about which datasets to pursue.

For these early-stage and exploratory tasks, researchers indicated that L1 or L2 synthetic data is often sufficient. The main value of this type of synthetic data lies in its accessibility: it captures basic structural and statistical features such as variable types, available values, distributions, and patterns of missingness, and can be safely shared outside secure environments. This enables experimentation, skill-building, and early-stage planning with minimal governance barriers.

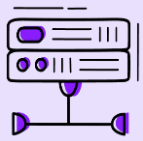
Use Cases of Synthetic Data in Sensitive Data Research

Data Discovery & Research Planning
Provides researchers with a realistic preview of dataset structure and variables before applying for access, enabling better-informed decisions about pursuing full data access.



Code Development
Allows researchers to develop and test analytical code during approval processes, enabling immediate analysis once real data access is granted.

Training & Teaching
Enables students and early career researchers to practice data science skills in educational settings without compromising sensitive information.



Supporting Innovation
Allows development and testing of new technical solutions, such as federation, before deployment to real data environments, while governance is established.

Privacy-Preserving AI Models
Enables training and release of AI models from secure environments while preserving privacy and reducing information leakage risks.

Enhancing Analysis
Augmenting real datasets by balancing classes, increasing sample sizes, imputing missing values, and simulating new scenarios for model testing.

Importantly, researchers stressed that even at these lower fidelity levels, synthetic data must reflect the authentic characteristics of real datasets, including their imperfections, inconsistencies, and complexities. Overly cleaned or simplified synthetic data that presents an idealised version of the dataset was seen as failing to prepare researchers adequately for the challenges they would encounter when working with real data.

Training & Teaching

When it came to training and teaching, requirements varied depending on the learning objectives and the depth of analytical understanding being developed. For methods-focused training such as learning programming skills or becoming familiar with TRE environments and analytical tools, L1 or L2 synthetic data was considered adequate. It allows trainees to work with realistic structures and formats while avoiding disclosure risks, making it ideal for technical skill development. This type of training benefits significantly from the accessibility of lower-fidelity synthetic data, as it enables educators to distribute datasets freely to students without the logistical and governance challenges associated with secure data access.

However, for training that involves more analytical depth, such as teaching statistical inference, causal modelling, or building AI models, higher-fidelity synthetic data was viewed as necessary. In these cases, preserving realistic relationships and patterns between variables is critical for ensuring that lessons learned and results generated provide meaningful feedback to learners. Researchers noted that students learning advanced analytical techniques need to understand not just how to write code, but how to interpret results, assess model performance, and recognise patterns that are only visible when data retains its inherent structure and correlations. Without this realism, training risks becoming too abstract or disconnected from the practical challenges of working with real-world data.

There was some discussion about whether higher-fidelity synthetic data used for training purposes should remain within TREs or be made available outside secure environments. Some participants saw value in requiring students to work within TREs when using more realistic synthetic data, viewing this as an opportunity to familiarise learners with controlled data environments and prepare them for the practices and constraints they would encounter when conducting real research.

Others argued that this approach could create unnecessary hurdles and limit the flexibility needed for effective teaching, particularly in contexts where students may not have formal research affiliations or where the primary goal is skills development rather than research output.

Analytical Use Cases

In contrast to the early-stage use cases described above, applications that involve advanced analysis, private AI model development, addressing bias and representational issues, or certain federated analytics scenarios were seen as requiring higher-fidelity synthetic data. Here, the key value requirement shifts from accessibility to utility: high-fidelity synthetic data must preserve complex relationships and distributions to support realistic analyses while simultaneously protecting individual privacy. This is especially critical for releasing privacy-preserving AI models, where analytical validity must be balanced against disclosure risk. Without sufficient fidelity, models trained on synthetic data may fail to generalise to real-world scenarios or may introduce unexpected biases that undermine their intended purpose.

Researchers highlighted that higher-fidelity synthetic data must meet additional requirements beyond those necessary for lower-fidelity applications. These include larger sample sizes to ensure adequate representation, robust evaluation of statistical fidelity to confirm that key relationships are preserved, and careful governance to ensure that utility is maximised without compromising confidentiality. The generation and release of such datasets was recognised as requiring significantly more sophisticated technical approaches, including the application of robust privacy measures.

An important distinction emerged during discussions regarding the types of analytical work researchers felt comfortable conducting with synthetic data. There was notably greater comfort and acceptance of using high-fidelity synthetic data for methods-based research compared to inference-based research.

However, researchers working in machine learning fields expressed notably different perspectives on the analytical use of synthetic data compared to those from traditional statistical or epidemiological backgrounds.

Many machine learning researchers reported that synthetic data is already an established component of their standard workflows, and they routinely use it for analytical purposes beyond just method development. These researchers highlighted that synthetic data generation techniques such as SMOTE (Synthetic Minority Over-sampling Technique), data augmentation, and simulation-based approaches are commonly employed in machine learning practice to:

- Address class imbalance: Generating synthetic examples of underrepresented classes to improve model fairness and performance
- Augment training datasets: Creating additional training samples to improve model robustness and generalisability
- Enhance model performance: Using synthetic data to supplement real data in contexts where real data is scarce, expensive to collect, or subject to access restrictions
- Improve representation: Generating synthetic examples to ensure adequate representation of minority populations or rare scenarios

Federation & Supporting Innovation

Federated analytics emerged as a use case with more variable requirements depending on the specific objectives. For initial feasibility testing or proof-of-concept work to demonstrate the viability of federated approaches, some participants felt that L1 or L2 synthetic data could suffice. However, others argued that more realistic L3 synthetic data would be necessary to properly simulate computational demands, test federated analytics workflows, and uncover practical implementation challenges that only become apparent when working with data that reflects genuine complexity and scale.

Beyond federated analytics specifically, synthetic data was seen as playing a crucial enabling role in supporting the development and testing of new innovative technologies and infrastructure while governance frameworks are being established. This emerged as a particularly valuable application across all stakeholder groups, as it addresses a common challenge in data research: the long lead times required to establish appropriate governance structures for novel technological approaches.

Participants highlighted that synthetic data provides a "safe sandbox" environment that allows organisations to:

- Demonstrate technical feasibility: Test whether new technologies, platforms, or analytical approaches work as intended before investing in full governance processes
- Identify technical and operational challenges: Uncover implementation issues, performance bottlenecks, or interoperability problems early in the development process
- Build evidence for governance discussions: Generate concrete examples and proof-of-concept demonstrations that help governance teams understand what is being proposed and assess associated risks
- Iterate and refine systems: Develop and optimise new infrastructure without the delays and constraints associated with accessing real sensitive data
- Establish trust and understanding: Allow stakeholders to see new technologies in action, building confidence and familiarity before real data is introduced

Concerns

Despite recognition of the potential value of high-fidelity synthetic data for advanced applications, researchers expressed significant uncertainty and caution about using synthetic data as a direct basis for research findings or publications. This was particularly evident in discussions about training AI models aimed at improving performance or fairness. Researchers were reluctant to publish findings based on synthetic datasets for such purposes, citing concerns over the lack of transparency in how the data was generated, the potential for synthetic data to introduce or amplify misleading patterns, and its inability to fully capture the complexities and biases present in real data.

There was a strong preference expressed for validation of any findings derived from synthetic data against real data before those findings could be considered reliable or publication-ready. Without such validation, researchers worried that analyses could produce results that appeared plausible but were in fact artefacts of the synthesis process rather than genuine insights.

This concern was especially problematic when considering applications that could influence clinical practice, policy decisions, or resource allocation, where the consequences of acting on misleading findings could be substantial.

Several participants stated that if synthetic data were to be used for analytical purposes, they would prefer to generate it themselves, ensuring that they understood the assumptions, methods, and limitations involved, rather than relying on pre-generated datasets provided by others. This preference reflects both a desire for methodological control and a concern about the interpretability of synthetic data generation processes, particularly when complex AI-based methods are employed. Researchers emphasised that trust in synthetic data for analytical purposes would require comprehensive documentation of the generation process, quantitative assessments of fidelity, and clear guidance about the types of analyses for which the synthetic data is and is not appropriate.

Recommendation V.1

Clearly Define the Intended Use Case and Requirements of the Synthetic Data

To maximise the value of synthetic data, organisations must identify the specific intended use case and requirements at the outset. This involves clearly defining the task or challenge the data supports (such as training, method testing, or research planning) and pinpointing the primary value driver (e.g., faster access, reduced risk, or external sharing etc). These drivers should directly shape the technical requirements, the required level of fidelity, and the corresponding governance, thereby ensuring the generated data is fit-for-purpose and is truly valuable to the end user.

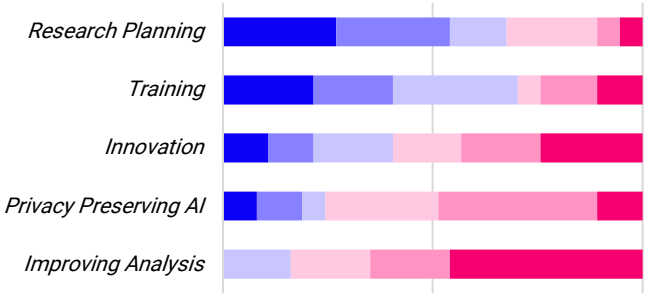
1.1.2 Data Custodian Perspectives

From the perspective of data owners and TREs, synthetic data offers significant operational benefits that enhance efficiency, reduce risk, and support innovation.

One of the most widely cited advantages from participants was the ability of synthetic data to optimise the data access pipeline. By allowing researchers to explore datasets and evaluate the feasibility of their analyses before submitting formal applications for access to the real data, synthetic data helps ensure that subsequent applications are more focused, technically realistic, and therefore more likely to meet approval criteria. This significantly reduces the administrative burden on TRE staff and data owners by decreasing the volume of incomplete, unsuitable, or iterative access requests, ultimately streamlining the process of granting access while maintaining rigorous governance standards.

Synthetic data was also seen as being able to play a critical role in researcher onboarding and training within TREs. By enabling researchers to become familiar with the secure environment, workflows, and analytical tools before they gain access to sensitive data, this reduces the need for intensive, hands-on support during actual analyses, freeing up TRE staff time and resources.

Fig 1.1 What do you see as the main benefits of synthetic data? (ranked from highest benefit to lowest)



(blue = highest benefit, red = lowest benefit)

In addition, synthetic data enables data custodians to experiment safely with new technologies, methods, and workflows. For instance, TREs can rigorously test federated systems, evaluate new technologies, or trial collaborative analytical tools using synthetic datasets before deploying them with real data. This capability allows innovation to proceed in parallel with the development of necessary governance frameworks, providing essential evidence on feasibility, security, and trustworthiness that can inform and accelerate future policy decisions.

Concerns and Scepticism

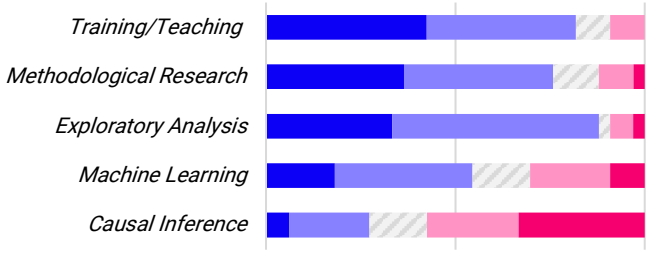
However, the perceived value of synthetic data was not universally accepted. Several data owners expressed that the concept remains abstract without concrete, validated, real-world examples of successful implementation and demonstrable impact. They wanted to see clear case studies demonstrating tangible operational and research benefits before committing substantial resources to its adoption and governance.

Others voiced strong scepticism around the use of synthetic data for complex analytical tasks or AI model development. Concerns centred on whether synthetic datasets could adequately reflect the statistical nuances, subtle biases, and intrinsic complexity of the real data, and on the potential risks of drawing misleading or erroneous conclusions that could impact public health or policy. Some critically questioned the rationale for using synthetic data at all when real data already exists, arguing that relying on synthetic datasets to address long access times is effectively treating the symptom rather than addressing the root cause - the fundamental inefficiencies and bottlenecks in the established data access processes. In some cases, this scepticism was compounded by a lack of familiarity with the rigorous technical methods used to generate and evaluate synthetic data, making it inherently harder for these stakeholders to trust its outputs.

Acceptability of Use Cases & Fidelity Levels

Overall, data custodians were far more confident in the value of synthetic data for operational purposes such as researcher training, code development, and feasibility assessment. In these contexts, the ease of sharing, and clear time-saving benefits were seen as strong, unambiguous justifications for adoption.

Fig 1.2 How comfortable would you feel about high-fidelity synthetic data being used for the following purposes? (ranked from highest benefit to lowest)



In contrast, synthetic data to support real-world analysis prompted greater caution and hesitancy. While they acknowledged its transformative potential for enabling privacy-preserving research, mitigating issues with bias and advanced AI model development, they were reluctant to support its use for analytical purposes without robust evidence of reliability, rigorous evaluation frameworks, and demonstrably strong governance safeguards in place. This hesitancy is likely driven not solely by evidence of technical limitations, but also by a lack of exposure to, and understanding of, the potential of synthetic data in these advanced AI use cases, suggesting that focused communication, practical demonstration, and targeted education could be key to building confidence.

1.1.3 Public Perspectives

Members of the public often expressed confusion regarding the complex, lengthy, and resource-intensive governance processes surrounding the use of real, sensitive data. The approval times, often extending many months for legitimate research proposals, were viewed as a systemic inefficiency.

The resulting sentiment was a shared belief that the focus on generating synthetic data was misplaced or premature. Participants asked why efforts and resources were being dedicated to creating an approximation of the real data, rather than being directed toward addressing the root cause of the delays: the bottlenecks and inefficiencies within the real data access system. This perspective reinforced the position taken by several data custodians. This convergence of opinion from both data owners and the public underscores a critical finding of the workshops: a lack of clear, compelling value proposition for synthetic data risks undermining its adoption.

If the perceived purpose is merely to skirt existing, albeit flawed, access rules, stakeholders will resist its use. For synthetic data to be truly accepted as valuable, its benefits must extend beyond simply reducing wait times, it must enable entirely new forms of safe, scalable research or provide critical operational relief that existing processes cannot. This finding directly leads to the necessity of rigorously defining the value proposition.

Recommendation V.2

Understand the Value Proposition

It is important to clearly define why synthetic data is needed, what specific problem it is intended to solve, and for which distinct stakeholder group. The workshops highlighted that different stakeholders face different challenges and therefore attribute value in fundamentally different ways. For researchers, synthetic data may primarily solve the problem of delays in data access, enabling earlier code development, feasibility testing, and skills training. For data owners and TREs, the problem may be operational, such as the burden of processing incomplete or unsuitable data access requests, or the inability to safely trial new technologies and analytical methodologies. If the need is not well defined, synthetic data risks becoming a "solution in search of a problem," being deployed where it is unnecessary, for example in situations where direct use of real data would be just as fast, more accurate, and entirely within governance limits.

Before committing to synthetic data generation, organisations should consider whether alternative approaches might better address the underlying problem such as streamlining governance processes or improving the detail and quality of existing metadata.

By explicitly linking the creation and use of synthetic data to a well-understood, measurable problem and the specific stakeholders affected, projects can ensure it delivers meaningful and tangible value rather than consuming resources without clear justification.

1.2 Aligning the Design, Generation and Deployment of Synthetic Data with the Intended Use Case.

A recurring theme across the workshops was that the value of synthetic data is context-dependent: its usefulness is determined by the task it is intended to support. Low-fidelity datasets excel at enabling early-stage work and skills development with minimal risk, while high-fidelity datasets are required for tasks where analytical utility is needed. Participants consistently emphasised that defining the intended use case upfront is crucial for guiding the generation, evaluation, and deployment of synthetic data. This ensures that the output is not only safe and accessible but also genuinely fit-for-purpose, providing tangible benefits to researchers without introducing unnecessary risk or overhead.

Across all workshops, participants consistently emphasised that the intended application should be the starting point for any synthetic data initiative. Different use cases demand varying levels of fidelity and associated requirements, from low-fidelity datasets that capture only basic structure and summary statistics, to high-fidelity datasets that preserve complex relationships and distributions. For example, early-stage activities such as code development, data discovery, or training may only require low-fidelity synthetic data, prioritising accessibility, speed of production, and minimal disclosure risk. In contrast, tasks such as enhancing analysis, developing private AI models, or supporting federated analytics require higher-fidelity outputs with greater analytical utility, often necessitating more sophisticated generation methods and stricter governance. These differences underline a shared conclusion across stakeholder groups: the use case should drive the choice of generation method, the design of utility evaluations, and the conditions for deployment, ensuring the resulting synthetic data is both suitable and able to deliver meaningful value in practice.

1.2.1 Expert Perspectives on Synthetic Data Generation Practices

The expert workshop provided insights into how synthetic data should be generated to ensure it meets the needs of different requirements.

A clear distinction was made between publicly-sourced synthetic data generation (using information already in the public domain such as metadata and aggregate statistics) and privately-sourced (using the actual private data itself).

Fig 1.3 What aspects do you consider the most important for low-fidelity synthetic data? (blue = highest importance, red = lowest importance)

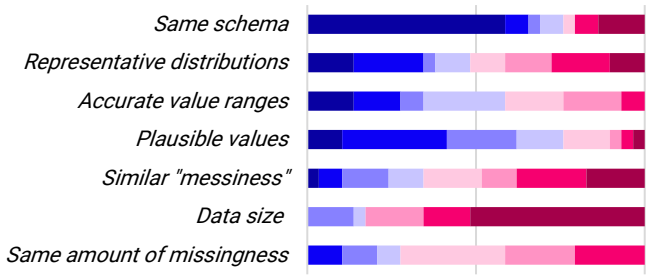
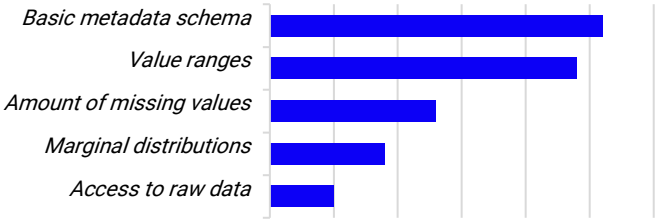


Fig 1.4 What is the minimum amount of information you need to create useful low-fidelity synthetic data?



For lower-fidelity synthetic data (L1 and L2), 95% of participants felt that only publicly-sourced information is required to create datasets that are still functionally useful. The essential inputs identified were basic metadata, value ranges for each variable, and the extent of missingness. Marginal distributions were considered less critical, and only a small fraction of participants (5%) believed that access to the raw underlying data was necessary. This indicates that effective low-fidelity synthetic data can largely be generated from aggregated summary statistics and metadata, allowing researchers to explore structure, plan their methods, and develop skills without compromising confidentiality.

Minimising reliance on raw data also reduces disclosure risks and simplifies governance, aligning with approaches such as NHS England's use of aggregate statistics, that have already passed disclosure control, to produce synthetic datasets.

For high-fidelity synthetic data, the generation process becomes significantly more complex. Experts reported relying primarily on methods such as Generative Adversarial Networks (GANs) and Bayesian networks, with other statistical approaches also applied depending on the dataset and use case. In selecting a method, three factors were repeatedly highlighted as critical: fidelity to the real data, privacy preservation, and interpretability/transparency of the generation process. High-fidelity synthetic data must preserve complex correlations, distributions, and interactions between variables, while simultaneously mitigating disclosure risk. Methods like Bayesian networks were frequently preferred because they provide high-quality outputs with transparent and traceable methods, allowing experts to explain and validate the generation process. GANs, while powerful for producing realistic outputs, were noted to pose challenges in interpretability and transparency making it more difficult to communicate to data custodians.

Participants also discussed practical considerations and trade-offs. High-fidelity generation often requires substantial computational resources, especially for large datasets or high-dimensional data, which can limit scalability. Decisions around the generation method must therefore consider not only utility and privacy but also operational feasibility. Experts emphasised that high-fidelity synthetic data is most suitable when analytical utility or privacy-preserving model development is critical, while less demanding tasks should leverage low-fidelity approaches to maximise accessibility and efficiency.

Finally, the workshop highlighted the importance of aligning generation approaches with the intended use case. Low-fidelity datasets can be ideal for exploratory work, training, and pipeline development, while high-fidelity data is better for tasks that demand accurate statistical inference or the development of privacy-preserving AI models. Experts stressed that a one-size-fits-all approach is ineffective; both fidelity and generation methods should be tailored to the problem being solved, balancing utility, privacy, interpretability, and resource constraints.

Real-World Examples of Synthetic Data Methods

Department for Education

ADR UK Longitudinal Educational Outcomes Synthetic Dataset

The LEO synthetic dataset was created by taking a random sample of the original data for each column, using a different randomisation key for each column. This approach preserves marginal distributions but doesn't maintain the relationships between the variables. This allows researchers to understand data structure, assess suitability, and draft code before accessing real data.

L2 Synthetic Data

Research Data Scotland

Scottish Public Sector Synthetic Datasets

RDS uses automated R code to create low-fidelity synthetic data from publicly available data dictionaries, ensuring that the synthetic data captures the structure, data types, value ranges and patterns of missingness. The focus is on supporting data discovery, training, and code development and have so far hosted two Census synthetic datasets from National Records Scotland.

L1 Synthetic Data

CPRD

Clinical Practice Research Datalink Primary Care Synthetic Datasets

CPRD developed synthetic datasets using a Bayesian Network-based approach, emphasising transparency and interpretability. They produced two high-fidelity datasets (COVID-19, cardiovascular) replicating complex clinical relationships to be used for analysis, and two medium-fidelity datasets (CPRD Aurum sample, CPRD Gold sample) resembling real-world data structure for research planning and training.

L3 Synthetic Data

NHS England

NHS England HES Synthetic Dataset

NHS England's HES artificial data uses a metadata scraper to extract anonymised aggregate statistics from the real data, which are then reviewed and signed-off through standard Statistical Disclosure Control procedures. A data generator then randomly samples from these anonymous aggregates to create artificial records which preserves some statistical properties of the data but no relationships.

L2 Synthetic Data

Recommendation V.3

Identify a Suitable Generation Method based on the Use Case & Requirements

Selecting the appropriate method for generating synthetic data is critical to ensuring it meets the intended objectives while balancing privacy, utility, and practicality. The choice of generation method should be guided by the specific use case and the fidelity required: low-fidelity synthetic data can often be produced using basic metadata and summary statistics, minimising disclosure risk and enabling rapid, accessible exploration or training. High-fidelity data, on the other hand, typically requires more sophisticated techniques such as Bayesian networks or Generative Adversarial Networks (GANs) to capture complex relationships and distributions, particularly when supporting analytical utility or privacy-preserving AI development. Experts emphasised that method selection should also account for interpretability, transparency, computational efficiency, and scalability, as these factors impact governance, trust, and operational feasibility. By aligning the generation method with the intended use and requirements from the outset, organisations can ensure that synthetic data is fit-for-purpose, delivering tangible value to researchers, TREs, and data owners without introducing unnecessary risk or overhead.

1.3 Assessing Different Types of Utility

1.3.1 Types of Utility

Evaluating the utility of synthetic data is essential to understanding whether it delivers real value for the intended task. Across the workshops, it became clear that the concept of "utility" is context-dependent, with different stakeholders interpreting it in different ways. In the synthetic data literature, utility is often assessed primarily in terms of analytical performance - how well models, statistical analyses, or machine learning tasks performed on synthetic data replicate results obtained from real data. While analytical utility is important for certain high-fidelity use cases, workshop participants emphasised that it represents only one dimension of value.

Two additional forms of utility emerged as equally important depending on the use case: **functional utility** and **operational utility**. Functional utility assesses whether synthetic data is suitable for a given task, such as software testing, pipeline development, or data exploration. These tasks may not require high analytical fidelity, but they do require the data to perform the intended function effectively. Functional utility can often be evaluated qualitatively, for example by considering whether researchers can complete tasks more efficiently, understand the dataset structure better, or develop skills in coding and analysis without unnecessary friction.

Operational utility, on the other hand, focuses on the impact of synthetic data on workflows, governance, and resource allocation. From the perspective of TREs and data owners, operational utility measures whether synthetic data improves access, reduces time spent on applications, lowers the burden on support staff, or enables safer experimentation with new technologies. This form of utility is particularly relevant for evaluating whether synthetic data delivers systemic improvements beyond individual research outcomes.

Many participants highlighted that current approaches to utility assessment are too narrowly focused on analytical performance, particularly machine learning benchmarks. While these metrics are useful for certain high-fidelity applications, they do not capture the broader spectrum of value that synthetic data can provide.

Participants stressed the need for more comprehensive evaluation frameworks that include other types of utility, tailored to the specific use case. By aligning utility assessment with the intended purpose of the synthetic dataset, organisations can ensure that synthetic data is genuinely useful, both for researchers to complete their tasks and for data custodians seeking to optimise operations and governance.

Recommendation V.4

Evaluate Appropriate Type of Utility

When assessing synthetic data, it is essential to align the evaluation approach with the intended use case and stakeholder needs. Rather than focusing solely on analytical utility, data practitioners should consider multiple forms of utility, including functional utility - whether the data effectively supports specific tasks such as code development or pipeline testing, and operational utility - whether it improves access, reduces delays, or lowers resource burden. Selecting the appropriate utility type ensures that evaluations capture what truly matters for the context in which the synthetic data will be used. For early-stage or exploratory tasks, functional and operational utility may be more relevant than precise statistical fidelity, whereas analysis or AI development will require stringent analytical utility measures. By explicitly defining the utility type at the outset, teams can ensure that synthetic data delivers real value, meets stakeholder expectations, and avoids unnecessary emphasis on metrics that are irrelevant for the task at hand.

Types of Utility in Synthetic Data

Analytical Utility
Measures how closely synthetic data preserves the statistical properties of the original dataset, including distributions, correlations, and relationships between variables. This is crucial for high-fidelity applications like statistical modelling, hypothesis testing, or machine learning, where results depend on accurate replication of underlying data patterns.

Functional Utility
Assesses whether synthetic data is suitable for specific tasks such as software testing, pipeline development, code writing, or data exploration. These tasks may not require high statistical fidelity but need the data to perform the intended function effectively, enabling researchers to work efficiently and safely without needing a close statistical match to the original.

Operational Utility
Focuses on how synthetic data impacts workflows, governance, and resource allocation from the perspective of data owners and TREs. It measures whether synthetic data improves access, reduces application processing time, lowers support staff burden, or enables safer experimentation with new technologies, delivering systemic improvements beyond individual research outcomes.

UK Synthetic Data Community Group
Developing Standards & Resources to Support Trust in Synthetic Data
To support the adoption of synthetic data, we aim to create standards for reporting generation and evaluation processes, as well as publish learning resources to provide stakeholders with an understanding of synthetic data. This enables trust in synthetic data to be used in a variety of applications in sensitive data research.
Standards
Setting standards for the evaluation and reporting of synthetic data.
Resources
Creating resources to help researchers, data owners and the public understand synthetic data.
Continue the conversation on our Slack Channel:
A Community for Advancing the Utilisation of Synthetic Data in Sensitive Data

2

Safe

Principle 2: Synthetic data safety requires both technical privacy protections to minimise disclosure risk and transparent communication to address reputational concerns that often pose greater barriers to release than technical risk itself. Organisations must conduct proportionate disclosure risk assessments aligned with fidelity levels and intended access contexts, while also building stakeholder confidence through transparent generation methods, clear documentation, and explicit consideration of the trade-offs between privacy, fairness, and utility.

02. Safe

The Safe principle recognises that synthetic data safety requires both rigorous technical measures and stakeholder confidence-building. While experts agreed that well-generated synthetic datasets often have very minimal technical disclosure risk, particularly when using transparent generation methods and robust privacy evaluation, data custodians emphasised that reputational risk often poses a greater barrier to release than disclosure risk itself. Concerns about public perception, potential misinterpretation, misuse, and negative media coverage can prevent organisations from releasing synthetic data even when privacy evaluations show acceptable technical risk levels.

Summary of Key Recommendations:

01. Consider the Context in which the Synthetic Data will be Accessed and Used

The intended access context, whether within secure TREs, or released publicly, determines appropriate privacy protections and governance arrangements. Organisations should use the access model type to inform decisions about fidelity levels, evaluation rigor, and additional safeguards like user agreements, recognising that different contexts warrant different protective measures.

03. Consider the Balance between Privacy, Fairness and Utility

Privacy-preserving techniques can inadvertently harm fairness by excluding minority populations, while maintaining high fidelity for representativeness may increase privacy risk. Organisations must make context-dependent judgements about these trade-offs based on their specific use case, explicitly documenting the rationale for decisions and any implications for different stakeholder groups or downstream applications.

05. Prioritise Transparent Generation Methods and Use Only the Minimum Information Necessary

Use only the minimum information from source data necessary to achieve intended purposes, reducing disclosure risk and simplifying governance. Prioritise transparent, interpretable generation methods as participants emphasised these build significantly greater confidence among data custodians and governance teams who must scrutinise generation processes, understand what information is retained or excluded, and defend release decisions.

02. Carry out a Risk Assessment

Before generating synthetic data, conduct a risk assessment, such as a DPIA, evaluating potential disclosure risks, source data sensitivity, characteristics of potential attackers, and how risks may evolve over time. Risk assessments should be documented, revisited as circumstances change, and used to inform appropriate mitigation measures and governance arrangements.

04. Assess Disclosure Risks Appropriately against Anonymisation Principles

Disclosure risk assessment should be proportionate to fidelity level and source data sensitivity: low-fidelity synthetic data typically requires only basic identity disclosure sanity checks, while high-fidelity synthetic data demands more comprehensive evaluation to address the GDPR anonymisation principles: singling out, linkability, and inference, to ensure regulatory compliance and defensible release decisions.

06. Identify Whether Reputational Risk is the Primary Barrier rather than Disclosure Risk

Clearly distinguish whether barriers to synthetic data release stem from disclosure risk or reputational concerns, as these require different mitigation strategies. Where evaluation demonstrates low risk meeting GDPR standards, yet approval remains difficult, the barrier is likely reputational, requiring focus on stakeholder communication, additional safeguards, and confidence-building rather than technical methods that won't address underlying concerns about perception, misunderstanding, or institutional caution.

02. Safe

Ensuring that synthetic data is safe is essential for its responsible development, governance, and deployment. Throughout the workshops, participants emphasised that safety is not an inherent property of synthetic data but a socio-technical outcome, shaped by how the data is generated, evaluated, contextualised, and governed. While synthetic data is often presented as a privacy-enhancing technology that can reduce or eliminate the risks associated with sharing real data, workshop discussions made clear that this safety must be demonstrated, not assumed, and should not just focus on disclosure risk. Participants agreed that synthetic data should be developed through a transparent privacy-by-design approach, evaluated against GDPR-defined anonymisation standards, governed through clear roles and responsibilities, and communicated transparently in ways that maintain public trust.

Experts expressed consistent agreement that for well-generated synthetic datasets, actual disclosure risk is often very low. However, data owners repeatedly emphasised that the biggest barrier to release is often not technical disclosure risk but reputational risk, driven by uncertainty about how synthetic data will be perceived, understood, or potentially misinterpreted by external audiences, including the public, media, and researchers. This distinction between technical and reputational risk emerged as an important finding: even when privacy can be quantified and controlled, concerns about public perception, the possibility of "naive intruders" mistakenly believing they have identified themselves in synthetic datasets, and the potential for damaging media coverage can prevent organisations from releasing higher-fidelity synthetic data that could deliver significant research value.

The Safe section therefore addresses two complementary but distinct aims: minimising disclosure risk through rigorous technical evaluation and generation methods, and managing reputational risk through transparent communication, stakeholder engagement, and governance frameworks that support confidence, accountability, and responsible sharing.

Safety, in this context, encompasses not only the privacy properties of synthetic datasets themselves but also the broader ecosystem of practices, standards, and relationships that determine whether synthetic data can be trusted and used responsibly.

This section explores perspectives from experts, data custodians, and members of the public on what it means for synthetic data to be safe. It explores the types of disclosure risks inherent in synthetic data and how evaluation approaches should differ based on fidelity level; the influence of source data characteristics and generation transparency on risk profiles; the operational and governance challenges of approving high-fidelity synthetic data; the difficulties in communicating privacy assessments to non-technical decision-makers; and the documentation, training, and accountability structures needed to ensure synthetic data is used appropriately. By addressing these different aspects, organisations can create synthetic data that is not only technically safe but also trusted, appropriately governed, and responsibly deployed in ways that enable its full potential while safeguarding privacy and maintaining public confidence.

2.1 Disclosure Risks in Synthetic Data

2.1.1 The Regulatory Context of Synthetic Data in the UK

Under UK GDPR, synthetic data is not automatically classified as non-personal data. Instead, whether synthetic data is considered personal data depends on whether it meets the standards for effective anonymisation. This distinction is important: synthetic data cannot simply be assumed to be safe or non-identifiable; rather, it must be evaluated to demonstrate that it does not pose risks to individual privacy.

The Information Commissioner's Office (ICO) addresses synthetic data in two key guidance documents. In their Privacy Enhancing Technologies (PETs) guidance, the ICO recognises synthetic data as a technique that can assist with data protection compliance. The ICO also briefly references synthetic data in their anonymisation guidance as a potential method for anonymising data.

For data to be considered effectively anonymised, it must satisfy three key tests:

1. **Singling out:** It should not be possible to isolate information relating to a particular individual from the dataset.
2. **Linkability:** It should not be possible to link records in the dataset with records in other datasets, in a way that identifies individuals.
3. **Inference:** It should not be possible to deduce information or attributes about individuals in the dataset with significant probability.

These principles form the foundation for assessing whether synthetic data can be treated as anonymous and therefore outside the scope of data protection regulations. The regulatory expectation is that organisations generating synthetic data will conduct thorough risk assessments and be able to demonstrate that their synthetic datasets meet these anonymisation standards before releasing them outside secure environments or treating them as non-personal data.

Furthermore, the ICO emphasises that the utility-privacy trade-off is context-dependent:

> *"You should consider the purposes and context of the processing when you decide what risk mitigation measures are appropriate balanced against what you need to fulfil your purposes. For example, if you intend to use synthetic data in a secure setting (eg a trusted research environment), then the risk of attack may be reduced. You could consider less stringent measures (eg adding less noise than if you were releasing the information to the world at large when using differential privacy)."*

This context-dependent approach to risk assessment emerged as a central theme across all workshops. Participants consistently emphasised that the appropriate level of privacy protection cannot be determined in isolation from how the synthetic data will be accessed, used, and governed. The "means reasonably likely to be used" principle in GDPR anonymisation assessments explicitly recognises that context matters: data accessed only within a secure TRE by trained, accountable researchers under suitable governance faces materially different risks than data released openly on the internet.

However, while the ICO's guidance acknowledges that TRE environments may reduce risk and justify lighter privacy measures, many expert workshop participants strongly cautioned against defaulting to TRE-only access for synthetic data unless truly necessary. They argued that confining synthetic data within TREs with approval processes and technical restrictions fundamentally undermines its value proposition for many key use cases. If synthetic data requires the same access pathways, waiting times, and usage constraints as real data, it offers little practical benefit for the preparatory and exploratory tasks, such as code development, data discovery, and training,—that participants identified as its most valuable applications.

Experts emphasised that for these use cases, the primary value of synthetic data lies precisely in its accessibility: the ability to download it, work on personal devices, share it with collaborators, and iterate rapidly without governance delays. Restricting synthetic data to TREs could negate these benefits, making it no more beneficial than accessing the real data and defeating the purpose of creating it in the first place. Several participants expressed frustration that risk-averse governance approaches might lead to synthetic data being over-protected, rendering it effectively unusable for the very purposes it was designed to support.

Conversely, data owners expressed more varied perspectives. While some acknowledged the importance of accessibility and recognised that lower-fidelity synthetic data could be safely released outside TREs, others felt considerably more comfortable with TRE-based access, viewing it as an additional layer of protection that reduces both disclosure and reputational risk.

For data owners unfamiliar with synthetic data or working in particularly sensitive contexts, the controlled environment of a TRE provided reassurance and accountability that made approval decisions more straightforward.

Ultimately, these contrasting perspectives highlight the need to find an appropriate balance between safeguards and accessibility, tailored to the specific characteristics of each synthetic dataset. The goal should be to apply the minimum level of access restriction necessary to manage genuine risks, rather than defaulting to the most conservative option. For low-fidelity synthetic data with demonstrably minimal disclosure risk, lighter-touch governance models that enable external access may be entirely appropriate. For higher-fidelity datasets or particularly sensitive contexts, TRE-based access or controlled release with user agreements may be warranted. The key is ensuring that access decisions are proportionate, evidence-based, and aligned with both the risk profile of the synthetic data and the intended value it aims to deliver.

Recommendation S.1

Consider the Context in which the Synthetic Data will be Accessed and Used.

The context in which synthetic data will be accessed - whether within a secure TRE by vetted researchers, downloaded by students, or released publicly - is critical for determining appropriate privacy protections and governance. Organisations should explicitly document the intended access model and use this to inform decisions about fidelity levels, evaluation rigour, and additional safeguards such as user agreements or access controls.

Workshop participants consistently emphasised that regulatory compliance should be embedded into the design and evaluation of synthetic data from the outset. This means selecting generation methods that incorporate privacy protections, conducting comprehensive disclosure risk assessments that address anonymisation principles, and documenting the rationale and evidence supporting the classification of synthetic data as anonymous.

Without this approach, organisations risk misclassifying synthetic data as non-personal when it may, in fact, still pose disclosure risks and therefore remain subject to GDPR obligations. Several expert participants emphasised that conducting a Data Protection Impact Assessment (DPIA) should be standard practice when generating synthetic data, particularly when the generation process involves accessing real personal data. Some organisations represented in the workshops have already implemented this approach, requiring DPIAs to be completed and approved before synthetic data generation proceeds. This aligns with the ICO's expectation that organisations will conduct thorough risk assessments.

The ICO's PETs guidance also specifically mentions differential privacy as a potential protection mechanism for synthetic data, noting that organisations:

> *"You could protect any records containing outliers from these types of linkage attacks with other information through suppression of outliers (data points with some uniquely identifying features); or differential privacy with synthetic data."*

However, the guidance also acknowledges important caveats:

> *"it may reduce the utility of the information and introduce a degree of unpredictability in the characteristics of the information," and emphasises that "if you do not have the required expertise, you should consult an external expert in setting an appropriate privacy budget when using differential privacy."*

These concerns were strongly echoed in the expert workshop, where differential privacy emerged as a contentious topic. Participants highlighted the significant practical difficulties of setting appropriate privacy budgets (epsilon values) when generating synthetic data.

An invited speaker presented recent research demonstrating that the privacy-utility trade-off in differential privacy is not linear, further complicating the already challenging task of selecting appropriate parameters. This non-linearity means that small changes in privacy budget settings can have disproportionate and unpredictable effects on both privacy protection and data utility, making it extremely difficult to find an optimal balance.

Many expert participants expressed scepticism about whether differential privacy is a beneficial mechanism for synthetic data generation in practice. Concerns centred on several key issues: the high degree of technical expertise required to implement it correctly; the lack of clear guidance on setting appropriate privacy budgets for different contexts and use cases; the degradation of utility when privacy budgets are set conservatively; and the difficulty of communicating and justifying differential privacy parameters to non-technical stakeholders such as information governance teams and data owners.

Several participants noted that even when differential privacy is correctly applied during generation, the resulting synthetic data must still undergo empirical privacy evaluation to confirm that it meets anonymisation standards. This reinforced the view that what ultimately matters is not the theoretical privacy guarantee provided by differential privacy, but rather the demonstrated privacy properties of the synthetic data itself as assessed through comprehensive disclosure risk evaluation.

Despite this existing regulatory guidance, however, data owner workshop participants expressed significant concern about the lack of detailed, practical guidance and established best practices for synthetic data generation and release. Many felt uncertain about how to interpret and apply the ICO's principles in specific contexts, particularly when assessing whether synthetic data meets anonymisation standards or determining appropriate risk mitigation measures for different fidelity levels and use cases. Participants emphasised that more concrete guidance, worked examples, and greater engagement with the ICO would be highly beneficial.

In response to these concerns, we established focused regulatory workshops bringing together key experts, information governance teams, legal professionals, data-holding organisations, and regulatory representatives to explore what is needed in terms of regulatory clarity and guidance for synthetic data. This work is ongoing, and we expect to publish a dedicated report in the coming months specifically addressing how synthetic data should be treated under GDPR, including practical frameworks for compliance and risk assessment.

Recommendation S.2

Carry out a Risk Assessment

Before generating or releasing synthetic data, organisations should conduct a comprehensive risk assessment. The assessment should evaluate any potential disclosure risks, consider the sensitivity of source data and characteristics of potential attackers, and account for how risks may evolve over time. Risk assessments should be documented, revisited as circumstances change, and used to inform appropriate mitigation measures and governance arrangements.

The ICO's PETs guidance also addresses the risk of perpetuating or amplifying bias. The guidance states:

> *"If you are generating synthetic data from personal information, any inherent biases in the information will be carried through. For example, if the underlying dataset is not representative of your population of interest (eg your customers), neither will the synthetic data which you generate from it. You should: ensure that you can detect and correct bias in the generation of synthetic data, and ensure that the synthetic data is representative; and consider whether you are using synthetic data to make decisions that have consequences for people (ie legal or health consequences). If so, you must assess and mitigate any bias in the information."*

These concerns came up in the public workshop when discussing high-fidelity synthetic data intended for analytical purposes.

Members of the public expressed significant unease about the possibility that synthetic data used to train AI models or inform policy decisions might encode and perpetuate existing inequalities or underrepresentation of minority groups. They emphasised that if synthetic data was to be used for consequential decision-making it was essential that the data accurately represented all affected populations, not just majority groups.

Notably, public participants also raised concerns about an additional source of bias: the privacy-preservation methods themselves. They worried that techniques used to protect privacy, such as removing or suppressing outliers and rare values, could disproportionately exclude or erase minority populations from synthetic datasets. Public participants felt that individuals from minority groups are more likely to appear as statistical outliers or rare cases in health and administrative datasets, whether due to rare conditions that affect certain ethnic groups, unique combinations of demographic characteristics, or underrepresentation in the source data itself. By removing these cases to reduce re-identification risk, synthetic data generation could inadvertently make minority populations invisible, effectively prioritising the privacy of the majority over the representation and needs of minorities.

This created a tension in the public's view: while they valued privacy protection and understood its importance, they were deeply uncomfortable with the idea that achieving privacy might come at the cost of excluding or marginalising already underrepresented groups. Several participants questioned whether this trade-off was ethically acceptable, particularly if the resulting synthetic data were used to train algorithms or inform policies that would affect those very populations who had been excluded from the data. They emphasised that any synthetic data used for consequential purposes must find ways to preserve both privacy and representativeness.

Synthetic Data for Medical Device Development
Beyond UK GDPR, specific regulatory considerations apply when synthetic data is used in the development of AI as a Medical Device (AIaMD). In July 2025, the MHRA and PHG Foundation published "Synthetic data for development of AI as a medical device (AIaMDs): Regulatory considerations," a guidance document developed by an expert working group to support manufacturers in justifying and assessing their use of synthetic data in regulatory submissions.

This guidance emphasises three overarching principles for regulatory acceptability of synthetic data: fidelity (how well synthetic data captures the statistical and structural properties of real-world data), representativeness (ensuring synthetic data reflects the full diversity of the target population), and transparency (clear explanation of the generation process, assumptions, and intended use). These principles underpin the MHRA's framework for when and how synthetic data should be used in medical device development.

The MHRA guidance adopts a risk-based approach, recognising that "the level of evidence and certainty required is generally proportional to risk: higher-risk devices may require stronger justification for use of synthetic data and higher levels of certainty regarding evidence on performance, whereas greater uncertainty may be tolerated for lower-risk devices." Importantly, the guidance states that "in the majority of cases, evidence for regulatory purposes should not be based purely on synthetic data, unless there is a very strong justification to do so."

To support manufacturers in providing such justification, the guidance categorises legitimate uses of synthetic data into three types. Technical justifications include dataset boosting to enhance model performance and address underrepresented populations (though the guidance stresses this must not perpetuate existing biases), benchmarking AI architectures, testing edge cases and rare scenarios not well-represented in real data, and quantifying uncertainty through simulation techniques. Ethical and regulatory justifications centre on facilitating access when real data is unavailable or restricted, and addressing representational gaps in real datasets. Lifecycle justifications include enabling pre-market evaluation of extended use cases and supporting post-market safety monitoring.

This categorisation demonstrates that synthetic data serves multiple distinct purposes in medical device development, from addressing data scarcity to enabling safer experimentation, but its appropriateness must be assessed in relation to both the specific use case and the risk profile of the device.

Perhaps most notably, the MHRA guidance takes a specific position on the tension between privacy protection and data utility when generating synthetic data for medical device development. The guidance states:

> *"in the context of AIaMD regulation, fidelity should be prioritised over the incorporation of privacy measures like censoring, suppression, aggregation or differential privacy during the synthetic data generation process. Privacy requirements should be fulfilled using alternative privacy approaches such as secure access to synthetic data by authorised users, so that fidelity is not compromised."*

This principle reflects a fundamentally different approach to managing the privacy-utility trade-off compared to many other synthetic data applications. Rather than embedding privacy protections within the synthetic data itself through techniques that degrade or limit the data, the MHRA framework proposes that privacy should be managed through governance and access controls around the synthetic data.

Recommendation S.3

Consider the Balance between Privacy, Fairness and Utility Carefully

Privacy-preserving techniques can inadvertently harm fairness by excluding minority populations who appear as outliers, while maintaining high fidelity for representativeness may increase privacy risk. Organisations must make context-dependent judgements about these trade-offs based on their specific use case, and document any implications.

2.1.2 Assessing Disclosure Risks in Synthetic Data

During the expert workshop, participants explored three main categories of disclosure risk that are often considered when evaluating the privacy of synthetic datasets: identity disclosure, membership disclosure, and attribute disclosure. These categories map closely onto the GDPR principles of singling out, linkability, and inference, providing a practical framework for assessing whether synthetic data is effectively anonymised.

Identity disclosure occurs when an individual can be directly identified from the synthetic dataset, typically through the presence of unique or quasi-identifying combinations of variables. This form of disclosure has traditionally been the primary focus of privacy assessments in synthetic data, with organisations commonly employing identity-based disclosure methods such as measuring the distance to the closest record (DCR) in the real dataset, or checking for exact matches between synthetic and real records.

Membership disclosure refers to the risk that an attacker can determine whether a specific individual's data was included in the original dataset used to generate the synthetic data. It exploits patterns or statistical properties in the synthetic data to infer the presence or absence of particular individuals in the source data.

Attribute disclosure occurs when sensitive information about an individual can be inferred from the synthetic data. For example, if the synthetic data reveals strong associations between certain demographic characteristics and sensitive health conditions, an attacker could use this information to make probabilistic inferences about individuals in the dataset known to share those characteristics.

Assessing Disclosure Risks in Synthetic Data

	Singling Out	Linkability	Inference
Traditional Definition	The ability to isolate a record, usually from unique combinations of attributes, that distinguish an individual from the other records.	The ability to link at least two records concerning the same individual.	The possibility to deduce, with significant probability, the value of an attribute given a set of other attribute values.
Definition in Context of Synthetic Data	Attacks typically target outliers or unique attribute combinations where synthetic generators are more likely to capture these rare/unique characteristics, making them susceptible to being replicated or appearing very similar to the original data.	Where an adversary connects a record in the synthetic dataset to a separate, external dataset by matching shared characteristics. A successful link reveals new sensitive information from the synthetic data (like income or health status) that was not present in the other data.	The ability to use a set of attribute values from an individual in the source data, to infer, with high confidence, sensitive unknown information about the same individual using the synthetic data.
Example	A dataset contains a row for a "19-year-old female, widowed, occupation: MP." While "19-year-old" and "female" are common, the combination of being a teenager, a widow, and an MP is statistically unique. An adversary viewing this dataset can instantly isolate this record as an outlier. If a real person fits this rare description, the attacker knows this specific synthetic row represents them.	A record lists "Male, Age 40, Admission Date: June 1st, Diagnosis: Gambling Addiction." A social media post from June 2nd says: "Checked into the Wellness Centre yesterday to focus on my health." The adversary links the Admission Date and Age from the synthetic data to the public post. By making this link, the adversary learns sensitive information that the reason is specifically for Gambling Addiction.	The synthetic data shows a strong correlation: "Every individual in the 'Night Shift' category who visits 'Clinic A' is being treated for Sleep Apnoea." The adversary knows their colleague, Dave, is in the source dataset, is on the Night Shift, and was seen at Clinic A. Even without finding a specific row for Dave, the adversary uses the rule learned from the synthetic data to infer that Dave has Sleep Apnoea.
Evaluation	Identity disclosure assessments to ensure outliers or unique records aren't replicated in the synthetic data.	Assessments such as nearest neighbours to evaluate similarity between source and synthetic datasets and whether they can be successfully matched.	Typically assessed through membership and attribute inference disclosure tests.

For low-fidelity synthetic data (such as L1 and L2 synthetic datasets), we found that organisations typically employ identity-based disclosure metrics primarily as a sanity check. Because these datasets preserve minimal information about the underlying data and do not retain complex relationships between variables, the disclosure risk is inherently low. However, for high-fidelity synthetic data, which preserves more complex statistical relationships and distributions across multiple variables, identity-based disclosure measures are insufficient on their own. The increased realism and analytical utility of high-fidelity datasets means they may inadvertently encode patterns that enable more sophisticated attacks. In these cases, participants strongly emphasised that membership and attribute inference attacks must be explicitly evaluated.

An important insight from the workshop came from a presentation by the Imperial Computational Privacy Group, which demonstrated that many synthetic datasets identified as private using traditional identity disclosure measures failed to withstand membership inference attacks. This finding highlighted a significant gap in current evaluation practices: while distance-based metrics such as DCR are widely used and provide some assurance against direct re-identification, they do not adequately capture the full spectrum of privacy risks, particularly those related to membership and attribute inference.

Recommendation S.4

Assess Disclosure Risks Appropriately against Anonymisation Principles

Disclosure risk assessment should be proportionate to the fidelity level of synthetic data and the sensitivity of source data. For low-fidelity synthetic data (L1 and L2) created from public information, basic identity disclosure checks are typically sufficient. For high-fidelity synthetic data, or synthetic data created from private information, comprehensive evaluation must include evaluations such as membership inference testing and attribute disclosure analysis to address all three GDPR anonymisation principles: singling out, linkability, and inference.

2.2 Characteristics that Influence Disclosure Risk

The disclosure risk of synthetic data is not determined solely by the evaluation methods used; it is also fundamentally shaped by the characteristics of the source data from which it is derived, the method used to create the synthetic data, and the way it is documented. Across all workshops, participants highlighted several key factors that influence the inherent risk profile of synthetic datasets.

2.2.1 Source Data

Sensitivity of the source data was identified as a primary consideration. Datasets containing highly sensitive information, such as detailed health records, criminal justice data, or genomic information, inherently carry greater disclosure risk and reputational consequences if privacy is compromised. The presence of outliers or rare values in the source data also significantly increases risk, as these cases are more likely to be identifiable, even in synthetic form. Additionally, the identifiability status of the source data - whether it is fully identifiable, pseudonymised, or already anonymised - directly impacts the level of privacy protection that must be built into the synthetic generation process.

An important distinction emerged across all three workshops between publicly-sourced synthetic data and privately-sourced synthetic data.

Publicly-Sourced Synthetic Data

is generated using only publicly available metadata or summary statistics that have already undergone standard disclosure control processes. Because this approach does not involve direct access to individual-level records, experts reported that it carries minimal disclosure risk. Data owners expressed significantly greater comfort with this type of synthetic data, viewing it as a lower-risk option that still provides meaningful utility for tasks such as data discovery, code development, and training. The public also expressed a strong preference for publicly-sourced synthetic data, viewing it as more transparent and less of a privacy risk.

Privately-Sourced Synthetic Data

is generated directly from real datasets held within secure environments, using methods that access and learn from individual-level records to preserve more complex statistical relationships. While this approach enables the creation of higher-fidelity synthetic data with greater analytical utility, it also introduces greater privacy risk and may require more stringent governance and evaluation.

This distinction has significant implications for how synthetic data should be governed and communicated. Publicly-sourced synthetic data, because of its minimal disclosure risk, can often be released more freely and with lighter governance controls. Privately-sourced synthetic data, while valuable for more advanced analytical purposes, requires robust privacy evaluation, transparent documentation, and carefully calibrated access controls to ensure it is used responsibly.

Data custodians expressed markedly higher levels of confidence and comfort when discussing publicly-sourced synthetic data compared to privately-sourced alternatives. This confidence stemmed primarily from the fact that publicly-sourced synthetic data is built entirely from information that has already been approved for public release through standard statistical disclosure control procedures. Because the underlying inputs such as metadata, value ranges, and aggregate summary statistics, have already passed through rigorous disclosure checks and are deemed safe for publication, data custodians felt that generating synthetic data from these sources introduced minimal additional disclosure risk beyond what already exists in the published statistics themselves.

Several data custodians explicitly stated that they viewed sharing publicly-sourced synthetic data as carrying essentially the same level of disclosure risk as sharing the aggregate summary statistics on which it is based.

This perspective was particularly important in building institutional confidence: if the component parts (the aggregate statistics) were already considered safe to release, then synthetic data generated from those components could reasonably be expected to maintain that same level of safety. This logic provided a clear and defensible rationale for approving the release of publicly-sourced synthetic datasets, as it aligned closely with established practices for publishing statistical information that organisations were already familiar with and comfortable implementing.

However, data custodians were careful to distinguish between disclosure risk and reputational risk. While they felt confident that publicly-sourced synthetic data carried very low disclosure risk from a privacy and GDPR compliance perspective, they acknowledged that reputational risk remained somewhat a concern.

Fig 2.1 How comfortable would you be releasing low-fidelity synthetic data outside of a TRE?

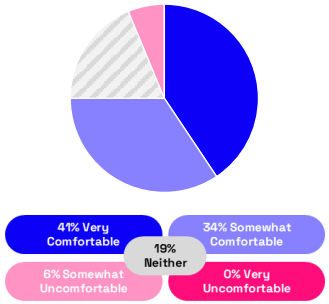


Fig 2.2 How much of a privacy risk do you consider L1 synthetic data to have?

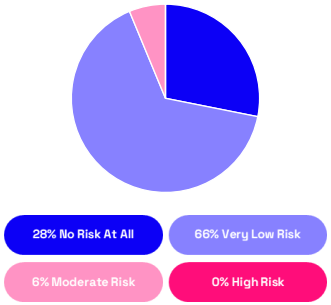


Fig 2.3 How much of a privacy risk do you consider L2 synthetic data to have?

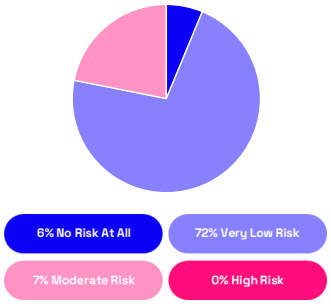
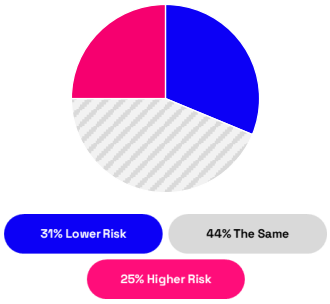


Fig 2.4 How does the risk compare to standard release of summary statistics?



2.2.2 Generation Method

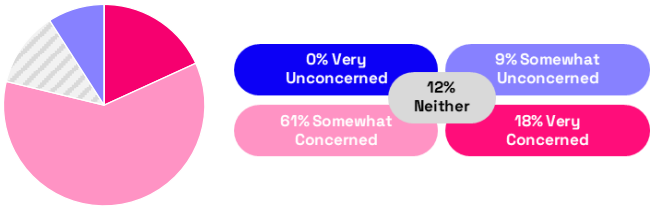
Transparency in how synthetic data is generated emerged as a key factor influencing perceptions of disclosure risk and trustworthiness. Across the expert and data owner workshops, participants consistently emphasised that transparent, interpretable generation methods are critical for building confidence in the safety of synthetic datasets, particularly when engaging data custodians who must make decisions about release and use.

From the expert workshop, it was clear that most organisations are gravitating toward more transparent generation methods such as Bayesian networks or statistical modelling approaches, rather than AI-based techniques like Generative Adversarial Networks (GANs), which are often perceived as opaque and difficult to explain. A presentation from the Clinical Practice Research Datalink (CPRD) at the MHRA showcased their use of Bayesian networks to create high-fidelity synthetic data, and many participants expressed comfort with this approach due to its clarity and traceability. Importantly, the CPRD source data was already fully anonymised, which further reduced concerns about disclosure risk.

Similarly, a speaker from the Alan Turing Institute described their method for generating synthetic data for University College London Hospitals (UCLH), which was built entirely around transparent generation using only metadata and summary statistics. This approach ensured that no individual-level data was accessed during the synthesis process, reinforcing privacy by design and making it significantly easier to gain approval from governance bodies.

Data owners reinforced the importance of transparency, stating that they felt far more comfortable with methods that could be clearly explained and understood. Complex, black-box AI models raised concerns about hidden risks, lack of auditability, and the difficulty of assessing whether privacy had been adequately protected. In contrast, transparent methods enabled data owners to scrutinise the generation process, understand what information had been retained or excluded, and make informed decisions about whether the synthetic data was safe to release.

Fig 2.5 How concerned would you be about the privacy risk of high-fidelity synthetic data generated by AI models?



This emphasis on transparency highlights the principle of privacy-by-design: synthetic data generation should incorporate privacy protections at every stage, from method selection through to evaluation and release. By prioritising interpretability and traceability, organisations can ensure that synthetic data is not only technically safe but also perceived as trustworthy by the stakeholders responsible for its governance.

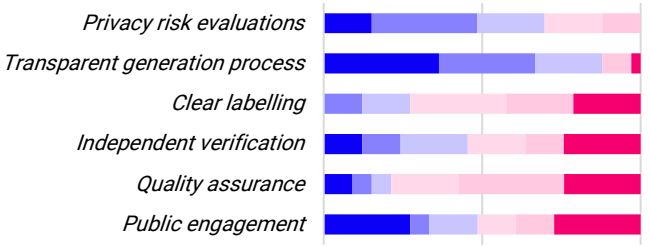
During the expert workshop, participants also discussed the potential role of differential privacy as a privacy-enhancing technique in synthetic data generation. Differential privacy is a mathematical framework that provides formal privacy guarantees by introducing carefully calibrated noise into data or statistical outputs, making it theoretically possible to quantify the maximum privacy loss from any individual's participation in a dataset.

While differential privacy offers compelling theoretical advantages, many workshop participants expressed significant reservations about relying on it as a primary or sole privacy guarantee for synthetic data. A recurring concern was that differential privacy is often poorly understood and incorrectly implemented, even by technical practitioners.

Participants noted that selecting appropriate privacy parameters (such as the epsilon value, which controls the privacy-utility trade-off) requires deep expertise and careful consideration of the specific context and use case. In practice, differential privacy can be applied too loosely, resulting in insufficient privacy protection, or too strictly, resulting in synthetic data with severely degraded utility that fails to serve its intended purpose.

Furthermore, participants highlighted that even when differential privacy is correctly applied, the output must still be evaluated to confirm that it meets anonymisation standards and does not introduce unexpected disclosure risks. As a result, experts emphasised that what truly matters are the post-generation evaluations of the synthetic data itself. Regardless of whether differential privacy or other privacy-enhancing techniques are employed during generation, the synthetic dataset must undergo empirical assessment against disclosure risks. These evaluations provide concrete, interpretable evidence of whether the data is safe to release and whether it meets GDPR anonymisation standards. Participants felt that relying solely on claims of differential privacy, without robust evaluation, risks creating a false sense of security and may not satisfy the transparency and accountability expectations of information governance teams, data owners, or regulators.

Fig 2.6 What would increase your trust in synthetic data?



2.2.3 Documentation

Another consideration raised by experts was the balance between transparency and security in documenting generation methods. While transparency is essential for building trust and enabling scrutiny, experts cautioned that releasing overly detailed information about the generation process could inadvertently provide attackers with insights that could be exploited in re-identification attempts.

Documentation should therefore be sufficiently detailed to enable understanding and validation, but should avoid disclosing information that could compromise privacy, such as exact parameter settings, random seeds, or detailed information about the presence of specific rare cases or outliers. The goal is to provide enough transparency to support informed governance decisions and user understanding, while not creating a roadmap that could be used for adversarial purposes.

Recommendation S.5

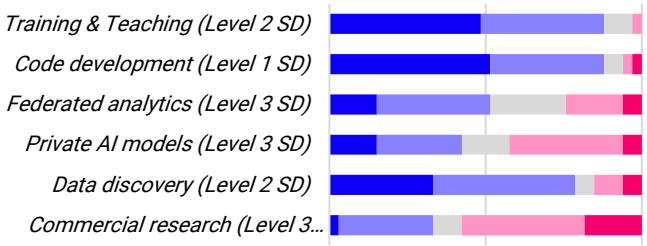
Prioritise Transparent Generation Methods and Use Only the Minimum Information Necessary

When generating synthetic data, organisations should consider the principle of data minimisation with a commitment to transparency and interpretability in generation methods. This approach serves to minimise disclosure risk, simplify governance processes, and maximise stakeholder confidence and trust.

Synthetic data generation should use only the minimum amount of information from the source data that is necessary to achieve the intended purpose and support the defined use case. This principle of data minimisation serves both to reduce disclosure risk and to simplify governance and approval processes.

Organisations should also prioritise generation methods that are transparent, interpretable, and easily explainable to non-technical stakeholders. Workshop participants consistently emphasised that transparent approaches build significantly greater confidence among data owners and information governance teams than methods such as opaque, black-box AI techniques. Transparent generation methods enable data custodians to scrutinise the generation process, understand what information has been retained or excluded, and make informed decisions about whether the synthetic data is safe to release.

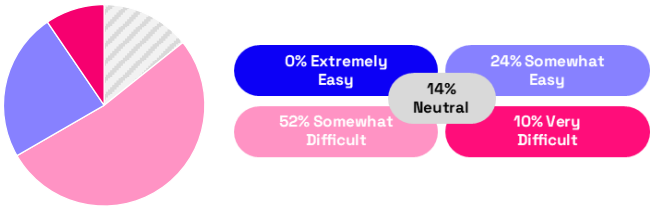
Fig 2.7 How comfortable would you feel about releasing synthetic data for the following purposes?



2.3 Challenges in Communicating and Interpreting Disclosure Risk

Even when robust privacy evaluation methods are applied, effectively communicating the results to data owners and governance teams remains a significant challenge. Expert participants reported that current evaluation metrics are difficult to interpret, making it hard to determine appropriate thresholds for acceptable risk and even harder to explain findings in a way that enables informed decision-making.

Fig 2.8 How easy do you find privacy metrics to interpret?

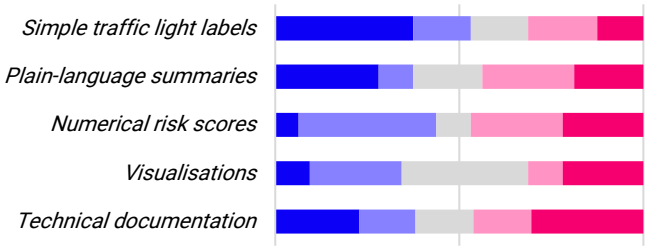


Metrics such as distance to closest record (DCR) and membership inference attack success rates provide quantitative measures of disclosure risk, but they often lack intuitive meaning for non-experts. Data owners didn't feel comfortable being asked to make decisions about synthetic data release based on metrics they did not fully understand, with little guidance on how to weigh different types of risk or set acceptable thresholds.

To address this, data owners suggested that they would feel more confident making decisions if risk evaluations were presented using clear, traffic-light-based risk categories (e.g., low, medium, high), accompanied by plain-language explanations and visual comparisons.

This highlights the need for synthetic data practitioners to invest not only in technical evaluation methods but also in communication strategies that translate complex privacy assessments into accessible, actionable insights for decision-makers. Clear documentation, intuitive risk categorisation, and transparent explanations of what evaluation results mean in practice are essential for ensuring that data owners can confidently approve and govern synthetic data releases.

Fig 2.9 What format would help you to best understand the privacy risk of a synthetic dataset?



2.4 Concerns about Reputational Risk of Synthetic Data

Despite the potential value of high-fidelity synthetic data, many organisations and data owners have struggled to gain approval for its creation and release. The workshops revealed that most organisations have, to date, focused primarily on generating and releasing low-fidelity synthetic data because it carries minimal disclosure risk and is significantly easier to explain and justify to decision makers.

Presentations from the Department for Education on their low-fidelity synthetic data for the Longitudinal Education Outcomes (LEO) dataset, Research Data Scotland on their low-fidelity offerings, and the Ministry of Justice on synthetic data made available through the UK Data Service all illustrated this trend. In each case, organisations felt comfortable releasing structural or univariate synthetic data but hesitated to progress to higher-fidelity datasets due to concerns about increased risk and the challenges of securing approval.

A key barrier identified by participants was the difficulty teams face in understanding and assessing the risks associated with high-fidelity synthetic data. Unlike low-fidelity datasets, which clearly preserve minimal information and carry negligible disclosure risk, high-fidelity synthetic data involves nuanced trade-offs between utility and privacy that are not always straightforward to evaluate or communicate. This uncertainty makes it harder for decision makers to confidently approve release, particularly in the absence of clear precedents or established frameworks.

Importantly, the workshops identified that there are actually two distinct types of risk preventing organisations from progressing into high-fidelity synthetic data: disclosure risk and reputational risk.

Disclosure risk refers to the technical probability that synthetic data could enable re-identification of individuals, inference of sensitive attributes, or determination of whether someone's data was included in the source dataset, assessed through quantitative privacy evaluation methods against GDPR anonymisation standards.

Reputational risk refers to the potential for institutional harm arising from public perception, media coverage, or stakeholder misunderstanding of synthetic data releases such as patients mistakenly believing they've identified themselves, even when technical disclosure risk is demonstrably low.

Interestingly, many participants, particularly experts, felt that with appropriate generation and evaluation methods, the actual disclosure risk of high-fidelity synthetic data can be well-controlled. However, they acknowledged that reputational risk remains a significant barrier. Data owners expressed concerns that members of the public might mistakenly believe they had identified themselves in a synthetic dataset, leading to accusations of data leaks and damaging media attention, even if no actual privacy breach had occurred. This concern was particularly acute for organisations handling sensitive data related to criminal justice, health, or vulnerable populations, where public trust is paramount.

One expert presenter highlighted the distinction between a motivated intruder (an attacker actively attempting to re-identify individuals) and a naive intruder (someone who mistakenly believes they are in the dataset). Data owners noted that members of the public or inexperienced researchers mistakenly believing they have identified a person even though that's not the case, represented one of their primary concerns: even if the risk of actual re-identification is low, the perception of a breach can cause significant reputational harm.

A compounding factor is the lack of real-world use cases demonstrating the value and safety of high-fidelity synthetic data in the UK context. Without established examples of successful deployment, organisations are reluctant to take the risk of being early adopters. Participants emphasised that building trust and acceptance will require demonstrating tangible benefits through real-world case studies, supported by transparent evaluation and governance frameworks that clearly show how privacy and responsible use are being ensured.

Some participants in the expert workshop expressed frustration at what they perceived as a paradox in how synthetic data is treated compared to real anonymised data. They pointed out that many organisations routinely release real anonymised health datasets with minimal controls or safeguards, often making them publicly available with little more than basic registration requirements. Yet organisations demonstrate extreme caution and hesitancy when considering the release of synthetic data. This apparent inconsistency seemed illogical from a purely technical risk perspective, particularly when synthetic data had been explicitly designed with privacy protections and had undergone comprehensive evaluation to demonstrate its safety. However, other participants offered an important clarification: the heightened caution around synthetic data is not primarily driven by concerns about disclosure risk, but rather by reputational risk and the potential for misunderstanding and misuse. Unlike established practices for releasing anonymised data, which benefit from years of precedent, regulatory guidance, and public familiarity, synthetic data remains relatively new and poorly understood by many stakeholders. There is concern that the public, might misinterpret the release of synthetic data as evidence of careless data handling or that a privacy breach has occurred. Furthermore, the risk of misuse, such as researchers inappropriately using synthetic data for analytical purposes it was not designed to support and subsequently publishing misleading findings, carries consequences that extend beyond individual projects to potentially damage the credibility of the data owner and undermine trust in data sharing more broadly. This recognition highlights that governance of synthetic data must address not only technical safety but also communication, education, and mechanisms to prevent and respond to misunderstanding and misuse.

Recommendation S.6

Identify Whether Reputational Risk is the Primary Barrier rather than Disclosure Risk

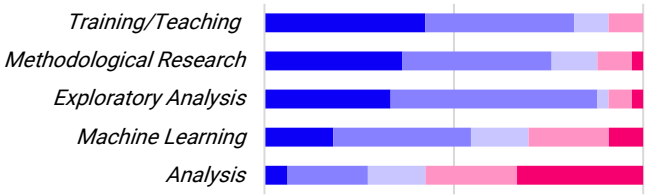
Organisations considering the generation and release of synthetic data should first clearly identify whether the primary barrier to release is technical disclosure risk or reputational risk, as these require fundamentally different mitigation strategies. Where rigorous privacy evaluation demonstrates that synthetic data meets GDPR anonymisation standards and carries quantifiably low disclosure risk, yet approval remains difficult to secure, the barrier is likely reputational rather than technical. In such cases, organisations should focus their efforts on building confidence, managing perceptions, and preventing misunderstanding rather than investing solely in additional technical safeguards that may not address the underlying concerns.

Data owners also expressed varying levels of comfort with different synthetic data use cases, with their acceptance closely tied to both the perceived risk of the application and their understanding of how the synthetic data would be used. There was broad acceptance and enthusiasm for synthetic data supporting data discovery, teaching, and code development, where the functional purpose is clear, the risk of misuse or misinterpretation is relatively low, and the benefits in terms of accelerating research and building skills are immediately apparent. These preparatory use cases were seen as straightforward applications where synthetic data could deliver genuine value without introducing significant concerns about validity or misuse, particularly when lower-fidelity datasets were employed.

However, data owners were notably more cautious about high-fidelity synthetic data being used for direct analysis or developing privacy-preserving AI models. In these contexts, concerns centred on whether synthetic data could adequately capture the complexity and nuance of real data, whether findings might be over-interpreted or presented as definitive evidence without appropriate validation, and whether the development of AI models on synthetic data might introduce unexpected biases or fail to generalise.

This heightened caution reflected data owners' awareness of their responsibility not only to protect privacy but also to ensure that data released under their stewardship is used in ways that produce valid, trustworthy research outcomes that do not mislead or cause harm. Some data owners explicitly stated that they would only feel comfortable approving high-fidelity synthetic data for analytical purposes if there were clear requirements for findings to be validated against real data before publication, and if the synthetic data's limitations were prominently disclosed in any resulting outputs.

Fig 2.10 How comfortable would you feel about high-fidelity synthetic data being used for the following purposes?



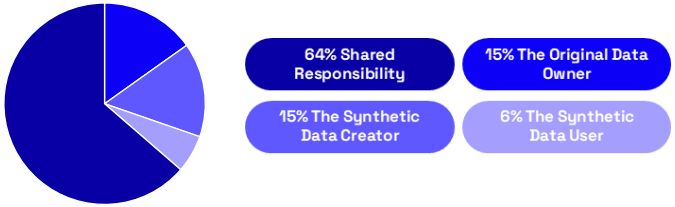
2.5 Accountability, Responsibility, and Data Ownership

The question of who is responsible for the safety and appropriate use of synthetic data emerged as a significant concern, particularly among data owners. Many data owners expressed unease at the prospect of taking ownership of a synthetic dataset they had not created and were not involved in generating. They questioned whether they should be held accountable for privacy risks or misuse of synthetic data that had been produced by external researchers or third-party organisations without their input or oversight.

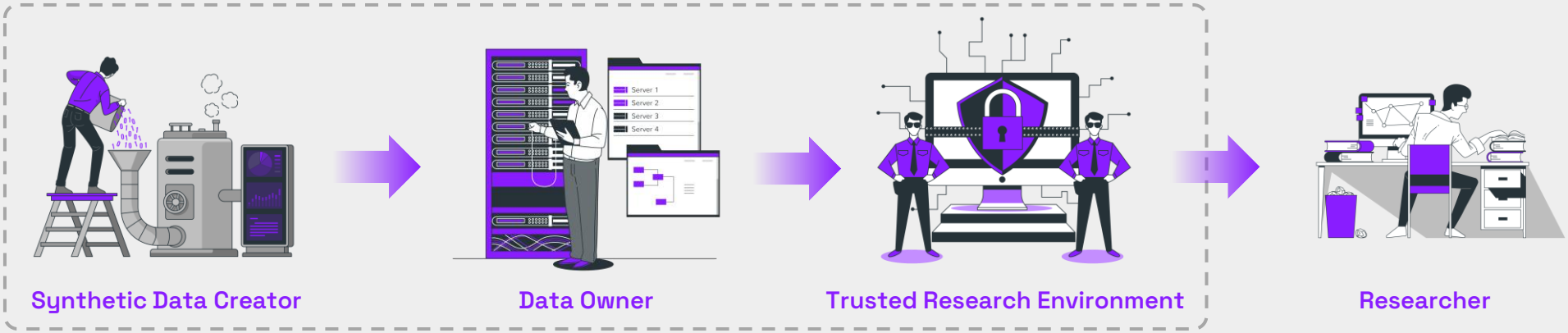
Conversely, there was a recognition that responsibility cannot rest solely with the data owner, particularly when synthetic data generation involves specialised technical expertise that data owners may not possess. Participants ultimately converged on the view that responsibility should be shared among three key parties: the data owner, the synthetic data creator, and the Trusted Research Environment (TRE) or host organisation facilitating access.

Under this shared responsibility model, the data owner retains oversight of the source data and must approve the release of synthetic datasets, ensuring they align with ethical and legal obligations. The synthetic data creator bears responsibility for employing robust generation and evaluation methods, transparently documenting the process, and ensuring the dataset is fit-for-purpose and privacy-preserving. The TRE or host organisation is responsible for implementing appropriate governance structures, facilitating safe access, and ensuring that users are educated and held accountable for responsible use.

Fig 2.11 Who should be held responsible for the safety of synthetic data?



These three roles may all be separate, or some/all may be the same.



Synthetic Data Creator

Responsibility to ensure synthetic data is created that protects privacy, evaluated appropriately, and sufficiently documents processes, limitations and risks in a clear accessible way.

Data Owner

Responsibility to approve the release of synthetic data given the information provided by the creator and select suitable safeguards comfortable to them and their risk appetite.

Trusted Research Environment

Responsibility to facilitate safe and responsible access to the synthetic data, ensuring users are educated and implementing appropriate safeguards desired by the data owner.

Researcher

Responsibility of the end user to understand the appropriate uses and limitations of the synthetic data, avoiding misuse and handling it in a responsible way

3

Transparent

Principle 3: Synthetic data generation processes, methods, assumptions, limitations, and evaluation results must be clearly documented and communicated to enable informed decision-making across all stakeholder groups. Documentation must serve multiple audiences, providing technical detail for experts while offering accessible summaries for non-technical stakeholders including data custodians and the public.

03. Transparent

Transparency is critical for building trust in synthetic data. Stakeholders, from data owners to end users, must be able to understand how the data was created, what it represents, and where its limitations lie. Transparency ensures that synthetic data is not treated as a "black box" and that decisions made using it are grounded in informed understanding. Transparency also helps foster trust and increase the acceptability of synthetic data among data owners, TREs, and end users.

Summary of Key Recommendations:

01. Ensure Synthetic Data is Clearly Documented and Labelled

Every synthetic dataset must be accompanied by comprehensive, well-structured documentation and clear labelling that identifies it as synthetic. This documentation is essential for building trust, preventing misuse, enabling informed decision-making across diverse stakeholder groups, and demonstrating the responsible data stewardship to maintain public confidence.

02. Clearly Define Intended Use Cases and Disclose Limitations

Documentation must explicitly state what the synthetic data was designed to support and what it should not be used for, accompanied by concrete examples. Limitations should be presented prominently using clear, non-technical language, enabling researchers to assess fitness-for-purpose, avoid wasted effort on inappropriate applications, and ensure methods will transfer successfully to real data environments.

03. Provide Clear Visual Comparisons Between Synthetic and Real Data

Include accessible visual comparisons that enable researchers to quickly and intuitively assess similarity between synthetic and real data. While quantitative similarity metrics provide precise assessments, visual representations are essential for researchers to rapidly develop intuitive understanding of data quality and fitness-for-purpose without requiring deep expertise.

04. Document Data Provenance and Any Preprocessing Steps

Provide information about the source data's origin and all preprocessing, cleaning, or filtering applied before synthetic generation, enabling researchers to understand exactly what state the data was in and how it was transformed. This provides essential context for interpreting how the synthetic dataset relates to the real data researchers may eventually access.

05. Provide Accessible Summaries and Contextualised Risk Communication

Adopt a multi-layered documentation approach combining technical detail with accessible communication tailored to different stakeholders: lay explanations of generation methods for data owners and governance teams, interpretable risk categories with real-world context rather than abstract metrics, and technical appendices for expert review. This ensures data custodians can confidently defend release decisions, make informed approvals, and the public can maintain trust, while still providing detail necessary for approval.

06. Explicitly Document All Assumptions in Generation and Evaluation

Transparently disclose all assumptions made during generation (e.g., distributional assumptions, constraints, modelling limitations) and evaluation (e.g., privacy threats considered, attacker knowledge assumed, success/failure thresholds). Clear documentation of assumptions prevents overconfidence in results, enables users to assess whether assumptions align with their use case, and supports nuanced interpretation of synthetic data capabilities and limitations.

03. Transparent

Transparency emerged as a key factor across all workshops, consistently identified as essential for building trust, enabling informed decision-making, and ensuring the responsible use of synthetic data.

Without transparency, synthetic data risks becoming a "black box" that stakeholders are reluctant to trust, approve, or use. Researchers need transparency to assess whether synthetic data is fit for their intended purpose and to understand any limitations that might affect their work. Data owners and governance teams need transparency to evaluate privacy risks, make informed approval decisions, and take accountability for releases. Members of the public need transparency to understand how their data is being used and to maintain confidence that privacy protections are being respected.

The workshops revealed that transparency requirements differ across stakeholder groups, reflecting their distinct roles and concerns. Researchers prioritise understanding the quality, utility, and practical limitations of synthetic data. Data owners require clear explanations of generation methods and disclosure risks, presented in accessible language that enables confident decision-making. Experts emphasise the importance of documenting technical assumptions, evaluation methodologies, and known limitations to support rigorous assessment and reproducibility. Despite these differences, all groups agreed that comprehensive, well-structured documentation is non-negotiable for synthetic data to be trusted and used responsibly.

This section explores transparency in synthetic data, examining what different stakeholders need to know, how information should be communicated, and what challenges organisations face in achieving meaningful transparency. It addresses the specific documentation requirements articulated by researchers, data owners, and experts; the importance of clear communication about generation methods, evaluation approaches, and limitations; the need for transparency to be tailored to both technical and non-technical audiences; and the critical role of transparency in managing both disclosure risk and reputational risk.

3.1 Why Transparency Matters

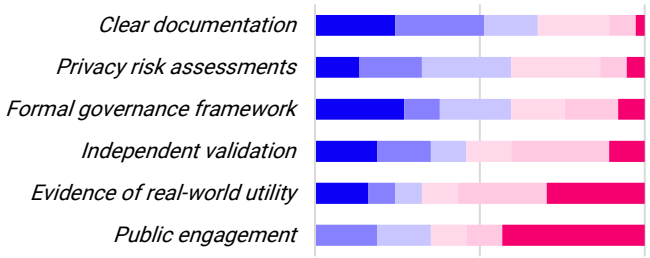
3.1.1 Building Trust and Confidence

Across all workshops, participants consistently stated that they would not trust synthetic data without comprehensive documentation explaining how it was created, evaluated, and validated. This requirement was not simply about technical detail, it was about confidence. Researchers needed to know they could rely on synthetic data for their intended tasks without introducing unintended biases or errors into their work. Data owners needed assurance that releasing synthetic data would not compromise privacy or damage their organisation's reputation. Members of the public needed to understand that their data was being handled responsibly and that synthetic data represented a genuine privacy protection.

The absence of transparency creates uncertainty, which in turn creates risk aversion. When stakeholders do not fully understand how synthetic data was generated or what guarantees exist around its privacy and utility, they default to caution, either refusing to approve releases, declining to use the data, or raising concerns about its legitimacy.

Transparency addresses this by making the generation process visible, the evaluations interpretable, and the limitations explicit. In doing so, it enables informed decision-making and reduces the perceived risk associated with synthetic data.

Fig 3.1 What would increase your trust in synthetic data?



3.1.2 Enabling Informed Use and Preventing Misuse

Transparency is also essential for ensuring that synthetic data is used appropriately. Without clear guidance on what synthetic data is suitable for and what it should not be used for, there is a significant risk of misapplication. Researchers may attempt to use low-fidelity synthetic data for analytical tasks that require high-fidelity datasets, leading to invalid or misleading results.

From the data owner workshop, misuse emerged as one of the most significant concerns surrounding synthetic data release, particularly the risk that users might present findings derived from synthetic data as if they were real-world results when the synthetic data wasn't intended for that. Data owners expressed anxiety about scenarios where low-fidelity synthetic data, created solely for code development or exploratory purposes, might be inappropriately used to generate analytical findings that are then published, cited, or used to inform policy decisions. Such misuse could lead to fundamentally flawed conclusions being treated as evidence-based insights, with potentially serious consequences.

This concern was particularly problematic because the consequences of misuse extend beyond the individual researcher to affect the reputation and trustworthiness of the data owner and the credibility of the source dataset itself. If synthetic data were to be misused in high-profile research that subsequently proved invalid or misleading, the resulting media attention and loss of public confidence could damage not only the perception of synthetic data but also broader efforts to enable responsible data sharing and research. Data owners emphasised that even a single well-publicised case of misuse could undermine years of work building trust with patients, the public, and regulatory bodies.

Documentation that clearly defines the intended use cases, explains the generation method and its implications for utility and privacy, and explicitly states the limitations of the synthetic data ensures that users can make informed choices. This is particularly critical for high-fidelity synthetic data, where the balance between utility and privacy is more complex and where inappropriate use could have serious consequences.

Data owners stressed that preventing misuse requires more than just stating limitations in technical documentation; it requires making those limitations impossible to ignore or misunderstand. Suggestions included prominently displaying warnings or disclaimers with synthetic datasets, requiring users to acknowledge specific limitations before gaining access, and clearly labelling all synthetic data files and variables with indicators that they are synthetic embedded in filenames, variable names, or metadata.

The risk of misuse varies significantly depending on the fidelity level and intended purpose of the synthetic data. Low-fidelity synthetic data, designed for code development, training, or data discovery, carries a higher risk of being inappropriately used because it is often more accessible and may be assumed by non-expert users to usable for analytical work. Data owners felt strongly that documentation for these datasets must include explicit, prominent statements that analytical findings derived from them are not valid and should never be presented as real-world evidence.

For high-fidelity synthetic data that is explicitly designed to support certain analytical tasks, such as hypothesis generation or algorithm development, the risk profile is different. Here, the concern is not that the data will be used for analysis but that findings will be over-interpreted or presented without appropriate caveats. Data owners emphasised that documentation for high-fidelity synthetic data must clearly communicate that while it may be suitable for preliminary analysis or methodological development, any findings must be validated against real data before being published or used to inform decisions. Some participants suggested requiring explicit statements in any outputs or publications acknowledging the use of synthetic data and confirming that findings have (or have not) been validated against real data.

To further mitigate the risk of misuse, many data owners supported the use of formal user agreements or terms of use that explicitly prohibit certain applications and require users to acknowledge their understanding of the synthetic data's limitations (more on this in the Accessible section). These agreements serve both as a deterrent to misuse and as a mechanism for accountability if misuse does occur.

However, participants recognised that agreements alone are not sufficient; they must be paired with clear, accessible documentation that genuinely enables users to understand what they are agreeing to and why the limitations matter.

Ultimately, preventing misuse requires a combination of clear communication, appropriate access controls, and ongoing monitoring and engagement with users.

Recommendation T.1

Ensure Synthetic Data is Clearly Documented and Labelled

To ensure synthetic data is trusted, used appropriately, and governed responsibly, comprehensive documentation must be provided with every synthetic dataset. By providing comprehensive, well-structured documentation, organisations can build trust in synthetic data, prevent misuse, enable informed decision-making, and demonstrate the responsible stewardship necessary to maintain public confidence and acceptance.

3.2 Researcher Perspective: Understanding Quality, Utility, and Limitations

For researchers, documentation requirements centred on practical usability: they need to understand whether synthetic data will work for their specific tasks, what differences they should expect compared to real data, and what adjustments they may need to make to their analytical workflows. Researchers emphasised that their primary concerns are quality, utility, and fitness-for-purpose rather than the technical details of generation algorithms or privacy. They need to know if the synthetic data accurately reflects the structure and characteristics of the real data, including its imperfections; whether it preserves the statistical properties necessary for their intended use case; and what preprocessing or filtering was applied that might affect how their code or analysis performs when eventually applied to real data.

Transparency, from the researcher perspective, is fundamentally about enabling confident, efficient work: ensuring they can trust the synthetic data for appropriate tasks, avoid wasting time on unsuitable applications, and develop methods that will transfer successfully to the real data environment.

3.2.1 Defining Fidelity and Intended Use Cases

Participants emphasised the need for synthetic datasets to be accompanied by a clear statement of their fidelity level, using a consistent classification system. This classification provides researchers with an immediate understanding of what the synthetic data preserves and what it does not, enabling them to assess whether it is appropriate for their needs.

Without a clearly defined fidelity level, researchers reported that they would struggle to determine whether synthetic data was fit for their intended purpose. A dataset described simply as "synthetic" provides insufficient information: does it preserve only structure, or does it also maintain statistical distributions? Are relationships between variables retained, or is each variable independent? These questions fundamentally shape whether the data can support a given research task, and researchers should not have to infer the answers through trial and error.

Researchers strongly advocated for the adoption of standardised fidelity classifications to create consistency across different synthetic datasets and organisations. When fidelity levels are clearly defined and consistently applied, researchers can develop expectations about what different levels mean in practice, streamlining their ability to select appropriate datasets and plan their work accordingly.

Equally important was the need for documentation to explicitly state what the synthetic data is suitable for and what it should not be used for. This clarity prevents inappropriate use and helps researchers understand the boundaries within which the synthetic data can be reliably employed.

Researchers also expressed a desire for use case examples to be included in documentation, illustrating how the synthetic data has been or could be used in practice.

These examples provide concrete guidance and help users understand the practical applications and boundaries of the dataset.

By clearly defining fidelity levels and documenting both appropriate and inappropriate use cases, organisations enable researchers to make informed decisions, reduce the risk of misuse, and ensure that synthetic data delivers value in contexts where it is genuinely fit-for-purpose. This transparency not only protects the integrity of research but also safeguards the reputation of data owners and maintains trust in synthetic data as a legitimate research tool.

Beyond defining appropriate and inappropriate uses, researchers stressed the critical importance of clearly disclosing any specific limitations or considerations that might impact their work. Researchers were worried about potentially wasting time using synthetic data where they only discover the limitations after investing significant time, such as finding that certain variables had been excluded, rare categories suppressed, temporal patterns disrupted, or specific subgroups underrepresented.

Documentation should explicitly identify variables or relationships that were intentionally modified or excluded, known biases introduced during generation, temporal or sequential constraints, handling of missing data, subgroup representation issues, and any logical rules or constraints applied. Understanding these limitations is essential not only for avoiding wasted effort but also for ensuring that methods and code developed using synthetic data will transfer successfully to the real data environment. If synthetic data has been changed in ways the real data has not, researchers need to know so they can build appropriate error handling and validation checks into their workflows.

For high-fidelity synthetic data intended for analytical tasks, transparent disclosure of limitations becomes even more important: researchers must understand how well relationships have been preserved, whether certain patterns or interactions may be artefacts of the generation process rather than genuine features of the real data, and what implications they have for the validity and generalisability of findings.

Without this transparency, researchers risk drawing conclusions from synthetic data that would not hold when validated against real data, potentially wasting time on analytical approaches that appear promising in synthetic data but fail when applied to actual datasets.

Recommendation T.2

Clearly Define Intended Use Cases and Disclose Known Limitations

To ensure synthetic data is used appropriately and delivers genuine value to researchers, documentation must explicitly define the intended use cases the data was designed to support and transparently disclose known limitations that might impact research workflows or analytical validity. Documentation should provide both guidance about appropriate applications (e.g., "ideal for code development, pipeline testing, and feasibility assessment") and explicit warnings about inappropriate uses (e.g., "must not be used for statistical inference or publication"), accompanied by concrete examples of suitable applications. Limitations should be presented prominently using clear, non-technical language, enabling researchers to make informed decisions about fitness-for-purpose, avoid wasted effort, ensure successful transfer of methods to real data environments, and prevent inappropriate analytical use that could damage credibility and trust.

3.2.2 Comparisons to Real Data

Researchers consistently emphasised the importance of understanding how the synthetic data compares to the real data. This comparison is essential for assessing whether the synthetic dataset is suitable for their intended task and for understanding what differences or limitations they need to account for.

For low-fidelity synthetic data, researchers wanted to know whether the structural properties of the real data had been preserved, including variable types, value ranges, patterns of missingness, and any quirks or inconsistencies present in the original dataset. Visual comparisons for statistical properties, such as histograms showing the distribution of values, were identified as particularly helpful for quickly assessing similarity.

For high-fidelity synthetic data, researchers required more detailed statistical comparisons, including correlations between variables, distribution shapes, and the preservation of key relationships. Participants expressed a preference for quantitative metrics accompanied by visual representations, such as scatter plots, correlation matrices, or principal component analysis (PCA) comparisons, to provide an intuitive sense of how closely the synthetic data resembles the real data.

Importantly, researchers stressed that comparisons should not only highlight similarities but also clearly identify where the synthetic data differs from the real data. These differences may be intentional, such as the suppression of rare values for privacy protection, or they may be artefacts of the generation process. Either way, researchers need to understand them to avoid misinterpreting results or making incorrect assumptions about the data.

Recommendation T.3

Provide Clear Visual Comparisons Between Synthetic and Real Data

Documentation should include clear, accessible visual comparisons between synthetic and real data to enable researchers to quickly and intuitively assess similarity, understand preserved characteristics, and identify important differences. Quantitative similarity metrics should accompany visualisations to provide precise assessments, but visual representations remain essential to enable researchers to rapidly develop an intuitive understanding of data quality and fitness-for-purpose without requiring deep expertise.

3.2.3 Preprocessing, Cleaning, and Filtering

Participants stressed the importance of understanding any preprocessing, cleaning, or filtering that had been applied to the source data before the synthetic data generation.

Researchers were concerned that if the real data had been cleaned or filtered in ways they were not aware of, they might develop pipelines or code that would fail or produce unexpected results when applied to the actual dataset.

For example, if certain outliers were removed, rare values suppressed, or variables transformed before generation, researchers needed to know about these changes so they could account for them in their workflows. Participants stressed that synthetic data should ideally reflect the "messiness" of real data, including errors, inconsistencies, and missing values, to provide an authentic testing environment. If cleaning or filtering was necessary for privacy or technical reasons, it must be clearly documented so researchers can adjust their expectations and methods accordingly.

Synthetic data documentation should not only describe what is present in the synthetic dataset but also what was present in the source data and how the two differ as a result of the generation process.

This emphasis on preprocessing transparency connects directly to the broader need for clear data provenance documentation. Researchers need to understand the complete journey of the data: which version of the source dataset was used, when it was extracted, what state it was in at that point, and what transformations occurred before synthetic generation began.

Researchers emphasised that understanding preprocessing is particularly critical when synthetic data is intended for code development or pipeline testing. If researchers build data cleaning routines, quality control checks, or error handling based on synthetic data that has already been cleaned, their code may be inadequate when applied to messier real data.

Conversely, if synthetic data deliberately preserves realistic imperfections, such as duplicate records, inconsistent coding, format variations, or logical impossibilities that exist in real data, this should be explicitly documented so researchers understand they are working with authentically messy data rather than flawed synthetic generation.

Recommendation T.4

Document Data Provenance and Any Preprocessing Steps

Documentation must provide comprehensive information about data provenance and all preprocessing, cleaning, or filtering applied to the source data before synthetic generation, enabling researchers to understand exactly what state the data was in, how it was transformed, and how the synthetic dataset relates to the real data they may eventually access.

3.2.4 Evaluation and Quality Assurance

Researchers wanted to know how the synthetic data had been evaluated and what quality assurance processes had been applied. This included understanding what metrics were used to assess fidelity, quality, and utility, and what the results of those evaluations were.

While researchers did not necessarily need to understand every technical detail of the evaluation process, they did need enough information to assess whether the synthetic data meets the requirements necessary for their intended use. Visual summaries, such as comparison plots or summary tables, were identified as particularly useful for communicating evaluation results in an accessible way.

During discussions, a distinction was made about quality evaluation: the difference between structural quality and statistical quality, and how the importance of each varies with fidelity level.

Types of Quality in Synthetic Data

Structural Quality

Structural quality refers to how well the synthetic data replicates the basic structure and characteristics of the real data, independent of statistical relationships. This includes aspects such as:

1. Correct variable names, types, and formats
2. Realistic value ranges that match the real data
3. Appropriate patterns of missingness (both in frequency and structure)
4. Preservation of logical constraints (e.g., dates of death occurring after dates of birth)
5. Presence of realistic data quirks, inconsistencies, and errors that characterise real-world datasets
6. Relational and temporal relationships across tables

Statistical Quality

Statistical quality, by contrast, refers to how well the synthetic data preserves the statistical properties and relationships present in the real data. This includes:

- Accurate univariate distributions (means, medians, variances, distribution shapes)
- Preservation of correlations and associations between variables
- Maintenance of multivariate relationships and conditional distributions
- Accurate representation of subgroup patterns and interactions
- Preservation of complex dependencies across multiple variables

Researchers emphasised that the evaluation approach should align with the fidelity level and intended purpose of the synthetic data. For low-fidelity synthetic data (L1), structural quality is paramount, while statistical quality beyond basic univariate properties is either limited by design or irrelevant to the intended use case.

3.3 Data Owner Perspective: Accessible Explanations and Risk Communication

For data owners and governance teams, transparency requirements are shaped by their dual responsibilities: making informed decisions about whether synthetic data is safe to release and being able to defend and explain those decisions to diverse audiences including regulators, senior management, ethics committees, and the public. Unlike researchers who focus primarily on utility, or experts who engage with technical details, data owners need documentation that bridges technical and non-technical understanding. They must be able to comprehend complex privacy evaluations well enough to assess risk yet also translate those assessments into plain language that non-specialists can understand and trust. Data owners expressed that their key transparency needs include accessible explanations of how synthetic data was generated that avoid overwhelming jargon; clear, interpretable communication of disclosure risks using intuitive categories rather than abstract metrics; honest disclosure of limitations and appropriate use cases to prevent misuse and manage expectations; and documentation structured to serve both technical reviewers and non-technical decision-makers simultaneously. Crucially, data owners emphasised that transparency is not just about understanding the synthetic data itself, but about enabling accountability: they need documentation robust enough to demonstrate due diligence, justify approval decisions if challenged, and maintain public trust in their stewardship of sensitive data.

3.3.1 Lay and Accessible Explanations

Data owners consistently emphasised that they needed the generation process to be explained in plain, non-technical language that they could understand.

Many data owners are not technical experts in synthetic data generation or privacy evaluation, yet they are responsible for making decisions about whether synthetic datasets are safe to release and for defending those decisions to governance boards, regulators, or the public.

This accessibility requirement extends beyond the generation method itself to encompass all aspects of the synthetic data lifecycle, including evaluation, limitations, and governance. Data owners need to be able to confidently answer questions such as: "How was this data made?" "Why is it safe to release?" and "What can and cannot be done with it?" If documentation does not enable them to answer these questions clearly and convincingly, they are unlikely to approve release.

Fig 3.2 Should documentation be tailored to different audiences, or should it just purely contain technical detail?

Members of the public echoed this need for accessible, plain-language explanations, emphasising that transparency is fundamental to maintaining their trust in how sensitive data is being used. Participants in the public workshop expressed that while they did not need to understand technical detail, they wanted to be confident that their data was being handled responsibly and that appropriate safeguards were in place. Public participants stated that accessible explanations demonstrate respect for patients and data subjects, acknowledging their right to understand how their information contributes to research and innovation.

Both data owners and the public also highlighted the need for better defined categories and types of synthetic data. Participants noted that the term "synthetic data" is too broad and can carry different meanings across contexts, leading to confusion and misunderstanding.

3.3.2 Clear Communication of Risk

One of the most significant challenges identified in the workshops was the difficulty of communicating disclosure risk evaluations to data owners in a way that was both accurate and interpretable. Current evaluation metrics are often very technical and often lack intuitive meaning for non-experts.

Data owners didn't feel comfortable being asked to approve synthetic data releases based on metrics they did not fully understand, with little guidance on what constituted acceptable risk or how to weigh different types of risk against one another. They emphasised the need for risk evaluations to be presented using clear, accessible categories, such as a traffic-light system (low, medium, high risk), accompanied by plain-language explanations of what these categories mean in practical terms.

Participants also stressed the importance of explaining the implications of risk findings. What does a particular risk level mean for the organisation? What are the potential consequences if the risk materialises? What additional safeguards, if any, are in place to mitigate residual risk? These contextual explanations help data owners understand not just the technical assessment but the real-world significance of releasing the synthetic data.

Recommendation T.5

Provide Accessible Summaries and Contextualised Risk Communication

Documentation for synthetic data must serve multiple audiences simultaneously, requiring a multi-layered approach that combines technical rigour with accessible communication tailored to different stakeholder needs. This ensures that data owners, governance teams, and members of the public can make informed decisions and maintain confidence in synthetic data practices, while still providing the technical detail necessary for expert review and validation.

3.4 Expert Perspective: Assumptions, Evaluation Methodologies, and Technical Rigour

For experts, including synthetic data generators, privacy specialists, and methodological researchers, transparency requirements focus on technical rigour, reproducibility, and the ability to critically assess the validity of synthetic data generation and evaluation processes. Experts need comprehensive documentation of the technical decisions, assumptions, and trade-offs made throughout the synthetic data lifecycle to determine whether the methods employed were appropriate, whether the evaluations were sufficiently robust, and whether any limitations or gaps might affect the reliability of the results. Their transparency needs extend beyond what was done to include why particular choices were made, what alternatives were considered, and what implications these decisions have for the interpretation and use of the synthetic data. This includes detailed documentation of assumptions embedded in generation algorithms, full disclosure of evaluation methodologies, including what threats were tested, what attacker capabilities were assumed, and what thresholds were used to define acceptable risk; and sufficient technical detail to support reproducibility, peer review, and ongoing validation.

3.4.1 Documenting Assumptions

Experts emphasised the importance of clearly documenting all assumptions made during the generation of synthetic data. These assumptions can significantly influence the characteristics and limitations of the resulting dataset, and users need to understand them to interpret the data correctly and assess its suitability for their purposes.

Assumptions may relate to data distributions, such as assuming normality for continuous variables or independence between certain features. They may involve modelling constraints, such as limiting the complexity of relationships captured or excluding certain variables to simplify generation. They may also relate to preprocessing decisions, such as how missing values were handled, whether outliers were retained or suppressed, or whether certain records were filtered out before generation.

Just as generation processes involve assumptions, so too do evaluation processes. Experts stressed the need to clearly document what privacy threats were considered and why, what assumptions were made about attacker knowledge and capabilities, and what thresholds were used for declaring evaluation success or failure.

For privacy evaluations, documentation should specify what types of attacks were tested (e.g., identity disclosure, membership inference, attribute inference) and what scenarios were considered. For example, evaluations might assume that an attacker has access to publicly available metadata and aggregate statistics but not to the raw source data. Alternatively, they might assume a stronger adversary with access to external datasets that could be linked to the synthetic data.

Thresholds for acceptable risk should also be clearly stated and justified. If a particular threshold is chosen to determine whether synthetic records were sufficiently different from real records, the documentation should explain why this threshold was chosen and what it means in practical terms.

By being transparent about evaluation assumptions and scope, organisations enable users and reviewers to assess whether the evaluation was appropriate and rigorous, and to identify any gaps or limitations that might affect confidence in the results.

Recommendation T.6

Explicitly Document All Assumptions in Generation and Evaluation

Documentation should disclose any assumptions made during both the generation and evaluation processes. For generation, this may include assumptions about data distributions, relationships between variables, or modelling constraints. For evaluation, disclose what privacy threats were considered, assumptions about attacker knowledge, and thresholds used for declaring success or failure. Transparent assumptions prevent overconfidence in results and support more nuanced interpretation.

Checklist **Documentation Considerations**

Lay Executive Summary

☑ Plain language explanation of what fidelity level the synthetic data is
☑ Clear statement that the data is synthetic
☑ Brief overview of generation method
☑ Primary intended uses and unacceptable use cases

Source Data Information

☑ Origin and description of source data used
☑ Version of the source data
☑ Data provenance including any pre-processing, filtering or cleaning
☑ Persistent identifiers (DOI or URL)
☑ Treatment of Outliers

Generation Methods

☑ Specific generation method used
☑ Why specific methods were chosen
☑ Details of any constraints / logic that were embedded
☑ Any assumptions made

Evaluations

☑ Details of any quality assessments
☑ Details of any utility assessments
☑ Details of privacy assessments
☑ Details of any fairness assessments
☑ Any assumptions made
☑ How decisions were made and thresholds selected
☑ Visualisations
☑ Clear explanations and communication for non-technical stakeholders

Other

☑ Citation requirements
☑ Limitations
☑ User Responsibilities

4

Accessible

Principle 4: Synthetic data should generally have lower governance barriers and simplified approval processes compared to real data, with the level of accessibility proportionate to risk and fidelity. Lower-fidelity synthetic data with minimal risk should be made available outside of TREs wherever possible to deliver maximum value, while higher-fidelity data may require more carefully calibrated access controls that balance utility against risk through mechanisms such as user agreements, registration, or controlled access within secure environments.

04. Accessible

Making synthetic data accessible is fundamental to realising its value. If synthetic data cannot be accessed easily and appropriately, much of its potential benefit such as accelerating research workflows, enabling training, supporting data discovery, is lost. However, accessibility must be balanced carefully against safety, transparency, and responsible use.

Throughout the workshops, a clear consensus emerged: accessibility requirements differ fundamentally based on the fidelity level of synthetic data, the sensitivity of the source data, and the intended use case. Low-fidelity synthetic data, which preserves minimal statistical information and carries negligible disclosure risk, should be made readily accessible to maximise its utility for code development, data discovery, and training. In contrast, high-fidelity synthetic data, which more closely resembles real data and carries greater risks, requires more carefully calibrated access controls that balance utility against risk.

Importantly, participants emphasised that accessibility encompasses a spectrum of access models, each involving different levels of user verification, agreement to terms of use, and accountability mechanisms. The goal is to ensure that synthetic data reaches the users who can benefit from it, under conditions that protect privacy, prevent misuse, and maintain public trust.

This section explores the perspectives of researchers, data custodians, and members of the public on how synthetic data should be shared and governed. It examines the appropriate access models for different fidelity levels; the role of registration, user agreements, and accountability mechanisms in managing risk; the balance between automated and human-review approval processes; the specific conditions and restrictions that should govern synthetic data use; and the practical challenges of implementing accessible yet responsible sharing frameworks. By establishing clear, proportionate, and practical approaches to accessibility, organisations can ensure that synthetic data delivers maximum value while maintaining the safety and trust that underpin responsible data stewardship.

4.1 Access Requirements by Fidelity Level

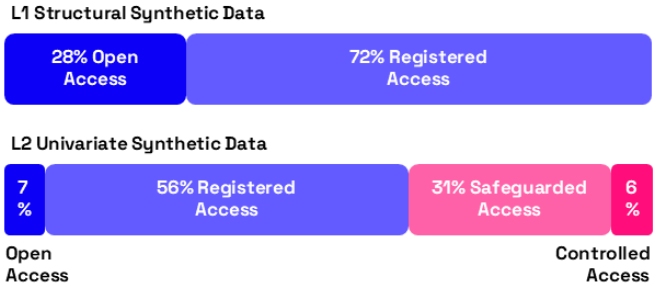
4.1.1 L1 Structural Synthetic Data

For L1 structural synthetic data, which preserves only the basic structure of datasets without any statistical properties, the majority of data owners felt that registered access was the most appropriate model. Some participants were comfortable with fully open access, viewing L1 synthetic data as carrying minimal risk and being no more sensitive than published metadata. However, the prevailing view was that at minimum, some form of registration should be required.

The rationale for requiring registration, even for very low-risk data, was twofold. First, registration provides a minimal but meaningful barrier that encourages users to engage thoughtfully with the data rather than treating it as entirely disposable. Second, it creates a record of who has accessed the data, enabling data owners to communicate updates, corrections, or guidance if needed, and providing a basic mechanism for accountability if concerns arise.

Importantly, registered access does not involve formal approval processes or lengthy applications. Users simply provide contact information and, ideally, acknowledge that they understand the data is synthetic and intended for specific purposes. This approach balances accessibility with a minimal level of user engagement and traceability.

Fig 4.1 What minimum level of access should be required for the following fidelity levels?



Access Models

To structure discussions around appropriate access mechanisms, the workshops explored five categories of access control, ranging from fully open to highly restricted:

Open Access:
No restrictions; synthetic data is publicly available for download without any registration or user agreement.

Registered Access:
Users must create an account and provide basic information (e.g., name, email, institutional affiliation) before downloading, but no formal approval process is required.

Safeguarded Access:
Users must agree to an End User Licence Agreement (EULA) and may require approval from the data owner before access is granted. Access is typically granted outside secure environments, but with clear terms of use and accountability mechanisms.

Controlled Access:
Synthetic data is only accessible within a Trusted Research Environment (TRE), with the same technical safeguards as real data but significantly lighter governance and approval processes.

Secure Access:
Synthetic data is governed and accessed in the same way as the real data, with full governance processes, approvals, and secure environment requirements.

4.1.2 L2 Univariate Synthetic Data

For L2 univariate statistical synthetic data, which preserves individual variable distributions but not relationships between variables, opinions were more divided. A significant number of data owners felt that registered access was sufficient, particularly when the synthetic data was generated using publicly-sourced information. In these cases, the data carries negligible additional disclosure risk beyond what is already present in published summary statistics.

However, a roughly equal number of participants argued for safeguarded access, requiring users to agree to an End User Licence Agreement (EULA) before download. This preference was driven by several concerns:

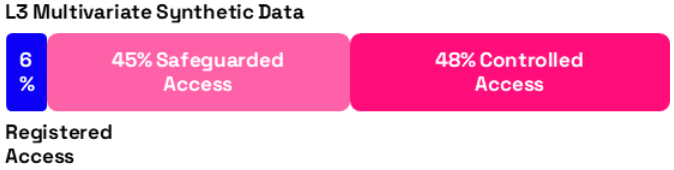
- **Source Data Sensitivity:** When synthetic data is privately-sourced (generated directly from real datasets rather than public statistics) data owners felt that safeguarded access with explicit user agreements was more appropriate, even if disclosure risk remained low.
- **Risk of Misinterpretation:** Participants felt that L1 synthetic data, because it contains no statistical properties, is clearly artificial and unlikely to be mistaken for real data. L2 data, by contrast, includes realistic distributions that could lead users to incorrectly assume it is suitable for analytical tasks. Requiring users to sign a EULA ensures they explicitly acknowledge the synthetic nature of the data and its intended limitations.
- **Preventing Unintentional Misuse:** Participants worried that researchers unfamiliar with synthetic data might use L2 datasets for statistical analyses, hypothesis testing, or even publication without understanding that such use cases are invalid. A EULA provides an opportunity to clearly communicate these limitations and establish accountability if misuse occurs.

Overall, the consensus for L2 synthetic data leaned toward safeguarded access with a EULA, particularly for privately-sourced datasets or contexts where there is greater potential for misunderstanding or misuse.

4.1.3 L3 Multivariate Synthetic Data

For L3 multivariate correlated synthetic data, which preserves complex relationships between variables, data owners felt that access should involve either safeguarded access with robust user agreements or controlled access within a TRE. Roughly equal numbers of participants supported each approach, with preferences depending heavily on individual risk appetite, organisational context, and the specific characteristics of the synthetic dataset.

Fig 4.2 What minimum level of access should be required for the following fidelity level?



Those favouring safeguarded access outside TREs argued that if the synthetic data has been rigorously evaluated and found to meet anonymisation standards, it should be possible to share it with appropriate safeguards such as formal user agreements, clear documentation of limitations, and accountability mechanisms. This approach enables broader accessibility and supports use cases such as advanced training, innovation development, and hypothesis generation, while still ensuring responsible use.

Conversely, those favouring controlled access within TREs emphasised that L3 synthetic data, due to its higher fidelity and closer resemblance to real data, carries greater reputational and disclosure risks. By keeping it within secure environments, organisations can maintain stronger oversight, reduce the risk of misuse, and benefit from the "means reasonably likely to be used" principle in GDPR, which considers whether data remains within a secure setting when assessing anonymisation.

From the expert workshop, participants acknowledged this tension. They felt that synthetic data should be made accessible outside TREs wherever possible to maximise its value, but recognised that from a regulatory and governance perspective, keeping higher-fidelity data within secure environments may reduce perceived risk and make approval more straightforward.

Ultimately, the appropriate access model for L3 synthetic data must be determined on a case-by-case basis, taking into account the sensitivity of the source data, the robustness of privacy evaluations, the intended use case, and the organisation's tolerance for reputational risk.

4.1.4 L4 Augmented Synthetic Data

For L4 augmented synthetic data, which is designed to enhance or extend real datasets and closely replicates their characteristics, there was universal agreement that access should be restricted in the same way as the real data. This level of synthetic data carries the same disclosure risks and governance requirements as the source dataset and must therefore be subject to the same approval processes, secure environment requirements, and use conditions.

L4 synthetic data is not intended to bypass governance or simplify access; rather, it serves specific analytical purposes such as class balancing for machine learning or simulating rare scenarios. As such, it should be treated as an extension of the real data, not as a separate, lower-risk resource.

4.2 Balancing Risk: Disclosure vs Misuse

A recurring theme in the data owner workshop was the distinction between disclosure risk and misuse risk, and how these different types of risk influence access decisions.

Data owners broadly agreed that sharing low-fidelity synthetic data (L1 and L2) poses minimal additional disclosure risk compared to sharing aggregated summary statistics or metadata. Because these datasets do not preserve complex relationships or individual-level patterns, the technical risk of re-identification is very low, particularly when generation methods are transparent and evaluations confirm anonymisation.

This recognition supported the view that low-fidelity synthetic data can and should be made relatively accessible, with lighter governance processes than those required for real data.

However, data owners emphasised that for low-fidelity synthetic data, misuse risk is often a greater concern than disclosure risk. Misuse encompasses scenarios such as:

- Researchers using synthetic data for analytical purposes it was never designed to support, leading to invalid or misleading findings
- Findings derived from synthetic data being published or cited as if they were based on real-world evidence

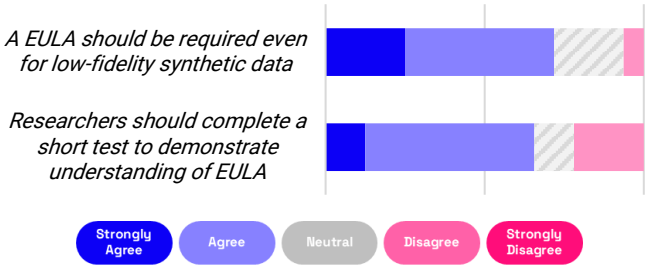
- Synthetic data being shared or redistributed without acknowledgment of its synthetic nature and limitations
- Users misunderstanding the purpose and scope of synthetic data, leading to inappropriate applications

To address these concerns, data owners felt strongly that even low-fidelity synthetic data should be accompanied by clear user agreements that explicitly define appropriate and inappropriate uses and establish accountability.

4.3 The Role of End User Licence Agreements (EULAs)

End User Licence Agreements emerged as a key mechanism for managing both misuse risk and accountability. Most data owners felt that a EULA should be required for all but the most basic structural synthetic data, and certainly for any privately-sourced synthetic datasets.

Fig 4.3 To what extent do you agree with the following statements?



4.3.1 Ensuring Understanding and Acknowledgment

The primary purpose of a EULA is to ensure that users explicitly acknowledge the synthetic nature of the data, understand its intended use cases and limitations, and agree to use it responsibly. By requiring users to actively engage with these terms before gaining access, EULAs help prevent unintentional misuse and provide a clear record of user responsibilities.

Data owners stressed that EULAs must be concise, clear, and accessible. If they are excessively long or filled with legal jargon, users are unlikely to read them carefully, undermining their effectiveness.

Some data owners even proposed requiring users to complete a short quiz or comprehension check before access to confirm they have understood the key terms and limitations. Proponents of this approach argued that it would ensure genuine engagement with the EULA and reduce the risk of misuse due to misunderstanding.

However, other participants felt this was unnecessary and could create friction that discourages legitimate use. They argued that the EULA itself, combined with clear documentation and labelling, should be sufficient, and that adding additional barriers risks undermining the accessibility benefits of synthetic data.

This debate highlights the need to balance accessibility with assurance of understanding. A practical middle ground might involve requiring explicit acknowledgment of key limitations (e.g., "I understand this data is synthetic and should not be used for statistical inference") as part of the access process, without imposing quizzes or tests which could prolong the process.

4.3.2 Accountability

Data owners also discussed whether user agreements should require organisational sign-off, particularly for higher-fidelity synthetic data or contexts where misuse could have significant consequences. Under this model, a senior colleague, supervisor, or institutional representative would be required to co-sign the EULA, increasing accountability and ensuring that organisations are aware of and responsible for their employees' or students' use of synthetic data.

This approach was seen as particularly relevant for educational or training contexts, where students or early-career researchers might not fully appreciate the limitations of synthetic data. By involving supervisors or institutional representatives, organisations can ensure appropriate oversight and reduce the risk of inappropriate use.

However, many data owners acknowledged that if someone intended to deliberately misuse synthetic data, a EULA would likely not prevent them from doing so anyway. A determined bad actor could simply agree to the terms with no intention of complying. Despite this recognition, data owners emphasised that EULAs and organisational sign-off still serve a valuable purpose: they primarily guard against unintentional misuse. Most misuse of synthetic data is likely to arise not from malicious intent but from misunderstanding, lack of awareness, or insufficient appreciation of the data's limitations. In these cases, requiring users to explicitly engage with terms of use, acknowledge limitations, and involve their organisations creates multiple opportunities for education and reflection, significantly reducing the likelihood of inadvertent misuse.

Furthermore, even if EULAs cannot prevent intentional bad actors, they establish clear expectations and provide grounds for accountability if misuse is discovered. They create a documented record of user responsibilities that can support consequences such as being banned from applying for real data access from the TRE.

However, participants also recognised that requiring organisational sign-off adds administrative burden and may slow access. As such, it should be reserved for contexts where the risk profile justifies the additional oversight, rather than being applied universally.

4.3.3 Data Retention

Another consideration raised by data owners was whether access to synthetic data should be time-limited, particularly when datasets are shared outside secure environments.

The rationale for time limits is that synthetic data, like any dataset, may become outdated, superseded by newer versions, or subject to revised risk assessments as understanding of privacy threats evolves. By requiring users to periodically renew access or delete old versions, data owners can ensure that users are working with the most current and appropriately evaluated synthetic datasets.

Participants acknowledged that enforcing time limits outside TREs is challenging, as data custodians have limited visibility into what users do with downloaded datasets.

However, they felt that including time limits as a condition in the EULA establishes a clear expectation of responsible use and provides grounds for accountability if misuse is discovered.

Practical implementation of time-limited access might involve periodic email reminders to users, automated expiration of download links, or requirements for users to reapply for access after a defined period.

Checklist Things to Consider Having in a EULA

- ☑ Definition of Synthetic Data and its Fidelity
- ☑ User Obligations
- ☑ Clear Identification and Acknowledgment of Synthetic Nature
- ☑ Intended Use Cases and Permitted Applications
- ☑ Prohibited Uses
- ☑ Acknowledgment of Limitations
- ☑ Attribution and Citation Requirements
- ☑ Non-Redistribution and Data Control
- ☑ Time Limits and Data Retention
- ☑ Consequences of Violation

4.3.4 Tailoring EULAs to Fidelity Level and Use Case

Data owners felt that EULAs should not be one-size-fits-all but should be tailored to the fidelity level and intended use cases of the synthetic data. The complexity, stringency, and specific requirements of a EULA should be proportionate to the risk profile and potential consequences of misuse.

For low-fidelity synthetic data (L1-L2) designed for code development, data discovery, or basic training, EULAs can be relatively straightforward, focusing primarily on ensuring users understand the synthetic nature of the data, acknowledge intended use cases and limitations, and agree not to use the data for analytical purposes. These agreements should be concise and accessible, avoiding unnecessary complexity that could create barriers to the rapid access that makes low-fidelity synthetic data valuable.

For higher-fidelity synthetic data (L3) that preserves complex relationships and may support certain analytical applications, EULAs require more comprehensive terms. This might include organisational sign-off requirements to ensure institutional oversight, validation obligations requiring analytical findings to be confirmed against real data before publication, time-limited access provisions to ensure users work with current versions, explicit acknowledgment of specific limitations that could affect analytical validity, and more stringent accountability mechanisms with clearly defined consequences for misuse.

Data owners emphasised that this graduated approach ensures that accessibility is not unnecessarily restricted for low-risk datasets, while higher-risk datasets receive appropriate safeguards commensurate with their potential for misuse or privacy concerns. The principle should be proportionality: match the stringency of EULA requirements to the actual risk and potential impact of misuse, ensuring that lower-fidelity synthetic data can be shared with minimal friction while higher-fidelity data receives the governance attention it requires.

This tailoring also extends to specific clauses within EULAs. For example, prohibited uses for L1 data might simply state "must not be used for any analytical or statistical inference," while L3 data might specify "preliminary analytical findings may be generated but must be validated against real data and clearly marked as derived from synthetic data."

4.4 Approval Process

The workshops also explored whether access to synthetic data should involve human review and approval, or whether it could be fully automated based on predefined criteria.

Proponents of automated approval argued that one of the main benefits of synthetic data is speed of access. If every request must be manually reviewed and approved, delays are reintroduced, undermining the value proposition. For low-fidelity synthetic data with minimal risk, automated approval based on registration and agreement to a EULA was seen as appropriate and efficient.

However, data owners and members of the public expressed some discomfort with fully automated access, particularly for higher-fidelity synthetic data or sensitive contexts. They felt that having a human review applications, even briefly, provides an additional safeguard against inappropriate use and ensures that data owners retain oversight of who is accessing their data and for what purposes.

Members of the public, in particular, felt more comfortable with a human-in-the-loop approach, viewing it as a sign of responsible stewardship and accountability. They expressed concerns that fully automated systems might allow access to users with inappropriate intentions or insufficient understanding, without any mechanism for detecting or preventing misuse.

4.5 Researchers Perspectives on Accessibility

From the researcher workshop, a resounding message emerged: for many of the most valuable use cases such as data discovery, code development, and basic training, the primary benefit of synthetic data lies in its accessibility. Researchers emphasised that if synthetic data requires the same lengthy approval processes, governance barriers, and access restrictions as real data, much of its value is negated.

This perspective was particularly strong for low-fidelity synthetic data (L1 and L2), where the primary purpose is functional rather than analytical. Researchers do not need these datasets to perfectly replicate statistical relationships; they need them to be structurally realistic, readily available, and usable in their own computing environments. The ability to download synthetic data, work with it on personal machines, share it with collaborators, and iterate rapidly on code development is precisely what makes it valuable for preparatory work.

However, researchers also acknowledged that accessibility must be appropriate to the risk profile of the data. They recognised that higher-fidelity synthetic data, particularly datasets that preserve complex multivariate relationships, may require more stringent controls. The key principle articulated by researchers was proportionality: access mechanisms should be as light-touch as possible while still ensuring responsible use and protecting privacy.

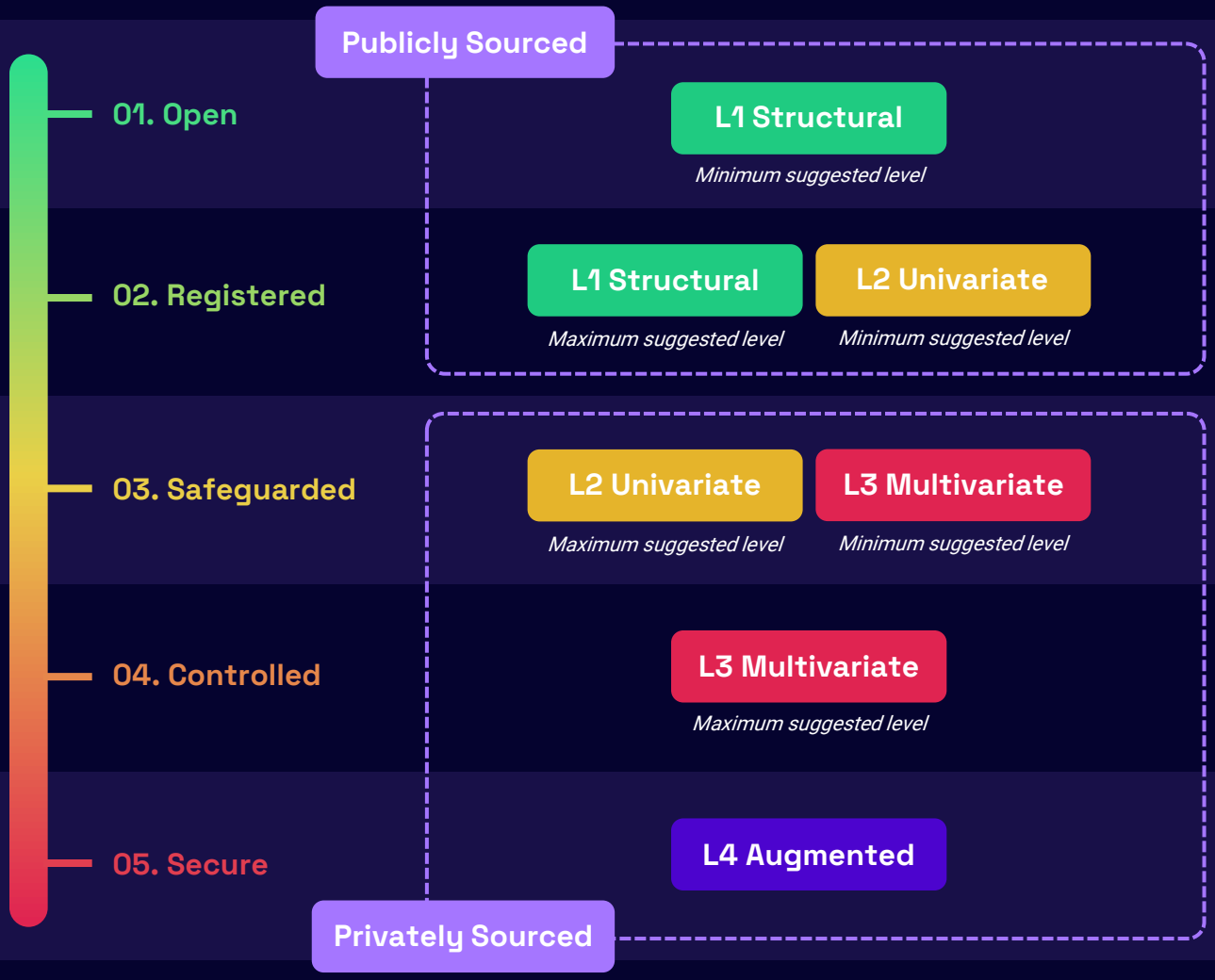
When it came to access for training and educational purposes, participants expressed somewhat divergent views on whether synthetic data should be accessible outside TREs or restricted to secure environments.

Many researchers argued that for training purposes, synthetic data should be accessible outside TREs to maximise flexibility, enable broader participation, and avoid imposing unnecessary technical and administrative barriers on students. Training datasets need to be readily available, easy to work with, and usable in standard computing environments where students are learning to code and analyse data.

Conversely, some participants felt that there is value in conducting certain types of training within TREs, even when using synthetic data. This approach familiarises students with secure environment workflows, access controls, and the practical realities of working with sensitive data. It also ensures that higher-fidelity synthetic data, which might be needed for teaching advanced analytical techniques, remains within controlled settings.

Ultimately, participants agreed that the appropriate access model for training depends on what is being taught and what level of fidelity is required.

Recommended Access Models for Synthetic Data Based on Data Custodian Perspectives



These recommended access model requirements reflect data custodian perspectives gathered during the workshop. It should be noted that L3 multivariate synthetic data is already frequently released under open access arrangements. The choice of access model ultimately depends on the context as well as the organisation's experience, risk appetite, and confidence in releasing synthetic datasets.

5

Representative

Principle 5: High-Quality synthetic data must reflect the structure, patterns, diversity, and characteristics of the original data in ways appropriate to its intended use case and fidelity level, whether that means capturing data quality issues and irregularities for code development or preserving statistical relationships and subgroup diversity for analytical applications.

05. Representative

The Representative principle addresses whether synthetic data maintains appropriate quality and faithfulness to the source data, ensuring it accurately reflects the structure, distributions, patterns, and diversity of the original dataset in ways that align with its intended use. This includes preserving realistic characteristics such as data messiness, irregularities, and patterns of missingness where appropriate, maintaining statistical properties like distributions and correlations when required for analytical purposes, and ensuring fair representation across different subgroups to avoid amplifying biases or creating algorithms that perform poorly for underrepresented populations. Quality assessment must be tailored to intended use cases, with evaluation criteria focused on whether the synthetic data is fit-for-purpose rather than applying universal standards.

Summary of Key Recommendations:

01. Ensure Plausibility and Logical Coherence in Synthetic Data

Synthetic data must maintain plausibility and logical coherence by embedding real-world constraints directly into generation algorithms wherever possible (e.g., ensuring birth dates precede death dates) and employing targeted manual inspection by domain experts to ensure it meets expectations.

02. Authentically Preserve Data Messiness and Irregularities

For use cases like code development and training, synthetic data should authentically preserve the messiness, errors, and irregularities present in real-world source data, including missing values, inconsistent formatting, and data quality issues, as these characteristics are essential for researchers to develop robust pipelines and error-handling mechanisms that will transfer successfully to actual datasets.

03. Assess and Address Bias in Synthetic Data for Analytical Use

For high-fidelity synthetic data intended for analytical or AI development purposes, systematically assess source data biases before generation to determine whether they should be preserved (reflecting genuine population patterns), mitigated (addressing historical inequities or systematic underrepresentation), or prominently documented for user awareness. After generation, evaluate synthetic datasets using appropriate fairness metrics with results transparently reported, and any users should validate any analytical findings or AI models on independent real-world datasets before considering them reliable for publication or deployment.

05. Representative

Ensuring that synthetic data is representative is fundamental to its usefulness across all fidelity levels and use cases. However, what "representative" means in practice differs significantly depending on whether synthetic data is intended for functional purposes, such as code development and training, or for analytical purposes, such as hypothesis generation or AI model development.

For lower-fidelity synthetic data, representativeness centres on structural fidelity and authentic messiness: capturing the real-world characteristics, quirks, imperfections, and challenges that researchers will encounter when working with actual data. For higher-fidelity synthetic data intended for analytical use, representativeness extends to statistical accuracy, preservation of meaningful relationships, and critically, ensuring that the synthetic data does not introduce, amplify, or perpetuate biases that could lead to unfair or discriminatory outcomes.

Throughout the workshops, participants emphasised that synthetic data which fails to be representative undermines its core value proposition. Overly cleaned, idealised, or simplified synthetic data may appear technically correct but fails to prepare researchers for the realities of working with messy, complex real-world datasets. Conversely, high-fidelity synthetic data that inadvertently amplifies biases or misrepresents minority populations can cause significant harm, both by producing invalid analytical findings and by perpetuating systemic inequities in research and AI applications.

This section explores what it means for synthetic data to be representative, examining both structural representativeness (authentically reflecting the characteristics, complexity, and imperfections of real datasets) and statistical representativeness (preserving meaningful patterns, relationships, and distributions without distortion or bias).

By addressing these dimensions holistically, organisations can create synthetic data that genuinely reflects the complexity, diversity, and reality of the data it represents, ensuring it delivers authentic value for its purpose while avoiding the introduction or amplification of harmful biases.

5.1 Structural Representativeness

For lower-fidelity synthetic data used in preparatory activities such as code development, data discovery, and basic training, representativeness is fundamentally about structural fidelity and authentic messiness. Researchers emphasised that synthetic data must reflect not only the basic structure of real datasets but also the imperfections, irregularities, and challenges that make real-world data difficult to work with.

5.1.1 Variable Naming, Types, and Value Ranges

Representative synthetic data must accurately replicate the fundamental structural elements of the source dataset, including variable names, data types, and the range of possible values for each variable. These basic structural features form the foundation upon which researchers build their analytical workflows and are essential for ensuring that code developed using synthetic data will function correctly when applied to real data.

Workshop participants emphasised several key aspects of structural accuracy:

- **Variable names and labels:** Synthetic data should use the same variable names and labels as the real dataset. Changes to variable naming conventions can break scripts and queries that researchers develop using synthetic data, creating frustration and wasted effort when transitioning to the real data environment.
- **Data types:** Each variable should have the same data type as in the source data, including distinctions between integers, floating-point numbers, strings, dates, categorical variables, and boolean values.
- **Value ranges:** For continuous variables, synthetic data should span the full observed range in the source data, for categorical variables, all categories present in the real data should be represented in the synthetic data.
- **Formatting conventions:** Dates, times, and other formatted values should follow the same conventions as the source data.

Participants noted that seemingly minor deviations in structural elements can have cascading effects on code functionality. For instance, if the real data contains a categorical variable with ten possible values but the synthetic data only includes eight of them, any code that expects all ten categories will need to be revised. Similarly, if date formats differ between synthetic and real data, date parsing functions may fail or produce incorrect results.

The importance of structural accuracy extends to more complex data structures as well. For datasets with hierarchical structures, such as students within schools or patients within hospitals, synthetic data should preserve realistic grouping structures and within-group relationships. For longitudinal datasets with repeated measurements over time, synthetic data should maintain the temporal structure, including realistic patterns in the timing and frequency of observations.

5.1.2 Logical Coherence and Realism

Synthetic data that contains impossible or nonsensical combinations of values undermines trust and usability, as researchers cannot be confident that the synthetic data authentically represents the structure and constraints of the real data.

Workshop participants highlighted several aspects of logical coherence that must be maintained:

Temporal consistency: Dates must follow logical sequences, such as dates of death occurring after dates of birth, and treatment start dates preceding treatment end dates etc.

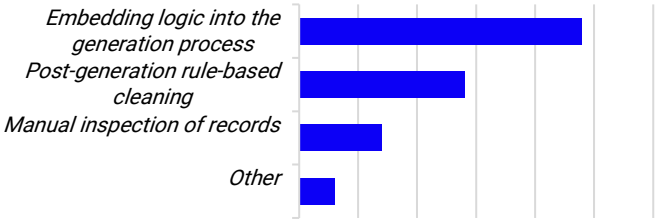
Referential integrity: For relational datasets, foreign key relationships should be maintained. If a patient record references a particular hospital or clinic, that hospital or clinic should exist in the corresponding reference table.

Plausibility: Variables should respect known relationships, such as pregnancy-related variables only appearing for individuals of appropriate sex.

Participants noted that maintaining logical coherence is particularly important for code development and federated query development. Researchers developing analytical pipelines, database queries, or federated analytics systems need to test their code against data that behaves according to real-world rules and constraints. Synthetic data that violates logical relationships can cause code to fail in unexpected ways or produce misleading test results, meaning that workflows may appear to function correctly during development but then break when applied to real data that respects actual domain constraints.

However, participants also acknowledged that some level of inconsistency may be appropriate if it reflects genuine issues in the source data. For example, if the real dataset contains miscoded or mislabelled records, replicating these issues in synthetic data can be valuable for teaching data quality assessment and cleaning techniques. The key distinction is between inconsistencies that reflect real-world data quality problems, which should be preserved, and unrealistic logic that would never occur in reality, which should be avoided and dealt with appropriate constraints.

Fig 5.1 How do you ensure plausibility of synthetic data?



The most common approach, employed by the majority of expert participants, was to embed logical rules and constraints directly into the generation process. This method involves specifying domain knowledge, rules, and known relationships as part of the generation algorithm itself, ensuring that synthetic records respect these constraints from the outset.

For example, generation methods might be configured to enforce that pregnancy-related variables only appear for individuals of appropriate sex, that temporal sequences respect chronological ordering, or that referential integrity is maintained across related tables. However, experts acknowledged that this approach requires significant domain expertise upfront to identify and correctly specify all relevant rules and constraints, and that overly rigid constraints can sometimes reduce the statistical fidelity of the synthetic data by artificially limiting the patterns the generation algorithm can explore.

Some participants described using post-generation cleaning and correction to address plausibility issues. Under this approach, synthetic data is initially generated with minimal hard constraints and then subjected to rule-based checks and corrections to identify and fix implausible combinations. For instance, records containing impossible date sequences might be flagged and corrected, or biologically implausible values might be adjusted to fall within acceptable ranges.

Several experts also emphasised the importance of manual inspection of synthetic records as a quality assurance step, regardless of which generation approach was used. This involves reviewing samples of synthetic data, either by visually examining individual records or by running domain-specific validation checks, to identify any implausible patterns or logical inconsistencies that automated methods might have missed.

Recommendation R.1

Ensure Plausibility and Logical Coherence in Synthetic Data

Synthetic data must maintain plausibility and logical coherence by respecting real-world constraints through embedding logical rules directly into generation algorithms wherever possible and employing manual inspection by domain experts as a targeted quality assurance step to identify subtle implausibilities that automated methods may miss.

For datasets with complex structures such as relational databases, longitudinal records, or hierarchical data, representativeness requires preserving the structural complexity and relationships that characterise the source data. This goes beyond simply replicating individual variables to encompass the ways in which different parts of the dataset connect and relate to one another.

Participants working with health datasets emphasised the importance of preserving realistic patient journeys and temporal patterns. For example, synthetic electronic health records should reflect realistic sequences of events, such as diagnosis following symptoms, treatment following diagnosis, and outcomes following treatment. Simply generating independent synthetic records for each timepoint would fail to capture the longitudinal structure that is central to the dataset's utility. Temporal patterns should include realistic intervals between events, such as the typical time between diagnosis and treatment initiation, or between successive follow-up appointments.

For relational databases with multiple linked tables, synthetic data should maintain referential integrity and realistic patterns of linkage across tables. If the source data includes patient records linked to appointment records, prescription records, and laboratory results, the synthetic data should preserve these linkages in realistic ways, ensuring that the pattern of connections between tables mirrors what would be seen in reality.

Workshop participants acknowledged that preserving structural complexity in synthetic data is technically challenging and may require sophisticated generation methods, particularly for high-fidelity datasets. However, they stressed that even low-fidelity synthetic data used for code development should reflect the basic structural architecture of the source data to ensure that analysis pipelines are designed appropriately. Researchers need to understand how tables relate to one another, how longitudinal records are structured over time, and how hierarchical groupings are organised, even if the synthetic data does not preserve all the detailed statistical relationships within and across these structures.

5.1.3 Patterns of Missingness

Missing data is a pervasive feature of real-world datasets, particularly in health, social science, and administrative contexts. However, missingness is not random; it often follows systematic patterns that researchers must understand and account for in their analyses. Representative synthetic data must therefore preserve realistic patterns of missingness that reflect the characteristics of the source data.

Researchers emphasised that the presence of realistic missingness is particularly important for code development and training. If synthetic data contains complete records with no missing values, researchers may develop analysis pipelines that fail when applied to real data. Similarly, students learning to handle missing data need exposure to realistic patterns of incompleteness to develop appropriate analytical skills, including techniques for handling missing values, assessing the potential impact of missingness on results, and selecting appropriate imputation methods when needed.

5.1.4 Messiness, Inconsistencies, and Data Entry Errors

Researchers consistently stressed that one of the most critical aspects of representative synthetic data is its inclusion of realistic imperfections. Real-world datasets are inherently messy: they contain inconsistent formatting, data entry errors, and various irregularities that reflect human and system limitations. These characteristics are not flaws to be eliminated in synthetic data; rather, they are essential features that must be preserved to ensure synthetic data serves its intended purpose.

Workshop participants stressed that while some of these errors might be corrected during data cleaning before analysis, the synthetic data used for code development and training should replicate the kinds of errors researchers are likely to encounter. This ensures that analysis scripts and data pipelines are robust and can handle real-world data quality issues rather than assuming perfectly formatted input.

However, participants stressed that synthetic data should contain plausible errors and inconsistencies that reflect genuine patterns in the source data, not introduce new categories of errors that don't exist in the original. The goal is to authentically represent the messiness of real data, not to create synthetic data that is artificially more problematic than the source. Researchers expressed some concern about wasting time and effort developing code to handle errors, edge cases, or data quality issues that are present in the synthetic data but absent from the real data they will eventually access. If synthetic data contains problems introduced during generation, such as encoding errors not present in the source, implausible value combinations that would never occur in reality, or structural inconsistencies that are generation artefacts rather than genuine features, then researchers may invest substantial effort building validation checks, error handling routines, and cleaning procedures that prove unnecessary or even counterproductive when applied to the actual dataset. This mismatch not only wastes development time but can also create false confidence in code that appears to work correctly on synthetic data but fails to handle the actual patterns of messiness present in real data.

Recommendation R.2

Authentically Preserve Data Messiness and Irregularities

Synthetic data should preserve authentic messiness, errors, and irregularities present in real-world source data as these characteristics are essential for enabling researchers to develop robust code and appropriate error-handling mechanisms that will transfer successfully to actual datasets.

5.1.4 Outliers and Extreme Values

Outliers and extreme values are common in real-world data and can have significant implications for analysis. Synthetic data that systematically excludes or suppresses outliers fails to represent the full range and complexity of the source data, potentially leading researchers to develop methods that are fragile or fail when encountering unusual cases in the real dataset.

Workshop participants noted that the treatment of outliers in synthetic data involves a careful balance between representativeness and privacy protection:

The representativeness perspective: Researchers need to develop analyses that can handle outliers appropriately. Training datasets that lack any extreme values fail to prepare them for this reality. Code developed on overly clean synthetic data may fail when encountering the variability and edge cases present in real data. Students learning statistical methods need exposure to outliers to understand concepts like robust estimation, outlier detection, and the impact of extreme values on different analytical techniques.

The privacy perspective: From a privacy standpoint, extreme and rare values can increase disclosure risk, particularly when multiple unusual characteristics co-occur in a single record. For high-fidelity synthetic data careful evaluation is needed to determine whether preserving specific outliers is essential for the intended use case and whether it can be achieved without unacceptable privacy risk.

Members of the public raised important concerns about the ethical implications of systematically removing or suppressing outliers in synthetic data. They emphasised that outliers are often not just statistical anomalies but represent real people, frequently those from minority populations, individuals with rare conditions, or people whose characteristics or experiences differ from the majority. Removing these outliers to reduce disclosure risk, while technically justifiable from a privacy perspective, can have the unintended consequence of systematically excluding already marginalised groups from synthetic datasets.

Public participants expressed worry that this practice could perpetuate health inequalities and research biases, as analyses developed using synthetic data lacking representation of minority groups or rare conditions may be inadequate or inappropriate when applied to the full diversity of the real population. They stressed that decisions about outlier handling should explicitly consider the equity implications: who is being excluded, what populations are being underrepresented, and whether privacy protections are inadvertently creating or reinforcing disparities in whose data contributes to research and innovation. This perspective highlighted that the treatment of outliers is not purely a technical privacy-utility trade-off but has significant ethical dimensions that must be transparently acknowledged and carefully weighed when generating and documenting synthetic data.

5.2 Statistical Representativeness

For higher-fidelity synthetic data intended for analytical purposes, representativeness extends beyond structural authenticity to encompass statistical accuracy and fairness. In these contexts, synthetic data must not only capture the distributions and relationships present in the source data but must also avoid introducing, amplifying, or perpetuating biases that could lead to invalid findings or discriminatory outcomes.

5.2.1 Addressing Bias and Ensuring Fairness

Although synthetic data can be used as a technique to mitigate bias and improve representation, the potential for synthetic data to reinforce, amplify, or introduce bias emerged as a significant concern across all workshops, particularly when synthetic data is intended for analytical purposes or AI model development. Data owners and members of the public expressed worry that high-fidelity synthetic data, if not carefully generated and evaluated, could perpetuate existing biases present in the source data or even create new biases through the generation process, leading to discriminatory outcomes in subsequent analyses or algorithmic applications.

Fig 5.2 How concerned are you that high-fidelity synthetic data could reinforce or amplify biases?

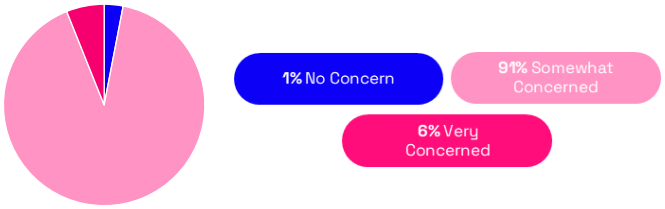
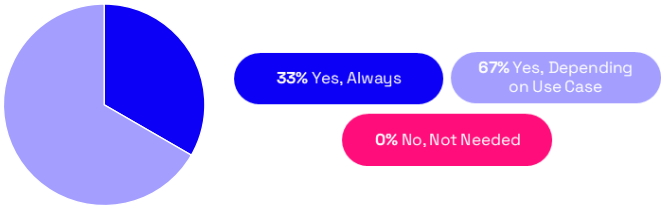


Fig 5.3 Should synthetic data be evaluated for fairness?



Data owners consistently emphasised that evaluating fairness metrics should be a critical component of synthetic data assessment before release. The majority of data owners felt that fairness assessment should depend on the intended use case, with their primary concerns focused on high-fidelity synthetic data designed to support analytical work, hypothesis generation, or AI model development. In these contexts, where synthetic data directly influences research findings or algorithmic outputs, the potential for bias to lead to invalid conclusions or discriminatory outcomes was seen as particularly problematic and warranting rigorous fairness evaluation. Some data owners even felt that fairness should always be assessed regardless of use case, reasoning that even lower-fidelity synthetic data used for training or code development shapes how researchers think about data and may inadvertently reinforce stereotypes or misrepresentations if biased patterns are present.

Members of the public raised additional concerns about how privacy preservation techniques themselves might disproportionately affect minority populations and vulnerable groups. They noted that practices such as suppressing rare value combinations, aggregating small categories, or removing outliers, while technically necessary for disclosure control, systematically reduce the representation of minority populations, people with rare conditions, and individuals whose characteristics differ from the majority.

Recommendation R.3

Assess and Address Bias in Synthetic Data for Analytical Use Cases

Synthetic data generation must explicitly address bias and fairness for high-fidelity datasets intended for analytical purposes or AI model development, through systematic assessment before generation and rigorous evaluation after generation. Before creating synthetic data, generators should assess what biases exist in the source data to determine whether they should be preserved, actively, or prominently documented (so users understand limitations and can make informed decisions about appropriate use). After generation, synthetic datasets intended for analytical or AI development purposes must be evaluated using appropriate fairness metrics, with results transparently reported in documentation alongside clear explanations of what the metrics mean and what limitations remain. This ensures that synthetic data does not inadvertently perpetuate or amplify existing biases, that decisions about bias preservation versus mitigation are made intentionally and transparently rather than by default, and that users have the information needed to understand how representative the synthetic data is across different populations and what implications this has for the validity and equity of analyses or models developed using it. Any analytical findings or AI models developed using synthetic data should also be validated on independent real-world datasets wherever possible before being considered reliable, published as research evidence, or deployed in practice, as synthetic data cannot guarantee that learned patterns will generalise appropriately to actual populations or that fairness improvements on synthetic data will hold in real-world applications.

Participants from both the expert and researcher workshops discussed their experiences using synthetic data for AI model development, particularly for data augmentation to address class imbalances, underrepresented populations, or limited sample sizes. Several experts described successfully employing synthetic data to generate additional training examples for minority classes in machine learning tasks, such as creating synthetic cases of rare diseases to improve diagnostic algorithm performance or generating synthetic examples of underrepresented demographic groups to reduce algorithmic bias.

However, participants who had employed synthetic data for augmentation emphasised the critical importance of validation using independent real-world datasets. They stressed that while synthetic data augmentation can improve model performance in controlled testing environments, it provides no guarantee that models will generalise appropriately to actual populations.

This validation requirement is particularly important when synthetic data is used to address bias or improve fairness. Researchers noted that synthetic data generated with the intention of creating more balanced representations may inadvertently amplify existing biases.

Checklist **Quality Assurance Checklist**

Structural Quality

☑ All variable names from the source data are present

☑ Variable types match the source data

☑ Data formats are consistent with source data

☑ Relational structure and referential integrity maintained (if applicable)

☑ Value ranges and categorical options align with real data

☑ Coded values use the same coding systems as source data

☑ Overall rate of missing values matches source data

☑ Errors and inconsistencies in the source data are replicated in the synthetic data (if appropriate for the use case)

☑ No new categories of errors introduced by generation process

☑ Temporal consistency maintained such as dates following logical sequences

Statistical Quality

☑ Distribution shapes match source data

☑ Correlations preserved within acceptable thresholds

☑ Subgroup-specific patterns represented

☑ Fairness metrics evaluated and minority populations and rare conditions adequately represented or limitations disclosed

L1 (Structural): Focus on schema, missingness frequency, messiness, plausibility
L2 (Univariate): Add univariate distribution assessment
L3 (Multivariate): Add relationship preservation, consider bias assessment for analytical uses
L4 (Augmented): Complete assessment including mandatory bias/fairness evaluation and validation requirements

Future Work

The development of the VSTAR framework marks a significant milestone in defining the principles for responsible synthetic data. However, the workshops highlighted that defining principles is only the first step; the challenge now lies in operationalising them and increasing adoption. To this end, the UK Synthetic Data Community Group has successfully secured a second round of funding from DARE UK. This next phase marks a strategic shift from recommendations to real-world adoption, focusing on overcoming the practical, regulatory, and cultural barriers that continue to inhibit the widespread acceptance of synthetic data in the UK research ecosystem.

1. Advancing Regulatory Clarity

One of the most significant challenges identified across our workshops was uncertainty around how synthetic data fits within existing regulatory frameworks, particularly GDPR and UK data protection legislation. While synthetic data offers considerable promise for reducing disclosure risk and enabling broader data access, questions remain about when synthetic datasets should be treated as personal data, and what safeguards are sufficient to meet regulatory expectations. This uncertainty can lead to overly cautious governance decisions, inconsistent practices across organisations, and missed opportunities for innovation.

To address this, we will convene a series of focused regulatory workshops bringing together regulatory representatives, legal experts, and information governance teams from across the UK data landscape. These workshops will aim to clarify the application of GDPR principles to synthetic data and develop practical guidance that data custodians can use when making governance decisions about synthetic data. By facilitating structured dialogue between regulatory bodies and data practitioners, we aim to build a shared understanding of acceptable practices, reduce ambiguity, and provide organisations with the confidence they need to deploy synthetic data responsibly and lawfully.

2. Expanding and Refining the Framework for Specific Contexts

To ensure the framework is genuinely suitable across the full spectrum of synthetic data applications, we will host targeted workshops to expand and refine the VSTAR principles for specific contexts.

One priority area is engagement with the NHS and Secure Data Environment (SDE) networks. NHS data underpins much of the UK's health research, and ensuring that synthetic data can be responsibly integrated into NHS data access workflows is essential for realising its potential to accelerate research while safeguarding patient privacy. Through dedicated workshops with NHS information governance leads, analysts, and SDE operators, we will explore how the VSTAR framework can be adapted to meet the specific governance, operational, and trust requirements of the NHS context.

A second priority is exploring how synthetic data can facilitate and accelerate data federation across the UK. Federated analytics and federated learning enable analyses to be conducted across multiple datasets without centralising data, but they often face significant technical and governance barriers.

Synthetic data offers the potential to support federated query development and trialling of new innovative technologies. We will convene workshops with federated data infrastructure leads and researchers to identify how synthetic data can address these challenges, what governance models are appropriate, and how the VSTAR principles can guide responsible use of synthetic data in federated contexts.

Finally, we recognise that the initial phase of this work focused primarily on tabular data. However, many research applications involve more complex data modalities such as neuroimaging, genomics, and free-text clinical notes, each of which presents unique challenges for synthetic data generation, privacy evaluation, and utility assessment. We will host workshops focused on these complex data types to understand their specific requirements, explore appropriate generation and evaluation methods, and refine the VSTAR framework to ensure it provides meaningful guidance for these high-dimensional and sensitive data modalities.

3. Establishing a Utility Framework

A clear message from our workshops was the need for better evidence on when and how synthetic data delivers real value. Questions remain about what fidelity levels are sufficient for specific use cases and how to evaluate whether synthetic data is truly fit-for-purpose. To address this, we will develop a utility framework grounded in empirical evidence.

Central to this effort will be a "SynthSprint" hackathon-style event, where researchers will test various fidelity levels of synthetic data against different tasks and use cases. Participants will work with synthetic datasets representing different generation methods, applying them to tasks such as code development, exploratory analysis, model training, and federated query design. Through structured feedback and performance evaluation, we will gather evidence on what types of synthetic data are genuinely useful for specific purposes.

The insights from the SynthSprint will inform a comprehensive utility framework linking use cases and required fidelity levels. This will provide data custodians with clear, evidence-based guidance on when to invest in higher-fidelity generation methods and how to assess whether synthetic data is meeting its intended objectives.

4. Standardisation and International Collaboration

We will form an expert working group to develop a Delphi consensus study aimed at establishing community-agreed best practices for generating and evaluating synthetic data. Through structured, iterative consultation, the group will work towards agreement on key questions. This consensus study will help establish a shared foundation of best practices that organisations can adopt with confidence.

We will also actively engage with international groups such as the Australian Synthetic Data Community of Practice to exchange insights, compare regulatory approaches, and explore opportunities for harmonisation. This international collaboration will ensure UK efforts remain informed by global best practices and contribute to the broader conversation on responsible synthetic data use.

Acknowledgements

We would like to extend our sincere gratitude to DARE UK for funding this work and enabling the UK Synthetic Data Community Group to advance the responsible use of synthetic data in Trusted Research Environments.

This report is the result of a truly collaborative effort, and we are deeply thankful to the 143 participants from 38 organisations who contributed their time and insights across our five stakeholder workshops. To the researchers, data custodians, synthetic data experts, and members of the public who engaged in these discussions, thank you for helping to shape the VSTAR framework. We also offer special thanks to our invited speakers for generously sharing their technical expertise and real-world implementation experiences, which provided the essential context for our findings.

We acknowledge the exceptional leadership and oversight provided by our Co-Chairs: Lewis Hotchkiss, Emma Squires, Emily Oliver, Sophie McCall, Cristina Magder, Anmol Arora, Fiona Lugg-Widger, Robert Trubey, Steve Moore, Timothy Rittman, John Gallacher, and Simon Thompson.

Finally, we wish to thank the internal project team, including Lois Owumi, Pawel Darda, Kafayat Adeoye, and Alieyeh Sarabandi, for their dedicated support in the delivery of this project.

Organisations Involved in these Workshops:

Administrative Data Research UK
Alan Turing Institute
Brunel University London
Cardiff University
Dementias Platform UK
Energy Systems Catapult
GSK.AI
Health Data Research UK
Imperial College London
Information Commissioners Office
LifeArc
MHRA – Clinical Practice Research Datalink
NHS England
Public Health Scotland
Public Health Wales
Queen Mary University of London
Queens University Belfast
Research Data Scotland
SAIL Databank
Scottish Government
Scottish Longitudinal Study
Swansea University
UK Biobank
UK Data Service
UK Government Department for Education
UK Government Department for Justice
UK Government Department for Work & Pensions
UK Longitudinal Linkage Collaboration
University College London
University of Birmingham
University of Bristol
University of Cambridge
University of Edinburgh
University of Essex
University of Exeter
University of Manchester
University of Oxford
Welsh Government

Funded by

DARE UK

For further information on this report or the Community Group, you can visit our website: www.syntheticdata.uk, or contact Lewis Hotchkiss (lewis.hotchkiss@chi.swan.ac.uk)