International Journal of Population Data Science

Journal Website: www.ijpds.org

What Are the Useful Characteristics of Synthetic Data? A Qualitative Study of Researcher Perspectives

Lora Frayling[1], Elizabeth Pattinson[2], and Emily Oliver[3,*]

Submission History

Submitted:	19/12/2025
Accepted:	28/08/2026
Published:	17/11/2026

[1] Health Data Insight CIC, Camlife 4, Cambridge Road, Fulbourn, Cambridge, CB21 5XE, UK
[2] Administrative Data Research UK (ADR UK), ESRC, Polaris House, North Star Avenue, Swindon, SN2 1UJ, UK
[3] Economic and Social Research Council (ESRC), Polaris House, North Star Avenue, Swindon, SN2 1UJ, UK

Abstract

Introduction
When access to real data is restricted by confidentiality requirements and lengthy approval procedures, researchers may consider ways to support their work prior to seeking access, such as using synthetic data. This is often the case with administrative datasets that contain sensitive information about individuals' interactions with government systems, including health, education, justice, benefits and income. While such datasets are highly valuable for research, they are typically securely held and strictly controlled. Synthetic data is increasingly being used to support the use of real-world data across various research activities. To optimise synthetic data creation and use, it is essential to define and assess its characteristics and understand how they relate to utility. While frameworks exist for understanding real data quality and utility, approaches to synthetic data characterisation remain limited, with a predominant focus on statistical fidelity.

Objectives
Our study aimed to identify which characteristics of synthetic data are important for utility from researchers' perspectives, the contexts in which they see benefits and what is needed to enable optimal use.

Methods
Drawing on real data quality frameworks, we conducted interviews with the researchers to explore their opinions and experiences using secure real data and synthetic versions. We interviewed 19 researchers with experience using secure administrative or health data for research, of whom 18 had some experience using synthetic data. Interviews were analysed qualitatively to identify key themes.

Results
We uncovered four areas of interest: (1) barriers to real data access, such as discoverability and governance; (2) versatility of synthetic data use cases, including research planning, education and collaboration; (3) dependencies affecting adoption, such as governance, cost, quality, suitability for use cases and safety; and (4) an underdeveloped ecosystem due to uncertainty in synthetic data usefulness, a lack of clarity on use and limited available information, which reduces confidence in synthetic data provision.

Conclusions
Based on our findings, we propose an initial synthetic data characterisation framework emphasising the relationship between synthetic and real data and the context of intended use. We recommend that synthetic data is developed with clear use cases, meets quality requirements, is well-documented, accessible and discoverable and is provided safely under proportionate governance. Embedding mechanisms for improvement, such as feedback loops and regular review, would further optimise the value and benefits of synthetic data for both providers and users.

Keywords
synthetic data, PET, privacy enhancing technology

*Corresponding Author:
Email Address: emily.oliver@esrc.ukri.org (Emily Oliver)

https://doi.org/10.23889/ijpds.v11i1.3406
November 2026 © The Authors. Open Access under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en)

Introduction

With digital data being routinely collected and advancements in computing, there is increased potential for data-driven research and innovation that benefits society. This is particularly relevant for administrative and healthcare data, which are collected for operational and management purposes and transformed by data holders into research-ready datasets that can provide valuable insights for healthcare, social science and public policy research [1]. However, such data often includes personal information and needs to be handled appropriately when used for research purposes, in line with UK GDPR and the Data Protection Act 2018, as well as relevant data protection frameworks. Consequently, these datasets are not typically accessible to researchers without appropriate ethical permissions, researcher accreditations and project-specific approvals, which can be time-consuming and difficult to obtain, posing significant challenges for researchers [2].

For these reasons, researchers are increasingly seeking access to synthetic data (SD) to use in place of real data (RD). SD is typically designed to support a variety of use cases across different research activities, such as data discovery, education, developing analytical code and running analyses [3]. However, balancing data privacy concerns with a need to support a range of researchers' use cases presents challenges.

Existing approaches to evaluating the usefulness of SD are typically limited to technical proxy measures, such as statistical similarity to RD or performance on specific downstream tasks [4–6]. While these metrics provide useful quantitative assessments, they do not account for the broader context in which SD is used. In contrast, RD is typically assessed with this broader context in mind, reflecting the research purpose and setting in which it is used. To our knowledge, no generalisable framework exists for comprehensively evaluating SD utility across diverse contexts of use. Without such a framework, data custodians may have difficulty in designing and providing SD that meets researchers' needs effectively and researchers may be uncertain whether a given synthetic dataset is suitable for their intended use.

Thus, developing a synthetic data utility framework is essential for broader adoption but first requires an understanding of how SD can benefit researchers. In our study, we make progress towards this end by:

1) Leveraging RD quality frameworks to investigate synthetic data characteristics

2) Studying SD use by interviewing researchers who have experience of using securely held RD and may have a need for SD

3) Suggesting a draft framework for SD characteristics that inform utility.

To contextualise our work, we consider the relationship between RD and SD and key concepts such as fidelity and utility.

Background – The Case for Synthetic Data

Researchers often face the challenge of assessing whether a real dataset will be suitable for their use case without being able to see what it contains [4, 5]. Data holders often provide data documentation to assist with this assessment [6]. However, for complex datasets it is difficult to capture all the essential data characteristics that give an indication of data *utility*, i.e., how well the data supports their use case [7]. Without direct access to RD, users must discover data properties by liaising directly with the data holder or apply for access without guarantee that the data is sufficient, both of which can be costly in terms of time and resources.

Synthetic Data Design and Use

SD can address these challenges by preserving key structural and statistical characteristics of real-world datasets without retaining disclosive information. Van Hoorn (2024) identifies a range of use cases across different stages of healthcare innovation including research (data exploration, method comparisons, replicable research), development (software, user, machine learning model testing) and deployment (demonstrations, user training). SD can also support workflows in which researchers develop code against the synthetic data, which is then executed on the real data by analysts employed by the data holder, who return the results following privacy checks. This approach, used with datasets such as the Simulacrum[1] and platforms such as OpenSAFELY[2], enables researchers to perform analysis on RD without requiring direct access [8, 9].

Beyond research preparation, SD has broader applications, for example in education, where researchers can gain practical experience with realistic data and running analyses without going through lengthy data access processes [10]. In machine learning, SD can augment limited real-world data, addressing class imbalance and bias in training datasets to improve model prediction and fairness [11]. In some cases, SD may even be used in place of RD to gain real-world insights [12]. These varied applications illustrate that SD has broad and flexible use cases.

Existing literature typically discusses SD utility in terms of *fidelity*, i.e., the extent to which SD preserves the characteristics of RD. In practice, fidelity may encompass multiple structural and statistical properties, however literature tends to focus on general statistical fidelity measures, such as similarity of data variable distributions, with an implicit assumption that higher fidelity directly translates to greater utility [13]. In addition, *task-based performance* is often assessed as a practical estimate of utility for specific analytical or machine learning tasks, thereby anticipating usefulness for similar applications [14]. While these approaches can offer valuable insights, they may overlook broader factors that impact the overall usefulness of SD.

For these reasons, we conducted a targeted user study to understand the key characteristics of SD that can enhance utility for researchers.

[1] https://simulacrum.healthdatainsight.org.uk/.
[2] https://www.opensafely.org/

Framework for Categorising Real Data Characteristics

To understand broader considerations for understanding relevant characteristics of SD, we draw on existing real data quality (RDQ) frameworks for evaluating *real data qualities*, i.e., characteristics that inform its utility.

Wang et al. (1996) identify four categories of data qualities that are important from the user's perspective: (1) inherent qualities of the data itself, regardless of its intended use, (2) representational qualities that relate to how the data is represented and understood, (3) contextual qualities relative to the specific use case and (4) accessibility [15]. The software-centric framework proposed by the International Standards Organization (ISO) more explicitly distinguishes system-dependent qualities that relate to the environment in which the data is accessed [16]. The Health Data Research (HDR UK) data utility framework categorises datasets across different dimensions such as technical quality, data documentation and accessibility [17].

In relation to synthetic data provision, we propose four types of RD characteristics that inform utility, building on these frameworks. We define inherent characteristics that can be measured directly from the data and we distinguish them from characteristics that depend on the full use context that includes the use case, the stakeholders involved and the system in which the data is accessed:

- **Inherent Characteristics:** Characteristics measured directly from the dataset that may be indicative of utility, e.g., accuracy (how close data values are to the truth), completeness (the proportion of values or records present) and uniqueness (absence of duplication of records).
- **Contextual Characteristics:**
  - **Use-dependent**: Characteristics that are evaluated relative to a specific use case, such as *appropriateness* (how suitable a dataset is for the intended use in terms of its structure and content)
  - **Stakeholder-dependent:** Characteristics that reflect the interaction between users, data providers and the data itself, such as *credibility* (whether values of the data are believed to be true)
  - **System-dependent**: Characteristics that relate to the technical and organisational environment in which data are stored, managed and accessed, such as accessibility, the usability of supporting infrastructure and quality of supporting data documentation.

User Study Towards a Synthetic Data Utility Framework

To explore SD utility, we asked participants to identify SD characteristics they considered important for its usefulness. Three qualitative studies have taken a similar approach, interviewing participants about SD created to address limitations of existing real data, as shown in Table 1.

Fornaca, C. (2022) interviewed eight participants on the potential use of SD to support clinical research and found a general lack of familiarity with SD that limited findings [4]. Van Hoorn, R. (2024) interviewed 24 experts on the utility, acceptance and adoption of SD for machine learning-based healthcare innovation, identifying broad use cases across the research workflow and a need to assure SD quality [18]. Key characteristics included fidelity and usability for downstream tasks to improve trust in utility, while validation on real data was considered essential. Rosenblatt, L. et al (2024) interviewed nine participants to understand the contexts in which SD could substitute for RD and what makes it useful. Similar concerns about validation were raised, however other applications that support use of RD were encouraged as a focus for future provision and work [5].

These studies demonstrate a growing interest in SD for research, but each is anchored to specific data types, use cases or SD generation methods, limiting the transferability of their findings. Our approach is deliberately agnostic to these aspects and grounded in RDQ frameworks as a foundation for reasoning about user perspectives of synthetic data (Table 1), enabling us to identify characteristics that hold across SD applications rather than those tied to a single context.

Grounding our work in RDQ frameworks reflected our expectation that characteristics relevant to SD utility would align with the categories we identified for RD, while remaining open to discover new characteristics specific to SD use and the relationship between synthetic and real datasets.

Method

Study Design

Our study aim was to learn about researchers' perspectives on and experiences with SD that are related to securely held real-world datasets. The research questions were informed by the RD characteristics categories outlined in the Background section. Specifically, we addressed:

R1) What challenges do researchers experience when accessing and using RD for research?

R2) What specific types of research activities could be supported by the use of SD?

R3) What are key characteristics that make SD useful for different research activities?

R4) How do factors such as the discoverability, accessibility and governance of SD impact its practical use?

To address these questions, we conducted a qualitative study using semi-structured interviews with researchers experienced in using securely held data.

Participant Recruitment

Participants were recruited through an open invitation circulated via SD-related networks, ensuring we interviewed researchers with an interest in or need for SD. Members of the ADR UK-led Synthetic Data Working Group reached out through their networks to identify volunteers. Eligible participants indicated prior experience in using secure administrative datasets for research.

Table 1: Summary of previous qualitative studies that investigate perspectives of researchers about synthetic data through interviews in comparison to our study

Paper	Method-agnostic	Application-agnostic	Domain-agnostic	Study informed by real data quality framework
Fornaca, C. 2022	**Yes**	No (Clinical research)	No (Healthcare)	No
van Hoorn, R. 2024	**Yes**	No (Machine Learning innovation)	No (Healthcare)	No
Rosenblatt, L. et al 2024	No (Differentially private)	No (Instead of real data)	**Yes** (No specified domain)	No
Our paper	**Yes**	**Yes**	**Yes** (Administrative data, no specified domain)	**Yes**

Volunteers willing to be interviewed completed a short survey capturing their prior experience with secure RD and SD and their career stage. Responses on prior SD use were used to assign participants to one of two versions of the interview guide. In total, 19 people responded and took part in the study.

Interview Design

Interview questions were informed by the research questions and aligned with categories of characteristics that inform RD utility, as described in Table 2. Two versions of the interview questions were created that align with the key interview categories: one for participants who initially self-identified as having used SD previously, and one for those who had not. The questions and interview guide were refined in consultation with the Synthetic Data Working Group whose members came from across government, academia and industry.

Using the interview guide and following practice interviews, four members of the working group conducted the interviews according to their availability. All interviews were held online, recorded and transcribed using Microsoft Teams.

Our full interview guide can be found in Supplementary Appendix 1.

Qualitative Analysis

Aligned with the exploratory objectives of the study, we applied a qualitative thematic analysis of the interview transcripts using standard qualitative analysis techniques, as proposed by Braun, V. & Clarke, V. (2006) [19].

First, the interviewers undertook pre-processing of the data, de-identifying transcripts (removing the participant names) and verifying transcription accuracy. Any incorrect words or acronyms were corrected. The cleaned transcripts were then imported into NVivo software (UKRI-licensed, Version R1/2020) where further organisation of the data was undertaken. This included numbering, classifying the

Table 2: Interview question categories

Question category	Topics covered	Characteristics type
Introductory questions	Warm up questions about affiliation.	N/A
Real secure dataset access	Experiences with real data and barriers faced with access.	System-dependent
Synthetic data use	Synthetic data use contexts, including benefits and limitations for those use contexts and whether there is expectation to get real data access.	Use-dependent
Synthetic data use support	Resource provision alongside the synthetic data.	System-dependent
Synthetic data characteristics	Synthetic data characteristics that may be important for participants, including similarity to real data.	Use-dependent Stakeholder-dependent
Synthetic data access	Perspectives and preferences for synthetic data discovery, mode of synthetic data access and synthetic data governance procedures. Expectations of researcher responsibility when using synthetic data.	System-dependent
General questions	General recommendations for future provision. Perspective on privacy risks from synthetic data.	General

Figure 1: Distribution of participants by sector

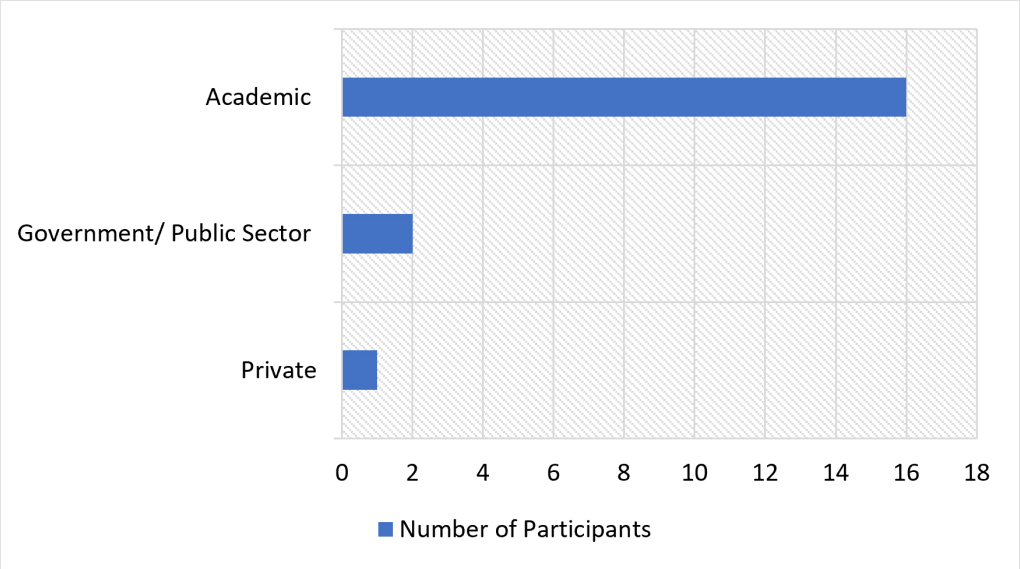
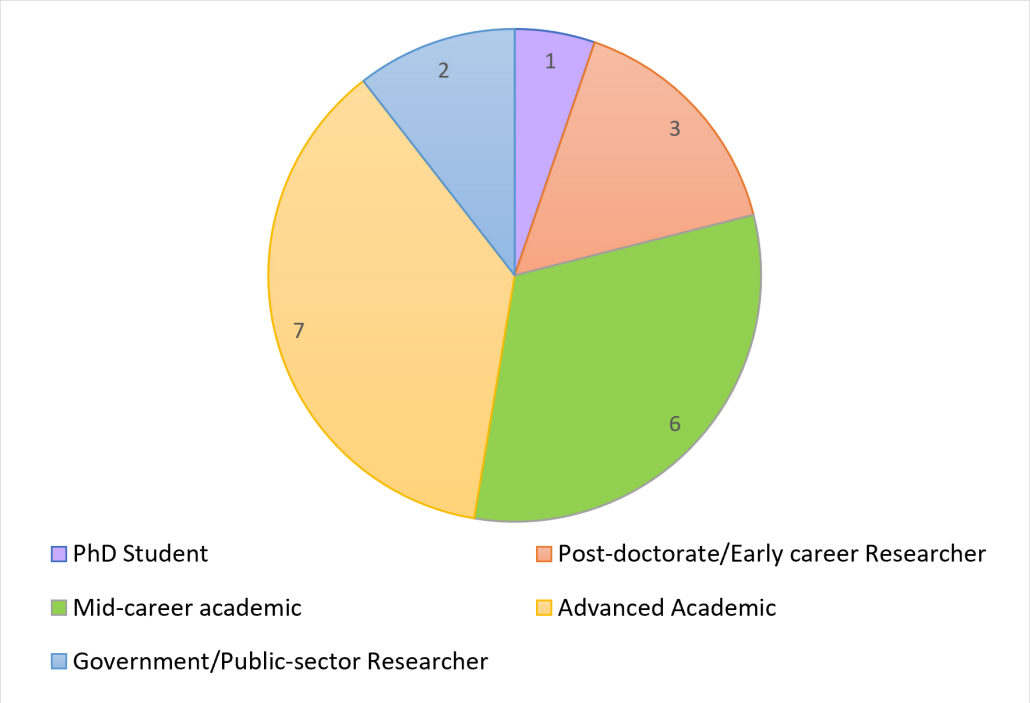


Figure 2: Career stage of participants, as identified through initial survey



participants by affiliation, career stage, and by their self-reported experience of using synthetic data.

We used an inductive thematic analysis approach to code the transcript. This involved the nuanced separation of sentences into meaningful units, assigning initial codes that captured both meaning and sentiment, and then grouping these codes into categories and themes. This approach ensures that themes emerge from participants' responses, rather than being constrained by the framing of the research questions.

The initial codes were developed by the lead analyst [EP], who then sorted codes into sub-categories and grouped codes by general topic. Two other members of the research team [LF and EO] reviewed all codes in the sub-categories to attain consensus of the coding interpretation, seek clarity and refine sub-categories where necessary. Once agreement was reached, sub-categories were then aggregated into broader categories and overarching themes.

The final interview codebook contained seven sections which, between them, captured all final category codes derived from categorising the original codes. Examples are provided in Supplementary Appendix 2.

Ethics

Ethical approval for the study was sought but deemed unnecessary following the completion of a self-assessment process approved by the UK Statistics Authority for the study

design. Approval was obtained from the UKRI Data Protection Team for the sharing of data between organisations involved in analysing the interview transcripts (approval reference UKRI-DSA-176). All participants provided informed consent prior to interviews, and participation was voluntary with assurances of confidentiality of their collected data.

Results

Of the 19 participants, 14 participants stated in the initial survey that they had used SD before while five stated they had not. However, through interviewing, we found that four non-SD users did, in fact, have some previous experience with synthetic data and nearly half of SD users had experience of creating SD themselves (8/19). Thus, the majority (18/19) in fact had some experience using SD.

From the initial survey, we also summarised the sectors that participants came from (Figure 1) and their career levels (Figure 2). During the interviews, all participants confirmed their occupation and that they had previous experience of using administrative or health data for research. The types of data they had used belonged to a variety of domains, including education, social care, hospital, economic and criminal justice data from the UK and other nations.

Overview of Themes

Our thematic analysis of the transcripts generated four overarching themes that capture the perspectives of participants on real and synthetic data use. In the following sections, we describe each theme in detail, including its associated sub-themes. We reference the number of participants who contributing to the related codes being discussed as a fraction of the total number interviewed (n=19), providing insights into how many shared specific views.

The four themes and their sub-themes are summarised in Table 3.

Theme 1: Barriers to Real Data Access

This theme covers the breadth of challenges that researchers encounter when accessing securely held real-world datasets. All participants confirmed that they had used or attempted to use securely held datasets. When asked about whether they had faced barriers, responses overwhelmingly revealed significant challenges, whether in relation to learning about the datasets, navigating the governance process or managing contextual barriers to physically accessing the data (18/19). Illustrative quotations are presented below:

> *"....it can be opaque that the application procedure needs to follow a particular sequence... it takes months and months because you need to wait for one to have been approved, before you could move on to the next step."*

> *"...you really don't understand all the data [...] you have to design your whole analysis without really understanding what data is available."*

> *"It will take so long for the data access to come through that it's not attractive for funders..."*

Data Visibility and Information

From the early stages of attempting to access secure RD, participants reported struggling to discover what datasets were available and what was in those datasets (10/19). They also commented on the inconsistent quality of support resources such as data catalogues and metadata. This impacted their capacity to learn about the datasets which affected both their ability to write feasible project applications and their confidence that the quality of the data was suitable for their needs. As one researcher explained:

> *"I think probably the main barriers are discoverability of the datasets and knowing that they exist and finding out kind of what sorts of data they contain so you can work out whether or not you can actually use them for the kind of research that you want"*

Navigating Governance and Application Processes

Several participants reported that the governance and the application process for secure RD were significant hurdles. Participants complained of excessive pape rwork and the "*back and forth*" with data controllers, describing application processes as "*opaque*" and hard to understand. Participants expressed their reliance on colleagues with previous experience of accessing a specific dataset. They desired clearer application processes with more explicit guidance to help manage their expectations.

Most explicitly associated with RD access processes were the implications this had for the researcher's timeline (12/19), particularly if they had a time-limited research grant. Participants perceived that the lengthy timescales for processing their applications were due to limited staff availability in government departments (5/19). While they appreciated the need for strict governance due to disclosure risk, they identified multiple associated challenges with this (8/19), such as increased time to access more granular data, institutional concerns around being accountable if data access controls were breached by a researcher and participants feeling they needed to overcome data owners' sense of distrust.

Varied Provision and Processes

Participants' experiences differed by contextual circumstances, of example, if access was limited to physical, rather than online, secure environments. In some instances, participants were restricted by geography and the need to access RD at a safe location, e.g., a safe pod[3]. They reported difficulties in getting to these locations, which generated further time and travel costs. Finding available appointments to use the safe locations was also a challenge. As a participant explained:

> *"...the fact that remote access is so limited in Northern Ireland and that the majority of research has to be done within the actual safe settings which can cause a lot of issues in terms of travel time, booking available slots, you know, et cetera."*

[3] https://safepodnetwork.ac.uk/

Table 3: Summary of overarching themes and sub-themes

Theme	Sub-theme	Summary
Barriers to accessing real data	Data visibility and information	Challenges with data discovery, documentation quality and feasibility assessments before applying for data access.
	Navigating governance and application processes	Complicated governance processes with insufficient guidance and support from data holders.
	Varied provision and processes	Challenges across modes of data access including geographical locations and TREs.
Versatility of synthetic data	Synthetic Data to support real data use	Clear use cases for SD prior to RD access, including data familiarisation, research planning, feasibility assessments, code writing.
	Value of Synthetic data on its own	Useful for educational training, but with limitations for replacing RD in decision-making.
	Collaboration	Enables collaboration across teams and organisations while maintaining patient privacy.
Dependencies for synthetic data adoption and utility	Access	SD should be more accessible and lower cost than real data, with appropriate governance and easy modes of access.
	Quality	SD characteristics should be tailored for different use cases and should include high-quality documentation.
	Safety	Appropriate governance should mitigate disclosure risks and protect against misuse.
Underdeveloped ecosystem	Varying usefulness	Experiences showed varying usefulness, often with insufficient characteristics for use cases and needs for SD modification.
	Lack of clarity and information	Absence of guidelines, checklists and clear support routes to data holders reduce utility of SD.
	Lack of knowledge and uncertainty	Limited user experience and knowledge mean preferences for SD are guided by instincts rather than knowledge.

In other instances, participants experienced easier access when their university had secure remote access agreements in place with a trusted research environment (TRE). This allows researchers to work elsewhere as long as they use a safe computer[4]. Some participants discussed access to data through TREs, i.e., secure computing environments with approved software where the data can be accessed remotely under strict controls. More experienced users stated that they had no access problems through TREs themselves; and the UK Data Service was cited as an example of being an easy and transparent TRE to use.

Theme 2: Versatility of Synthetic Data

The second block of interview questions moved the discussion to synthetic data rather than real data. We sought to understand how SD can be or has been used to support research activities. The participants we interviewed were overwhelmingly supportive of SD provision and identified multiple use cases (Figure 3).

Synthetic Data to Support Real Data Use

All participants identified scenarios where SD could support them in using RD. Their comments framed SD's value as an informational resource that indirectly enabled the use of RD by enabling them to become familiar with RD characteristics (14/19), assess feasibility of research (10/19) and conduct research planning for grant proposals (12/19). As one researcher explained:

> *"...the synthetic data might help me to realise that administrative data source isn't suitable for what I want to do and that would be useful."*

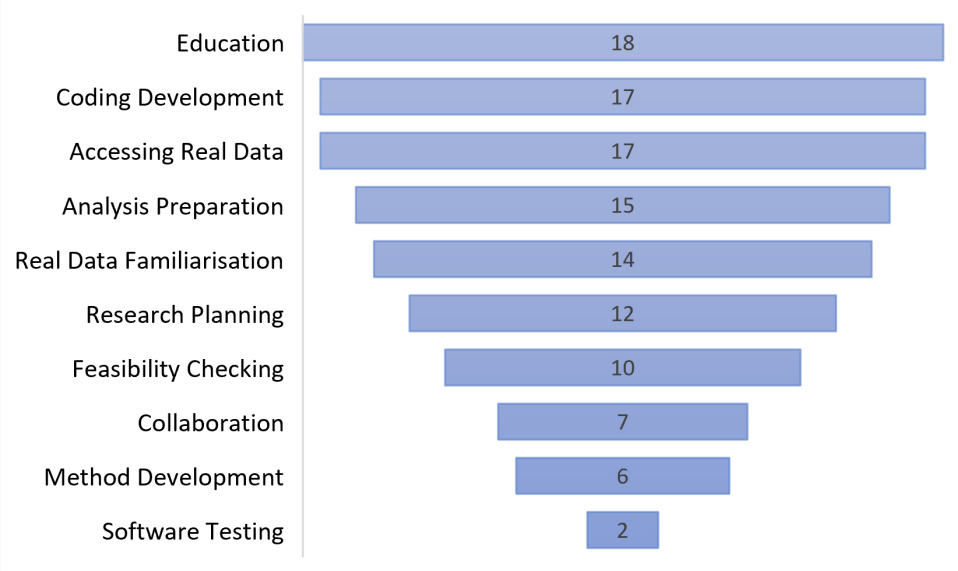
The majority of participants indicated a preference for SD to be signposted or provided as a resource alongside RD (15/19), as this researcher articulated:

> *"I would say it would be great if basically any kind of [...] dataset that's held by any public organisation, so all the health, all the government datasets, anything that can be used in research should have metadata and a kind of sample low fidelity dataset you can download and play around with."*

Participants also discussed examples where SD could specifically support them in accessing RD (16/19), whether in supporting them through application processes or in aiding them in preparing their research while waiting for access. They highlighted SD as a useful pre-application step to get access to the data or to speed up the process, for example, by streamlining discussions with data owners by

[4] https://cy.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/accessthedatasecurely

Figure 3: Use cases for SD identified by participants



quickly identifying data relevant to their research. Most participants discussed the potential for developing analytical code (17/19) and preparing their analyses (15/19). This was seen as particularly valuable in preparation for using RD in TREs, i.e., shadow coding, or for learning skills by simulating scenarios.

Value of Synthetic Data on its Own

Participants identified use cases not necessarily tied to the eventual use of RD. Educational uses were the most commonly cited (18/19). These included teaching, instructing or training another party, irrespective of context skills like coding, interactive learning and experimentation. As one participant explained in the context of teaching using General Practitioner (GP) data in a course:

> *"In the longitudinal [data] community, people are used to what are effectively a whole series of cross-sectional files [...] with GP data. It's a very complex sequence of events which are clustered around visits to GPs [...]. We need to train people how to interpret that and then how to bring that together with the cross-sectional data and make sense of it in longitudinal analysis. You don't need real data for that. You need realistic data and you don't need to go through a full TRE application process to get hold of it with all the governance and costs and barriers in it contains. I think it's really important to generate data for training."*

Another use case identified was methodological development, including applying and testing models, either using SD based on RD or where RD does not exist, and for sharing and demonstrating methods that are being developed (6/19). One participant mentioned an example of SD that was only partially based on existing data but augmented to capture more granular and complete geographies.

When asked about expectations for going on to use the RD, many of the participants did not expect this (8/19). For those who envisioned using SD on its own, a higher level of fidelity was frequently mentioned as a prerequisite for this use case, especially if replacing RD for research. Those with experience of this tended to mention cases where existing infrastructure had made the RD inaccessible.

Collaboration

Participants described collaborative uses of SD, such as for introducing less experienced researchers to using data (7/19), Additionally, participants described several examples where they recognised SD as a vehicle for direct collaboration (7/19). An effective model for this was described which enables the researcher to prepare analytical code which they can send to the data holder, who runs the code on the real data and delivers results to the researcher, thus bypassing the need for direct access. Examples included the Simulacrum, a synthetic cancer dataset, and OpenSAFELY, a secure analytics platform which provides SD to researchers to write their code [8, 9]. A participant clarified:

> *"What OpenSAFELY does is let you analyse sensitive data in the secure environments, but without you actually looking at any of the data itself or having direct access to or eyes on the data itself. And so you write code which then gets sent to the secure environment, and then you can request the results of that code to be released if it's non-disclosive [. . .]."*

In these use cases, participants relayed that the purpose of SD was closely linked to protecting privacy and was a means to provide an 'air gap' between analysts and the RD. Other collaborative uses identified included code sharing, reproducing methods and sharing data potential with external stakeholders.

Theme 3: Dependencies for Adoption and Utility

Participants discussed characteristics that would support high utility and broad use of SD.

Accessibility

Overwhelmingly, participants emphasised the need for SD to be easily accessible and, importantly, more accessible than secure RD. Participants made a connection between governance and the barriers it could create, favouring minimal governance (11/19). They also identified that the access process should be easier for SD than RD (9/19) and suggested they would be discouraged from using SD otherwise. One researcher commented:

> *"In general [...] if [...] practical access to synthetic data was made quite difficult, on the same level as actually the real data set, I think it could be quite off putting."*

Expectations around accessibility were also reflected in participants' references to the mode of SD access. In contrast to their experiences with real data access, participants framed expectations of SD access to be straightforward and quick (17/19). They expected it to be downloadable (12/19) and/or accessible remotely (11/19). They proposed that it should be open access or requestable via email (6/19). Accessing SD through secure environments, such as sidepods or TREs, was described as barrier more often than a preference.

The third way this theme emerged was in the discussion of costs and a willingness to pay for SD (10/19). This was often context-dependant, for instance if the provider or user was from the commercial sector or if there was a guarantee of the SD meeting sufficient quality standards for use. Nonetheless, participants still expected it to be cheaper than whatever they might expect to pay for the real data (although it is also worth noting that not all TREs charge for access to RD). These mirrored concerns about paying without a usefulness guarantee (4/19). Wider financial constraints imposed by limited institutional funding was acknowledged as a barrier.

Quality

Key quality requirements identified included the need for documentation (17/19) such as data dictionaries and metadata. Some participants stressed the importance of having documentation consistent with the RD (6/19), while others felt a need for clarity about how the SD had been created (8/19) and a way of ensuring or judging fidelity for themselves.

In terms of characteristics, 'realism' was a unanimous requirement discussed by all participants. For use cases related to RD use, this was largely described as similarity to the RD, i.e., fidelity, though this requirement was articulated in different ways and often mentioned using vague terminology. As one participant stated:

> *"It needs to look like the real data you know as well as in terms of how comprehensive it is as much as possible."*

Participants mentioned various types of fidelity that are important, which we have categorised as follows:

- Structural fidelity (15/19), i.e., how closely the SD structure is to the real dataset, e.g., having matching variables, same data types and content
- Statistical fidelity (14/19), i.e., how similar the statistical properties are, whether marginal statistical distributions, correlations or more complex multivariate distributions
- Fidelity to RD characteristics, such as:
    - Messiness and completeness of RD (8/19), i.e., how well the SD reflects errors and the proportion or missing values in the RD
    - Semantic and logical correctness (4/19), i.e., how well the SD reflects the logical and structural relationships between variables in the RD and matches its coherence, even if the RD itself contains inconsistencies.

Use cases related to education, e.g., for training and learning purposes, also referred to fidelity, however the emphasis was on retaining generic characteristics of real data that enabled teaching specific skills.

Over half of participants referenced reasons and/or advocated for low fidelity SD, i.e., SD that has low similarity to the RD (11/19), as this is easier to release from a governance perspective and still useful for specific use cases.

Participants acknowledged that there was inevitability going to be variation in the characteristics they would need (13/19), citing dependence on use cases, subject discipline and research requirements. They also referenced the desire for or potential value of varied levels of SD fidelity to cater to individualised needs.

Safety

Participants highlighted the need for SD to be safe to use. They identified a range of appropriate governance measures. While it was clear that they had varying levels of knowledge of this topic, conversations identified a need for scalable governance appropriate to the level of disclosure risk of the SD (13/19). Participants went on to stress the importance of modifying SD governance based on characteristics: primarily, higher governance for higher fidelity or sensitive data and fewer restrictions for lower fidelity and less sensitive data. However, this sentiment was presented in conjunction with a desire that governance measures ensuring safety should be balanced with usefulness and accessibility wherever possible (9/19).

Participants also expressed a desire for processes to assure SD is used for the right purposes. Examples included: verifying researcher identity and uses prior to provision; requiring consent from the data controller to share SD with others; and the establishment of generation and use criteria for each synthetic dataset. A potential misuse mentioned by participants was the misrepresentation of SD as RD. Suggested solutions included a requirement that researchers clearly communicate the synthetic nature of the data, enforced by clear terms of use, upfront disclaimers and guidelines for presenting findings based on SD.

Some conversations with participants included discussion about their own role in enabling safe use. Some acknowledged their own responsibilities, either in reference to their inherent ethical responsibility as researchers or 'standard' data responsibilities. On the other hand, there were still important risks that participants did not have a clear expectation or resolution for, such as accidently regenerating real individual-level data. There was a lack of insight around issues of sample size, the use of AI and anonymisation risks, all of which affect levels of trust in SD safety.

Theme 4: Underdeveloped Ecosystem

Participants' experiences and opinions had been shaped by the newness of the landscape.

Varying Levels of Usefulness

Despite their overwhelming support for SD, all participants indicated limitations for the uses cases they had cited, which included hypothetical scenarios and actual experiences, describing them in different ways. For example, one user highlighted where it had failed to be "*fit for purpose*" for their specific use cases, while others described needing to modify the SD to their needs before it was useful. The overall lack of available synthetic datasets was also mentioned. Largely, issues related to the lack of characteristic similarity that the SD bore to the RD (15/19). Participants discussed scenarios where lack of fidelity caused limitations for multiple use cases. Both lack of structural and statistical fidelity was cited as the cause of problems for analytical code development, as it can mean code fails to run on the real data. This is particularly the case if the SD does not preserve real data quality issues, as the user cannot take these into account when preparing code for data cleaning. Lack of statistical fidelity also caused participants difficulty in interpreting analytical outputs from the SD, in particular when they do not align with a user's logical and domain knowledge: in some of these cases, trust in the SD for their use case diminished quickly and they needed to get in touch with the data provider to understand whether implausible values were signs of fidelity issues or properties of the real data. One participant said:

> *"A survival by deprivation level where you expect that people with higher deprivation have lower survival that was reproduced by the dataset but not by age group, which was a bit strange. Actually, we sent some emails to the people in charge of Simulacrum and, and we had some discussion about this point."*

This preference for realism was stressed by participants, along with a sense that SD was a 'limited response' to replacing or alleviating real data access challenges (7/19). As one researcher described:

> *"I think the problem's actually deeper rooted, so we wouldn't need synthetic data in the first place if the access, the linkage, the creation of admin data would be easier".*

> *"It is so difficult to access it [real data] and I think synthetic data is kind of, to some extent, maybe firefighting a symptom of what is going on rather than being the solution. [...]"*

Lack of Clarity and Information

Participants described a lack of supporting resources surrounding SD, particularly general guidelines or checklists for users (11/19). They mentioned a lack of support resources (9/19), including visibility of SD information (e.g. lack of adverts, dedicated online spaces) and unclear routes for feedback or improvement of the SD.

Where participants flagged specific resources they had used, these tended to be varied, citing metadata, general papers on SD, use of open-source real data and colleagues or contacts with expertise. Many of the participants who flagged the lack of resources also considered existing ones insufficient to resolve specific tasks (7/9), such as extracting subsets from SD, which greatly impacted the utility of SD for participants' use cases:

> *"I didn't get a data library or metadata with the synthetic data and then when I requested one, the version didn't match the data I had the like the variable names and things. So, then my question was does the synthetic data actually match, in terms of variable names, the contents in the secure environment? Because otherwise what I'm doing is pointless and I wasn't really given sufficient reassurances around that."*

Lack of Knowledge and Uncertainty

There was a pattern of participants feeling uncertainty and being aware of their own lack of experience when it came to aspects of SD provision (14/19). Participants' sentiment on this was multifaceted, spanning discussions on payment, governance, the importance of fidelity, accessibility, provision and the potential for it to meet the needs of future use cases. This came through in conversations across all levels of experience. For example, one participant recognised their "*lack of understanding*" could be guiding their assumption that SD *"should be very easy to access*". Yet, SD users expressed uncertainty about SD owners' "*ambitions or intentions*". As a more experienced participant commented:

> *"Two years ago, I was working for an organisation who just finished a project on doing a feasibility study of synthetic data for a government department, and that was two years ago and nothing's happened [. . .] it's just confusing to me because as I said, [I could] put together a synthetic data set in a couple of hours. Like, 'What's the hold-up?' is the kind of question for the data owners really."*

Participants acknowledged not knowing enough about disclosure risk and responsibility for security of the data (7/19). The effect meant that personal preferences for SD were seemingly influenced by intuitive senses of what was safe. One participant advocated for low fidelity data largely due to the perception that SD made from RD could not be safe. Meanwhile, another was less concerned about higher fidelity

as they considered it to be the data owner's responsibility to guarantee that necessary permissions are in place.

Discussion

Our research was motivated by notable gaps within the literature on SD provision and use. In particular, we identified a lack of studies that examine SD characteristics relevant to utility across diverse use cases or that provide a framework for characterising utility grounded in researchers' perspectives.

Our study offers new insights that can begin to address what researchers value and need. It provides helpful points to standardise practice and the quality of future provision.

Synthetic Data Use

Our results suggest that SD has the potential to effectively support researchers across a broad range of research activities, including feasibility checking, developing analytical code in advance of RD access and for collaborating across teams. SD can also serve wider needs, such as providing accessible educational resources or tools for stakeholder engagement. These use cases are in line with previous SD literature and interview studies [3–5, 18]. In most examples given by participants, the SD had not been designed specifically for the use cases at hand, but for more general use. However, participants did indicate examples, such as the Simulacrum dataset, where a researcher-centred approach to its design has increased its utility.

Governance and Risk

Participants highlighted the importance of proportionate governance for SD with respect to its level of fidelity and disclosure risk. This is interesting as, at the time the study was conducted, there were no consistent standards for evaluating SD disclosure risk [20]. New governance frameworks have since emerged that can support SD providers to develop a consistent approach for doing so [21, 22]. We expect this, along with ethical and quality standards, will become established with a growing body of work moving towards establishing best practice for synthetic data generation and use, such as those by Raab et al. (2025) and Oliver et al. (2025) [23, 24]. Neither is there consistent provision of high-quality documentation provided alongside SD. This gap has been flagged by our participants and makes sense as it is a routine expectation for the provision of real data in research contexts [1].

Fidelity and Utility

Beyond governance and access, participants elaborated on data characteristics that providers should prioritise to ensure optimum usefulness. Most agreed that core structural fidelity, i.e., retaining the same data format, variables and relational structure, is essential for utility when supporting research with real data. For some participants, having this structural similarity alongside minimal statistical similarities, often referred to as 'low fidelity', was sufficient and preferable where it enables quick and safe access others emphasised the need for more detailed aspects of statistical fidelity, such as preserving distributions, rules, domain knowledge, and realistic errors in the data. This suggests that fidelity has multiple, richer dimensions than is typically discussed in existing literature, which focus on structural and statistical fidelity [13].

When discussing educational use contexts, such as skill development and learning, participants emphasised a need for 'realistic' data. In this case, the emphasis was not on fidelity to a specific real dataset but rather on the SD reflecting certain general properties of real data or real concepts, such that it can afford specific skill development.

The range in SD characteristics needed across different uses demonstrates the importance of providers defining clear intended use cases and selecting appropriate SD generation methods to ensure that resulting data is fit for purpose.

Ecosystem Factors

Despite the potential of SD, participants' experiences highlighted that the current landscape is not yet meeting researchers' needs and is not yet well enough understood to inform their demands. The undercurrent of uncertainty around SD poses a barrier to shared understanding of its utility, particularly where questions remain about what a given dataset can and cannot support, what researchers can expect from it and whether those expectations can be met safely. This uncertainty is compounded by the common experience of SD being created and made available without sufficient user support and by the inconsistent routes through which researchers discover available datasets and develop the expertise to use them. As a result, the broader adoption of SD, and the opportunity to establish good practice, is inhibited until there is greater clarity around availability of SD and how it can still be useful to researchers in an ethical way.

While these general gaps require a wider effort of inquiry and research, data holders can begin to address other dataset-specific uncertainties, by providing a consistent set of supporting resources around each synthetic dataset they release.

Synthetic Data Characteristics Framework

Our study provides insights that can inform the development of a comprehensive framework for SD utility. Earlier in this paper, we proposed inherent and contextual characteristics of RD, where the latter includes aspects that are use-dependent, stakeholder-dependent and system-dependent. These help us to understand RD and how it can be used.

We found that some aspects of SD are more complex than for the real data. This is because, while SD might support various use cases, its features need to be tailored to each use case to be useful. While RD are assessed with respect to real-world phenomena, SD are based on the RD or a real-world concept as the primary reference point. Its quality is considered against the real data it originates from.

When SD is used to support research on RD, the primary objective is to match key features of the RD, such as accuracy, completeness, consistency, validity, structure and statistical distributions, so that participants can meaningfully assess real data's potential utility. As already mentioned, we found the notion of SD fidelity is much richer and nuanced than typically

Table 4: Synthetic data characteristics framework

Synthetic data characteristic type		Description	Examples
Inherent	Fidelity	Fidelity to the real data for specific properties	Structural fidelity; Statistical fidelity; Relational fidelity; Semantic fidelity; Fidelity to real data characteristics
	Realism	Realistic representation of real-world processes or plausible dataset characteristics when real data are unavailable.	
	Enhancement	Deliberate divergence from real data properties that improves utility beyond what the RD offers	Reducing disclosure risk Improving class balance for ML Improving completeness Improving interpretability Improving usability Increasing sample size
Contextual	Use-dependent	Appropriateness for a specific use case	Feasibility testing Code development Analysis ML training Education
	Stakeholder-dependent	Characteristics whose importance depends on the needs or perspectives of different stakeholders	Credibility/ Reliability Interpretability Plausibility
	System-dependent	Characteristics that depend on processes or resources in the surrounding data ecosystem	Accessibility Timeliness Availability Security/safety Discoverability Governance Cost Data documentation/metadata Data use support Feedback mechanisms

acknowledged in existing literature. However, fidelity to RD is not always the priority. In machine learning, the aim may be to enhance the RD, such as by augmenting it to improve predictive performance or rebalancing data to reduce bias. Similarly, SD can fill completeness gaps in real data, such as by adding missing geographic information.

Thus, we propose similar inherent and contextual categories for describing SD characteristics, as presented in Table 4, but here we define inherent characteristics as those that are measurable directly from the SD with respect to either the real dataset or real-world phenomenon and include *fidelity*, *realism* and *enhancement* as examples. Additionally, we posit that a specific use-dependent category is still included, which we would expect to have characteristics that are defined for specific use cases.

As with RD, we rely on a two-way trust between the provider and user. In the case of SD, we found user trust to be heavily shaped by the fidelity of the data itself, the quality of provided documentation, the user's prior experience and communication with the data holder and SD provider. Trust was lost when users noticed implausible values, and when there was no clear route to challenge the provider and ask for information and explanations. Thus characteristics, such as plausibility, are included under stakeholder-dependent category.

Finally, characteristics dependent on the system in which SD is accessed and used, such as accessibility, documentation and appropriate governance, remain essential as with the RD, and are integral for responsible use and utility of the SD.

Limitations

Our study has several limitations. First, our findings were based on the viewpoints of 19 participants at a single point in time and reflect individual experiences, limiting generalisability to the wider data user population. Although participants had varying career stages, geography, domains of expertise and SD experience, this cannot capture the full diversity of possible synthetic data users.

In addition, participants were recruited through known networks of members of the Synthetic Data Working Group, including the study team, which may bias findings towards particular datasets, SD types and use cases. Voluntarily participation also introduces self-selection bias, implying an existing interest in synthetic data.

Separate interview guides were developed for SD users and non-users. However, during the interviews, 4 of the 5 participants initially classified as non-users reported some prior experience with SD. As a results, participants with SD experience were interviewed using slightly different guides dependent on their initial classification. While this may have introduced minor differences in wording of questions and topics explored, we could still capture similar opinions and experiences across participants. Given the limited representation of those without SD experience, we did not compare the two groups directly in our analysis.

Finally, our thematic analysis findings did not always distinguish between views grounded in participants' real experiences and those reflecting hypothetical scenarios.

Future research should expand the number of participants interviewed and investigate perspectives across different user profiles, use cases and contexts.

Conclusion

In the context of growing interest in and demand for synthetic data, we have found that there is a lack of consistency in both its provision and expectation for use. By taking the user perspective and drawing on interviews with 19 participants, we have unpacked the characteristics that synthetic data needs to optimise usefulness across a range of research activities and we have proposed an initial characterisation framework of SD based on our findings. The framework emphasises both the relationship between the synthetic and real data and the broader context in which synthetic data is intended to be used, providing a stronger understanding of how it can be characterised, evaluated and improved to meet user needs.

Broadly our findings suggest that to optimise usefulness of SD, users expect it to be:

1. **Created with intention** – developed with clear, well-defined use cases in mind
2. **High quality** – achieving the characteristics required to be useful for its intended use case
3. **Well supported** – accompanied by sufficient user support, resources and documentation comparable to real data provision
4. **Accessible and discoverable** – ideally available alongside the real data it is based on
5. **Safe** – explicit about responsibilities, with robust governance and appropriate user registration in place
6. **Improvable** – designed with mechanisms in place for feedback and continuous improvement.

We urge readers to consider these as guidance for future development of synthetic datasets. Weaving them into policies for provision through increased discourse between SD providers and users will help to raise standards and build trust, as well as open opportunities for more testing and use. This can complement other emerging guidance that is informed by perspectives from different stakeholders within the ecosystem, including data holders and the public [25, 26].

Future areas of work should explore SD utility across specific use cases and develop a SD utility framework that systematically maps out characteristics while keeping contextual factors in mind. Such a framework would enable more robust evaluation, more consistent provision and, ultimately, increase the adoption and benefits from synthetic data use.

Acknowledgements

We are grateful to the Synthetic Data Working Group (names available on request) for their support, to our interview team (Amy Dunlop, Carly McAllan, Lora Frayling, Emily Oliver), and to the interview participants who freely gave up their time to speak to us. Thanks also to Dr Kamran Soomro for his advice and support throughout the analysis and writing of this manuscript and to Emma Frayling for her guidance on thematic analysis.

Statement on Conflicts of Interest

None declared.

Ethics Statement

Ethical approval for the study was sought but deemed unnecessary following the completion of a self-assessment process approved by the UK Statistics Authority for the study design. Approval was obtained from the UKRI Data Protection Team for the sharing of data between organisations involved in analysing the interview transcripts (approval reference UKRI-DSA-176).

Data Availability Statement

Data is not available because consent to share interview transcripts publicly was not provided.

AI Disclosure Statement

The authors used Microsoft 365 Copilot for grammar checking and phrasing assistance. The output was reviewed and verified by the authors.

References

1. McGrath-Lone L, Jay MA, Blackburn R, Gordon E, Zylbersztejn A, Wiljaars L, et al. What makes administrative data research-ready? : A systematic review and thematic analysis of published literature. Int. J. Popul. Data Sci. [Internet] 2022 June [cited 2025 24];7.

Available from: https://ijpds.org/article/view/1718. https://doi.org/10.23889/ijpds.v7i1.1718

2. Moorthie S, Hayat S, Zhang Y, Parkin K, Philips V, Bale A, et al. Rapid systematic review to identify key barriers to access, linkage, and use of local authority administrative data for population health research, practice, and policy in the United Kingdom. BMC Public Health 2022 Dec.;22:1263. https://doi.org/10.1186/s12889-022-13187-9

3. Kokosi T, De Stavola B, Mitra R, Frayling L, Doherty A, Dove I, et al. An overview on synthetic administrative data for research. Int. J. Popul. Data Sci. [Internet] 2022 May [cited 2025 13];7. Available from: https://ijpds.org/article/view/1727. https://doi.org/10.23889/ijpds.v7i1.1727

4. Fornaca CA. Enabling Synthetic Data Usage for Medical Research [Internet]. 2022 [cited 2025 16]; Available from: https://escholarship.org/uc/item/82r9v8sc.

5. Rosenblatt L, Howe B, Stoyanovich J. Are Data Experts Buying into Differentially Private Synthetic Data? Gathering Community Perspectives [Internet]. 2024 Dec. [cited 2025 16]; Available from: http://arxiv.org/abs/2412.13030. https://doi.org/10.48550/arXiv.2412.13030

6. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 2016 Mar.;3:160018. https://doi.org/10.1038/sdata.2016.18

7. Karr AF, Sanil AP, Banks DL. Data quality: A statistical perspective. Stat. Methodol. 2006 Apr.;3:137–73. https://doi.org/10.1016/j.stamet.2005.08.005

8. Kafatos G, Levy J, Jose S, Hindocha P, Archangelidi O, Vernon S, et al. Leveraging Synthetic Data to Facilitate Research: A Collaborative Model for Analyzing Sensitive National Cancer Registry Data in England. Ther. Innov. Regul. Sci. 2025 Sept.;59:919–28. https://doi.org/10.1007/s43441-025-00820-z

9. Nab L, Schaffer AL, Hulme W, DeVito NJ, Dillingham I, Wiedemann M, et al. OpenSAFELY : A platform for analysing electronic health records designed for reproducible research. Pharmacoepidemiol. Drug Saf. 2024 June;33:e5815. https://doi.org/10.1002/pds.5815

10. Kuo NI-H, Perez-Concha O, Hanly M, Mnatzaganian E, Hao B, Di Sipio M, et al. Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project. JMIR Med. Educ. 2024 Jan.;10:e51388. https://doi.org/10.2196/51388

11. Shahul Hameed MA, Qureshi AM, Kaushik A. Bias Mitigation via Synthetic Data Generation: A Review. Electronics 2024 Oct.;13:3909. https://doi.org/10.3390/electronics13193909

12. Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. JMIR Med. Inform. 2020 Feb.;8:e16492. https://doi.org/10.2196/16492

13. Frayling L, Bharat SS, Pattinson E, Stock J, Lugg-Widger F, Gordon E, et al. A Review of Synthetic Data Terminology for Privacy Preserving Use Cases. Int. J. Popul. Data Sci. [Internet] 2025 Oct. [cited 2025 24];10. Available from: https://ijpds.org/article/view/2967. https://doi.org/10.23889/ijpds.v10i2.2967

14. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. JMIR Med. Inform. 2022 Apr.;10:e35734. https://doi.org/10.2196/35734

15. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manag. Inf. Syst. 1996 Mar.;12:5–33. https://doi.org/10.1080/07421222.1996.11518099

16. ISO/IEC 25012:2008: Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model [Internet]. 2008; Available from: https://www.iso.org/standard/35736.html.

17. Gordon B, Barrett J, Fennessy C, Cake C, Milward A, Irwin C, et al. Development of a data utility framework to support effective health data curation. BMJ Health Care Inform. 2021 May;28:e100303. https://doi.org/10.1136/bmjhci-2020-100303

18. van Hoorn RD. On the Acceptance, Adoption, and Utility of Synthetic Data for Healthcare Innovation.

19. Braun V, Clarke V. Using thematic analysis in psychology. Qual. Res. Psychol. [Internet] 2006 Jan. [cited 2025 24]; Available from: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa. https://doi.org/10.1191/1478088706qp063oa

20. Hyrup T, Lautrup AD, Zimek A, Schneider-Kamp P. A systematic review of privacy-preserving techniques for synthetic tabular health data. Discov. Data 2025 Dec.;3:1–32. https://doi.org/10.1007/s44248-025-00022-w

21. Inverarity C, Snaith B. Towards the ethical use of synthetic data in health research [Internet]. The Open Data Institute; 2025 [cited 2025 26]. Available from: https://theodi.hacdn.io/media/documents/Towards_the_ethical_use_of_synthetic_data_in_health_research_2025.pdf.

22. Arora A, Hotchkiss L, Myles P, Oliver E, Magder C, Comerford P, et al. Synthetic Data Release Framework under UK Data Protection Law [Internet]. Zenodo; 2026 [cited 2026 29]. Available from: https://zenodo.org/doi/10.5281/zenodo.20535817. https://doi.org/10.5281/ZENODO.20535817

23. Raab GM, McCall S, Cavin L. Four checks for low-fidelity synthetic data: recommendations for disclosure control and quality evaluation [Internet]. 2025 [cited 2025 13]; Available from: https://arxiv.org/abs/2503.14211. https://doi.org/10.48550/ARXIV.2503.14211

24. Oliver E, McCall S, Bremner R, Boyd A, Magder C, Myles P, et al. Considerations for the provision of synthetic forms of secure data: An ADR UK Synthetic Data Working Group Guidance paper. 2025 Dec. [cited 2025 19]; Available from: https://zenodo.org/doi/10.5281/zenodo.17980237. https://doi.org/10.5281/ZENODO.17980237

25. Magder C, Haaker M, Kasmire J, Zahid H, Ogwayo M, Economic and Social Research Council, et al. The Role of Synthetic Data in Research: Benefits, Costs, and Practical Insights from Data Owners and Trusted Research Environments Experts - A Project Report with Practical Recommendations [Internet]. Zenodo; 2025 [cited 2026 29]. Available from: https://zenodo.org/doi/10.5281/zenodo.15191271. https://doi.org/10.5281/ZENODO.15191271

26. Nollett C, Lugg-Widger F, Stock J, Brookes-Howell L, Fitzgibbon J, Johnson S, et al. Discussing Data: A UK-wide public consultation on the use of low fidelity synthetic data for research. Int. J. Popul. Data Sci. [Internet] 2025 Nov. [cited 2026 30];10. Available from: https://ijpds.org/article/view/2962. https://doi.org/10.23889/ijpds.v10i1.2962

Supplementary Appendix 1

What are the useful characteristics of synthetic data?

Interview guide

Study: Researcher Perspectives on Synthetic Data

The aim of the study is to learn about the perspectives of researchers on synthetic data. Specifically, we want to:

(a) Understand how synthetic data can be and/or has been used to support research

(b) Identify the characteristics that make synthetic data useful

(c) Assess how the discoverability, accessibility and governance of synthetic data impact its use.

Researchers who have used secure forms of data for research have been recruited as participants in the study and will be interviewed so that we can learn from their opinions and experiences. There are two types of researchers we will be speaking to: those who have used synthetic data before and those who haven't.

Interview plan

Prior to the interview, the participant will be sent (a) a briefing sheet and (b) a consent form that they should have signed and returned.

There will be 1 hour allocated for the interview, split into 3 sections:

- Pre-interview briefing (10 mins): time for the participant to get onto the call and be given a quick introduction to the study.
- Interview questions (30-45 min)
- Post-interview de-brief (5 mins): thank them for their time and explain what will happen next – send a follow-up email with a couple of questions.

Pre-interview briefing (10 mins)

1. Welcome interviewee and introduce yourself
2. Checks
    a. They have read the briefing sheet
    b. Quickly as a reminder, introduce the study again and the interview structure
    c. Reiterate that all transcripts will be held confidentially, and all outputs of the study will be anonymised
    d. Are they happy for us to start recording and proceed with the interview?
3. Start recording and begin the interview (questions below)

Example:

Hello, thank you for agreeing to be interviewed today. I'm [insert interviewer's name] and I am [describe self]. I'm helping to conduct the interviews as part of this study.

First, I would like to double check that you have read the briefing sheet we sent over. If not, could you take a moment to read through it.

As a reminder, this study aims to learn about researchers' opinions on synthetic data. Synthetic data is artificially created and usually based on a real dataset that is securely held and not easily accessible due to privacy concerns. Researchers can use synthetic data in many ways to support their research. We would like to learn from you today about your opinions and experiences.

Just to reiterate, with your permission, the interview will be recorded so that we can collect transcripts of the conversation. These will remain confidential, and all outputs of the study will be anonymised.

We expect the interview to take up to 45 minutes. Do you have any further questions before we begin?

Let me know if you are happy to proceed with the interview and I will start recording.

[Starts recording session]

Research questions overview (30-45 mins)

The questions are split into 7 categories:
[NOTE, you need two paths of questions: for those who have used the synthetic data before and those who have not.]

Question topics for those who have used synthetic data

- Introductory questions:
  - These are some introductory questions to establish a rapport with the interviewee
- Data use:
  - This section is to get a feel for which datasets they value and would like/need to get access to
  - What are the barriers for accessing these datasets?
    - Is there a cost associated with the data?
    - Is time an issue? Does data use require special skills, or regulatory permissions?
- Utility and support:
  - This section is to get a feel for which synthetic datasets they have used, why and how they have used them and their general experience
  - Do they expect to get access to the real data in the end? If so, how beneficial is it to conduct research using synthetic data? What aspect of synthetic data use has the highest benefit?
  - Also, whether there is any information, documentation / support they received alongside use and what they would find helpful.
- Characteristics
  - This section naturally leads on from asking about the use of synthetic data
  - Here we want to find from them what characteristics/properties of the synthetic data are useful for each of the use cases they have mentioned
- Discoverability
  - This section is asking about how easy they find it to discover synthetic data
- Access
  - Once they have actually found the synthetic data, how easy is it to actually access it?
- Governance
  - These questions relate to what they believe their responsibilities are using synthetic data and their opinion about whether there should be governance around synthetic data.
- Conclusion
  - Some general questions and final remarks

Questions

Questions for those who have used SD	Notes/prompts. . . .only use if needed
Introductory questions	
1. Just to start off, could you please tell us your name, so we have a record for the transcript?	
2. What is your affiliation? (Organisation/university)	
Data use	
3. Which secure datasets are you most familiar with? (up to 3)	
4. What barriers have you come across for accessing these datasets?	Costs Time Special data skills Regulatory permissions

Continued

Questions Continued

Questions for those who have used SD	Notes/prompts. . . .only use if needed
Utility and support	
5. You indicated that you have used synthetic data in the past. What kind of synthetic data did you use and why did you use it?	
6. How have you used it? *(Prompt only if needed: Examples might be learning about a dataset, data analysis training, planning research, doing research)*	Might get several answers here. Make a note of each one and ask them to expand on each one separately
7. Please can you expand on what you have used it for .... *(ask this for each of the uses they mention, eg training, planning research, testing code etc)*	No need for this if they have expanded above.
8. How was your experience using it for each of these purposes? (*Just include the ones they mention*).	
9. Did you expect or intend to get access to real data after using the synthetic data? If so, how beneficial is it to conduct research having the synthetic data to use first?	
10. What aspect of synthetic data use has the highest benefit?	
11. Did you have any issues or challenges when using the synthetic data? What were they?	
12. Was there anything available to help you deal with those issues? If so, did you use them and were they helpful?	
13. If you were going to use synthetic data, what would you want to be given to help you use it?	Prompt if needed: Metadata (variables, scope, time...) User guide Data dictionary
Characteristics	
14. What are the important characteristics of SD for you?	(could be size, scope, similarity etc)
15. Do you think it's important for the synthetic dataset to match the real one closely?	
Discoverability	
16. How did you discover the synthetic data you used? How easy was it to find?	
17. In your experience, how does the discoverability of synthetic data compare to real data?	
18. How easy do you think it should be?	
Access	
19. What is your experience of getting access to synthetic datasets?	
20. How long did it take to access the synthetic dataset?	
21. Would you expect to pay for access to a synthetic dataset?If so, how much would you expect and be willing to pay?	
22. In terms of physical location, where did you access the synthetic data and did you find the practical access convenient?	
23. Were there any other hurdles to accessing the synthetic dataset?	
24. What are your preferences regarding access to synthetic data?	
Governance	
25. What are your views on governance procedures when accessing synthetic data?	Such as signing 'conditions of use' forms
26. How do you think governance affects accessibility?	
27. What do you believe your responsibilities are when you have access to synthetic data?	
28. Have you encountered any guidelines or checklists provided for researchers using synthetic data?	If yes, please share them
Conclusion	
29. What recommendations would you make for the future provision of synthetic data (for research)?	
30. Do you have any thoughts on the real or perceived risk of disclosure associated with synthetic data?	
31. Any other points to add? Is there anything else you want to tell us about your experience.	

Question topics for those who have not used synthetic data

- Introductory questions:
  - These are some introductory questions to establish a rapport with the interviewee
- Data use:
  - This section is to get a feel for which datasets they value and would like/need to get access to
  - What are the barriers for accessing these datasets?
    - Is there a cost associated with the data?
    - Is time an issue? Does data use require special skills, or regulatory permissions?
- Utility and support
  - Do they think that synthetic data can be useful for them for doing research on these datasets?
  - Will the real data still be used in the end? If so, how much benefit is in using the synthetic data and how could that benefit be achieved?
- Characteristics
  - This section naturally leads on from asking about potential benefits of using a synthetic version of the data
  - Here we want to find from them what characteristics/properties of the synthetic data would be essential to attain.
- Distribution and access
- This section is asking about how the data would be best distributed and used. Governance
  - These questions relate to what they believe their responsibilities are using synthetic data and their opinion about whether there should be governance around synthetic data.
- Conclusion
  - Some general questions and final remarks

Questions for those who have NOT used SD	**Notes/prompts. . . .only use if needed**
1. Please tell us your name (for the transcript) 2. What is your affiliation? (Organisation/university)	
Data use 3. Which secure datasets are you most familiar with? (up to 3)	
4. What barriers have you come across for accessing these datasets? (Add prompts if they can't think of anything)	Costs Time Special data skills Regulatory permissions
Utility 5. You indicated that you have not used synthetic data in the past. Why not?	
6. If you were to use it, do you have any thoughts on what you might use it for? (Prompt if needed: Examples might be learning about a dataset, data analysis training, planning research, doing research)	Might get several answers here. Make a note of each one and ask them to expand on each one separately
7. Please can you expand on these . . . . (ask this for each of the uses they mention)	Only if you think more info is needed here
8. How effective do you think it might be for each of these purposes? (Just include the ones they mention). 9. Would you expect to get access to real data after using the synthetic data? If so, how beneficial would it be to have used the synthetic data to use first? 10. What aspect of synthetic data use do you think has the highest benefit? 11. Can you imagine any potential issues with using synthetic data?	
12. If you were going to use synthetic data, what sort of things might you want to be given to help you use it?	Examples if needed: Metadata (variables, scope, time...) User guide Data dictionary

Continued

Questions for those who have NOT used SD	Notes/prompts. . . .only use if needed
Characteristics	
13. What would you say are the important characteristics of SD for you?	
14. Do you think it's important for the synthetic dataset to match the real one closely?	
Discoverability	
15. Have you ever tried to find synthetic data? If so, how easy was it to find what you needed? If not, where might you expect to find it?	
16. How easy do you think it should be to find it?	
Access	
17. How long would you expect it to take to access a synthetic dataset?	Immediate Less than a week 1-2 weeks More than 2 weeks
18. Would you expect to pay for access to a synthetic dataset? If so, how much would you expect and be willing to pay?	
19. In terms of physical location, where and how would you expect to access a synthetic dataset?	Eg At home, in a safe pod, in a safe room at your university
20. Are there any other hurdles you think you might encounter when accessing a synthetic dataset?	
21. What are your preferences regarding access to synthetic data? -*only need this if they haven't already answered*	
Governance	
22. What are your views on governance procedures when accessing synthetic data?	Such as signing 'conditions of use' forms
23. How do you think governance affects accessibility?	
24. What do you believe should be the responsibilities of a researcher should be when they have access to synthetic data?	
25. Have you encountered any guidelines or checklists provided for researchers using synthetic data?	If yes, please share them
Conclusion	
26. What recommendations would you make for the future provision of synthetic data (for research)?	
27. Do you have any thoughts on the real or perceived risk of disclosure associated with synthetic data?	
28. Any other points to add?	

Post-interview de-briefing (5 mins)

That concludes the interview, so now I'll stop recording.

[Stops recording]

Thanks so much for giving us your time - it was really interesting to hear what you had to say. Just a bit of information on next steps: we will send a follow-up email just to debrief, with a couple of questions about whether you are happy to be contacted in future or if you know anyone else we could interview.

Is there anything you'd like to ask before we finish?

Thank you and goodbye!

Supplementary Appendix 2

Example reference	Base code	Sub-theme code	Category	Thematic code	Key Theme
1. Majority of the interview transcripts segmented into concise sections (1–3 sentences) to preserve the overall meaning. Text with no research relevance (e.g., participant names or identification remarks) was excluded from coding	2. Base codes created, representing detailed meaning and sentiment from the sections of the transcript.	3. Base codes were organised by overarching sub-themes, identifying where codes overlapped or demonstrated nuanced distinctions	4. Sub-theme codes were consolidated under broader topics to facilitate systematic review and organisation.	4. Sub-theme codes were assigned to overarching thematic codes, with multiple thematic labels applied where necessary to reflect overlapping themes.	6. Key themes were derived through a comprehensive review of all thematic codes and categories, highlighting sub-theme codes where needed to provide more detailed insights
Uh says lots of forms to fill out. You have to engage legal back and forth, and then there's the technical side as well that needs to be set up. but it it's the high cost, you know, in terms of time at the time and effort sort of thing, yeah.	High Time cost of in inefficient exchanges in RD access	Administrative and Bureaucracy Barriers	Barriers to Real data	Barrier to Real Data use	**Theme 1: Barriers to real data access** **Navigating Governance and Application Processes**
Reference and base code were organised into 2 sub-themes codes but only discussed in 1 Key theme.		Time			
So from my perspective the synthetic data, it's very nice is you can try out things and especially if you're a student, you can you can get a feel of large data set and things like that.	Educational Value of SD for learning and experimentation	Synthetic data for practical learning	Use cases	Synthetic data for education	**Theme 2: Versatility of synthetic data** **Value of Synthetic data on its own**
think it should be fairly straightforward. Well, you would hope so because you would think that has to be the point of synthetic data to make access easy or easier than it is to get access to the real data.	Expectations for Simplified Access to Synthetic Data Compared to Real Data	Synthetic Data- Easier to access in Comparison to Real Data	Access	Easy access to Synthetic Data	**Theme 3: Dependencies for adoption and utility** **Accessibility**
it's quite difficult to do if some of the simulated data properties do not match the real [data]. Summaries sort of population or global level private properties.	Difficulty matching Simulated Data Properties to Real Data	Issues using Synthetic Data due to lack of statistical representation	Characteristics	Importance of Realistic Synthetic Data	**Theme 3: Dependencies for adoption and utility** **Quality**
This sub theme was organised into three thematic codes, which are discussed across 2 Key themes.				Statistical fidelity	
				Limitations	**Theme 4: Underdeveloped Ecosystem** **Varying levels of usefulness**