International Journal of Population Data Science

Journal Website: www.ijpds.org

Discussing Data: A UK-wide public consultation on the use of low fidelity synthetic data for research

Claire Nollett[1,†], Fiona Lugg-Widger[1,†], Joshua Stock[1], Lucy Brookes-Howell[1], Jim Fitzgibbon[2], Sean Johnson[1], Kim Munnery[1], Michael Robling[1], Farheen Yameen[2], and Rob Trubey[1*]

Submission History	
Submitted:	18/12/2024
Accepted:	09/09/2025
Published:	27/11/2025

[1] Centre for Trials Research, Cardiff University, 7th Floor, Neuadd Meirionnydd, Heath Park Campus, Cardiff, CF14 4YS UK

[2] Public contributor

[†] Joint first authors

Abstract

Introduction

Synthetic data is an emerging tool for health and social care researchers which can be used to expedite research processes and enhance the safety of personal data handling. Although data owners across the UK and globally currently make synthetic datasets available, public awareness and perceptions of its use for research remain underexplored.

Objectives

We aimed to agree a set of recommendations for data owners on producing and releasing synthetic data and communicating with the UK public about this topic.

Methods

Forty-four public members were invited to four deliberative workshops to co-create a set of recommendations for data owners. Notes and transcripts from workshops 1-3, along with public member survey feedback, were reviewed to identify key topics.

Results

Thirty-seven individuals contributed. We developed ten recommendations spanning five areas: introducing synthetic data; explaining its purpose; creating datasets; managing access, use, and misuse; and clear communication. These were iteratively refined with public input during the final workshop.

Conclusions

This study offers a unique insight into public attitudes towards synthetic data use in health and administrative research. The recommendations developed are intended to support UK data owners in releasing synthetic data responsibly, aligning with public expectations and fostering trust in its use as a research innovation tool. Future research should continue to assess evolving public attitudes to both low and high-fidelity data as its use expands.

Keywords

synthetic data; public attitudes; public deliberation; public recommendations; public patient engagement

*Corresponding Author:
Email Address: trubeyrj@cardiff.ac.uk (Rob Trubey)

https://doi.org/10.23889/ijpds.v10i1.2962

Introduction

Synthetic data can be used to replicate the structure and patterns of real datasets while mitigating the risk of individual re-identification. It is an emerging tool for health and social care researchers in the UK and globally [1–3]. Synthetic datasets allow researchers to perform exploratory analysis, develop statistical codes, and generate hypotheses without the risks or regulatory hurdles associated with accessing real datasets. The fidelity of synthetic data exists on a spectrum, reflecting how closely it mirrors real-world data in terms of statistical properties, data distributions, and utility for downstream tasks. Low-fidelity synthetic data abstracts information to a greater extent, effectively minimising re-identification risks and enhancing privacy protection. Although the utility of the dataset may be reduced in some instances, it can still serve specific purposes such as improving data access requests to accelerate research processes and enhancing the safety of personal data handling. Its potential has been increasingly recognised by data owners such as Research Data Scotland [4], NHS England [5] and the Clinical Practice Research Datalink [6] who already make synthetic datasets available.

Advancements in synthetic data technology have accelerated significantly in recent years. However, despite these developments and calls for better communication on the topic [7], public awareness and perceptions of its use in research remain largely underexplored. Recent efforts to expand the use of synthetic data have focused on the perspectives of researchers and data owners, with less attention given to public opinion. A 2021 report by the Behavioural Insights Team highlighted this lack of public awareness and understanding of key terminologies and emphasised the need for large-scale public consultation to address these gaps and build trust [5]. Although some initial public engagement activities have taken place, a comprehensive examination of public attitudes remains crucial, particularly because synthetic data, though abstracted, is still often derived from real public data [8].

Public consultations empower individuals to play a more active role in shaping policies or recommendations that directly impact them. They foster trust between the public and government organisations and can inspire innovative ideas [9, 10]. Importantly, more in-depth consultations tend to produce more effective and better-informed policies [9]. We present here the first UK-wide public consultation on low-fidelity synthetic data, highlighting the diversity of perspectives on this topic and documenting how these discussions have shaped recommendations aimed at helping data owners and researchers build and maintain public trust in synthetic data.

Methods

Overview and aims

We conducted a public consultation on attitudes towards synthetic data using a series of deliberative workshops. Deliberative workshops allow members to give in-depth consideration to complex topics by working together, and with expert input, to gradually develop an agreed view or list of recommendations [11].

We aimed to agree a set of recommendations for data owners (organisations who provide public data to third parties) on producing and releasing synthetic data and communicating with the public about this topic.

Recruiting public members

Our aim was to engage approximately forty members of the UK public, without prior knowledge or expertise in administrative or synthetic data, divided into two groups of twenty. This approach ensured diverse perspectives while keeping online sessions manageable. Public members were selected to reflect a range of demographic backgrounds.

Recruitment was facilitated by Egality [12], a community engagement agency, who produced the advert through consultation with four community organisations across the UK (Ethnic Health Forum, Lancashire BME Network/Inclusive North, Latin American Community Association of Edinburgh and Race Council Cymru). The advert was disseminated via Egality and additional channels such as Administrative Data Research (ADR) Wales and ADR UK [8], Health Data Research UK Voices [13], NIHR People in Research [14] and Engage [15] in Northern Ireland.

Interested individuals completed an online Expression of Interest form, providing contact details, demographic information, and availability. Selection was anonymised by removing names and aimed to ensure diverse representation (across gender, age, ethnicity, level of education, country of residence and referral organisation) rather than statistical representativeness. We did not request information about prior knowledge of the topic. A 10% dropout rate was assumed when making the final selection of members.

Workshop design and implementation

Over five weeks in June and July 2024, we conducted a series of three interactive online workshops, each offered twice at different times of the day. The series concluded with a single final workshop (in September 2024), to which all public members were invited in order to help refine the recommendations as a group.

Workshops lasted two hours and were held online via Zoom to ensure geographical diversity and accessibility across the UK. This format enabled participation from individuals unable to travel and allowed for the involvement of a broader range of expert speakers. Members provided their consent by email for the use of anonymous quotes and were offered £25 per hour for their contributions in line with national guidelines [16]. Public members chose a daytime or evening workshop time slot but could switch between the two across the first three sessions.

Before the first workshop, a Question & Answer (Q&A) video about the project featuring our two PI partners was shared with all invited members, alongside a written welcome pack. The workshops were facilitated by an experienced public involvement lead (Author CN) with technical support provided by our senior project administrator (Author JS).

In the first session, participants and co-ordinators were randomly allocated to small breakout rooms to get to know one another. Subsequent workshops began with a recap of the previous session, presented by one of the project team (Authors RT, FLW). Workshops then followed a structured format: an informational presentation, followed by small breakout rooms where four to five members discussed the content in small groups. The combination of participants in these groups was changed each session and were facilitated by a member of project team who then fed back a summary to the whole group.

Workshops 1-3 featured short presentations from synthetic data experts, identified through the ADR UK synthetic data working group or by team members (Table 1). To conclude each workshop, members rated their attitudes toward synthetic data using an emoji-based activity, chosen for its simplicity and accessibility to all public members (Appendix 1).

Defining low-fidelity synthetic data: An introduction to the concept of synthetic data was provided through playing a 90 second animation developed by Research Data Scotland [17]. We then invited a subject matter expert to provide more detail which focused on introducing low-fidelity synthetic data. The definition focused on preserving data structure (data columns, types, values, ranges, etc.) and 1-dimensional distribution of columns.

Each workshop was designed to build on the previous one, informed by a review of written notes and transcripts from the prior session, along with input from our two PI partners, Project Management Group (PMG) and Project Steering Group (PSG) (Figure 1). Feedback from members via an online survey after each workshop (Appendix 2) was used to shape subsequent sessions.

Development of recommendations

Facilitator notes and transcripts from workshops 1-3 were consolidated and thoroughly reviewed by one author (CN) who looked for recurring themes across the outputs. Key points under each theme were then identified and a draft list of 11 recommendations under five themes was produced. These recommendations were then refined collaboratively by the project team, including PI partners, before being circulated to workshop members in advance of workshop 4 for discussion. Members who were unable to attend this final workshop were invited to provide input via email. The final recommendations were developed iteratively, incorporating workshop member input, revisions from PI partners, feedback from our PMG and input from the project oversight group. At Workshop 4, members suggested removing two recommendations and adding another (advising data owners to work with the public to ensure all of this information is accessible to people with a diverse range of needs), leaving a final list of 10.

Role of public involvement partners

Two members of the public were part of the core project team. Author JF was known to the team and helped to design the project prior to funding, while author FY joined after funding confirmation. Neither were familiar with synthetic data but both delivered valuable experience as public involvement partners in research.

The aim of their involvement was to ensure that the workshops were accessible to a diverse public audience, in terms of logistics and content, while ensuring the public's voice was consistently represented throughout the project. Key contributions included:

- Participating in weekly meetings to design the workshops, thinking through accessible language, workshop timings, and use of technology
- Drafting and featuring in the Q&A video and writing the welcome pack
- Attending workshop rehearsal meetings
- Facilitating breakout room discussions during the workshops
- Reviewing workshop members' feedback via the survey
- Overseeing project delivery at monthly management meetings
- Assisting in drafting the recommendations and dissemination materials

All input from PI partners was recorded in a simple public involvement log, maintained by the public involvement lead and reviewed or amended as necessary by our PI partners. Remuneration for time was offered in line with national guidelines [17]. We report the role of our PI partners in line with the GRIPP-2 SF [18].

Project governance and oversight

This was a public consultation, aligning with principles of public involvement and engagement rather than research. Following discussions with the Cardiff University School of Medicine Ethics Committee it was confirmed as not requiring ethical approval.

A Project Steering Group (PSG) was established to provide independent oversight and expert input throughout the consultation. The group comprised members with expertise in synthetic data, ethical data practices, and public engagement (including representation from ADR UK, UK Data Service, Research Data Scotland and academic experts). Its role was to advise on the structure, content, and facilitation of workshops, to review and comment on draft and final recommendations, and to guide dissemination and engagement strategies. The group met on three occasions (before the first workshop, during the development of draft recommendations, and at the dissemination planning stage), offering constructive feedback and ensuring that outputs were robust, credible and useful for data owners and custodians. Notes from these meetings were reported back to the project team and management group, with key contributions reflected in the design of workshops, refinement of recommendations, and planned communication activities.

Results

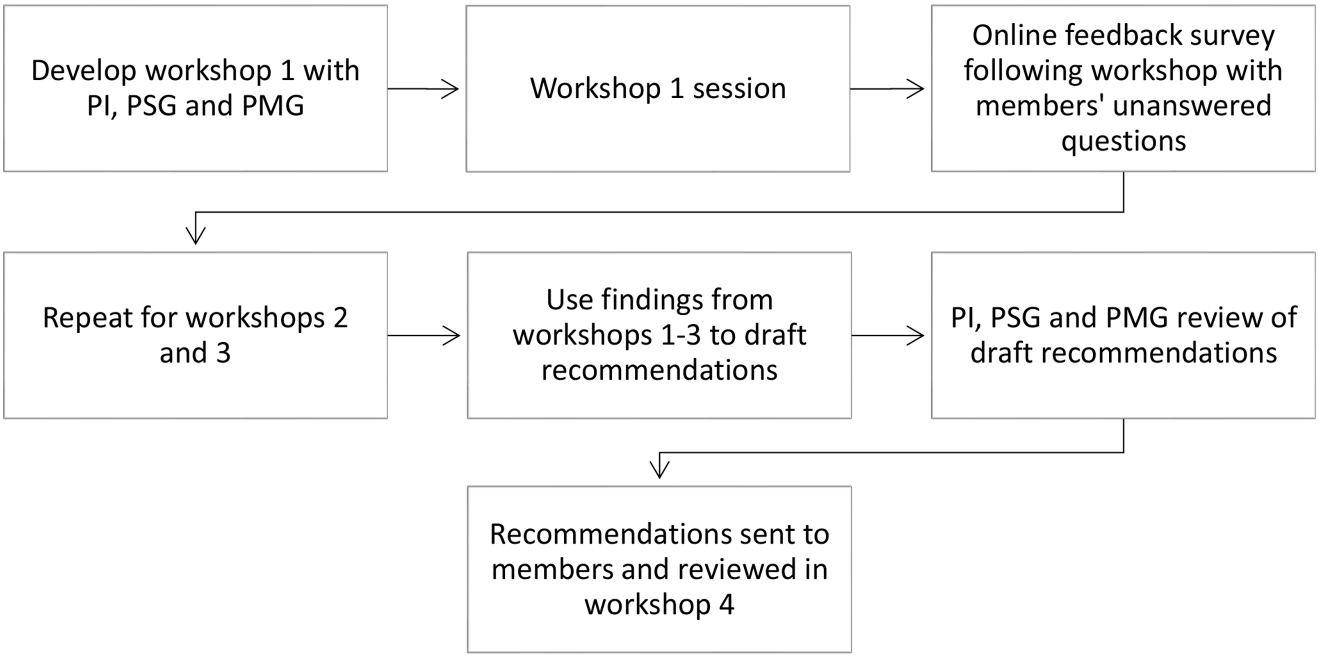
Recruitment of public members

We received 138 expressions of interest, 54 through Egality and their community groups and 84 from a range of other channels. Forty-four individuals were selected based on their demographic characteristics and background information (Table 2) and invited to take part in the workshops. The majority were from England to reflect its larger population

Table 1: Programme of four workshops

Work-shop	Topic	Presentations	Discussion points
1	**Understanding low fidelity Synthetic Data**	1. Scope for Influence 2. Introduction to Public sector data and Synthetic Data 3. How Synthetic Datasets are created	*What questions do you have for the speakers?* *What would you like to know more about in future workshops?*
2	**Using Synthetic Data in research: Benefits and Concerns**	1. Session 1 recap 2. Balancing public benefit & risks of synthetic data 3. Types of, and ways to access, Synthetic Data	*Have your questions from session 1 been answered?* *Which method of access to synthetic data would you be most comfortable with?* *Does this differ by type of data?*
3	**Communicating clearly about Synthetic Data**	1. Session 2 recap 2. Current terminology 3. Information to accompany a Synthetic Data file	*Have your questions from session 2 been answered?* *Do you have a preferred alternative word to synthetic data?* *How would you describe synthetic data to a friend?* *What information should data owners be providing alongside synthetic datasets?*
4	**Review the Draft Recommendations**	1. Recap video (sent in advance) 2. Next steps for the project 3. Our draft recommendations	*Do you agree with these recommendations?* *What refinements would you make?* *Have we missed anything?*

Figure 1: Process of developing the recommendations


Develop workshop 1 with PI, PSG and PMG
Workshop 1 session
Online feedback survey following workshop with members' unanswered questions
Repeat for workshops 2 and 3
Use findings from workshops 1-3 to draft recommendations
PI, PSG and PMG review of draft recommendations
Recommendations sent to members and reviewed in workshop 4

size (n = 28), followed by Scotland (n = 7) and Wales (n = 7) then Northern Ireland (N = 2). Twenty individuals were recruited through Egality's Community groups and 24 via other networks. Seven did not respond or later dropped out, leaving 37 attendees to one or more workshops (Table 3).

Considerations that formed the final recommendations

Based on the notes and transcripts from the first three workshops, we identified eleven key considerations across five

Table 2: Public member characteristics for those invited to contribute (n = 44)

Characteristic	N
Gender	
Man (including trans man)	14
Woman (including trans woman)	26
Prefer to self-describe	3
Prefer not to say	1
Age	
18-24	3
25-34	10
35-44	7
45-54	13
55-64	6
65-74	2
75-84	3
85+	0
Ethnicity	
Arab	3
Asian or Asian British - Indian, Pakistani, Bangladeshi	9
Asian or Asian British - Other	3
Black or Black British - African	6
Latin American	4
Mixed	1
Mixed - Black, African, and White	1
Mixed - Black, Caribbean, and White	1
Mixed - White and Latino	1
Prefer to self-describe	1
White	13
White (but not British or Irish)	1
Formal Education	
GSCE / O-Level / CSE	3
A-Levels / Level 3 BTEC / NVQ	6
Higher Education	33
Prefer not to say	2
Working Status	
Employed - Full time	13
Employed – Part time	9
Not employed	7
Retired	5
Other	7
Prefer not to say	3
Disability	
No	25
Yes	18
Prefer not to say	1

Table 3: Workshop attendance out of the 44 invited

Attendance	N
Workshop 1	37
Workshop 2	35
Workshop 3	33
Workshop 4	24

central topics. Following consultation with our PI partners and discussions in the final workshop, these were then refined to a set of ten recommendations. We present them below, with a summary of the discussions as context for each recommendation.

Introducing synthetic data

The term synthetic data was not well understood by or helpful to our public members who viewed the word as having negative connotations e.g. bogus, fake or mass produced. Many alternative words were suggested by public members (Figure 2) but there was no consensus with people voting both for and against many of these during workshop 3.

In light of this, we were unable to offer a definitive synonym that would be an acceptable alternative to use on public facing materials. Instead, the consensus from the final workshop was to focus on an accessible and easily translatable definition.

Recommendation 1: The term synthetic data is not well understood by the public. Provide a brief definition to explain: - that synthetic data is not real data - but that it is based on real data - that it is created in a way that minimises personal privacy risks

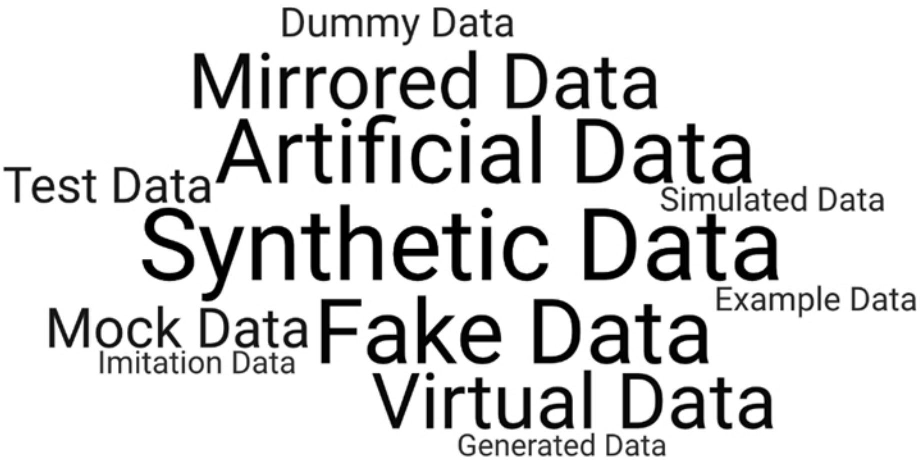
Explaining the purpose of your synthetic data

Some public members questioned the need to create or use synthetic data at all. Their queries centred around whether the process for accessing real data could be expedited and/or whether simply removing personal identifiers from real datasets would be sufficient and more economical time and money wise. Public members asked that data owners give explanations about how the creation and release of synthetic data benefits the public, the organisation and the researchers accessing the data. They could appreciate the advantages to the researcher but, even in the final workshop, were still reflecting on the benefits to research outcomes and people's everyday lives. They also stressed the importance of highlighting what the data can and cannot be used for. This was in part in relation to concerns of insurance companies or media misusing the information but also because the data is not an 'accurate' dataset and should not be used as a replacement to the 'real' data researchers are interested in accessing.

Recommendation 2: Make it clear what your synthetic dataset can and cannot be used for e.g. it can be used for training, or writing computer code, but not to answer research questions. Emphasise that real policy decisions are always made with real data.
Recommendation 3: Explain the benefits and impact of your synthetic dataset for:

- your own organisation
- researchers e.g. training and understanding the data
- the public e.g. if public money is being spent to create it, how do the public benefit?

Figure 2: Word cloud of terms discussed



Recommendation 4: Explain the personal privacy benefits that using synthetic data offers.

Creating your synthetic data

Public members had many questions related to how synthetic datasets are created and then checked before release. A preference for human involvement (as opposed to solely automated processes) in both generation and checking was expressed. Questions arose about the accountability of the person/people involved in the process and there was a sense of wanting to guarantee privacy, security and integrity of the original 'real' dataset. While technical validation protocols may be needed in practice, public members were primarily concerned about oversight and reassurance.

Recommendation 5: Provide a simple explanation for how your synthetic data is created. In particular, you should explain the role of humans versus automation (such as Artificial Intelligence) in the process.
Recommendation 6: Human oversight in checking personal privacy risks is important. Explain the quality and privacy checks you undertake before your synthetic dataset is released.

Access, use and misuse of synthetic data

Many public members viewed public data as power and were suspicious about who would want to access synthetic data and for what gain, as mentioned above. Their concerns centred around malicious misuse e.g. by conspiracy theorists or journalists with their own agenda, hackers who might be able to infer things about the real data set or private organisations such as insurance companies who might profit from using the data in some way. They also had reservations about researchers who might inadvertently use the data to answer real research questions.

Whilst a few members felt synthetic data had sufficient benefits and minimal risks to be made freely available (for example downloading direct from a webpage without the need to provide any contact details), there was not widespread support for, or confidence in, this option. Stemming from concerns about potential misuse, a common view was that it is 'better to be safe than sorry' when allowing access to synthetic data. Most public members were keen on a registration process to enable the data owner to know who is accessing the data and for what purpose. This information could then be summarised on their website enabling transparency about the number of requests and for what purpose which may in itself provide reassurance. A few people advocated for a stringent procedure, however the majority suggested compromising on a simple process: they acknowledged that this was not foolproof but wanted to maintain a balance between benefits and risks. The group recommended that during the process, potential users sign a simple agreement making clear what the data can and cannot be used for (to minimise accidental misuse) and for how long the dataset can be kept (to ensure it reflects the current real data).

Recommendation 7: There is not widespread support for a fully open access approach to synthetic data:

- use a simple registration process which records the requestor's name, email address and intended use (as a minimum).
- implement a simple user agreement covering the key terms and conditions such as allowed usage and how long synthetic data can be held.

Communicating synthetic data clearly

Synthetic data was an unfamiliar concept to most of our public members and definitions and explanations alone were not enough to facilitate good understanding. Public members requested real life case studies which could demonstrate why the researcher needed the synthetic data, how they used it and how they and the public benefited. They also highlighted the need to provide information in a variety of engaging formats such as animations or infographics with minimal text and to consider inclusive communications e.g. for people with sensory impairments, those whose first language is not English, neurodivergence. The materials should be developed with members of the public to maximise accessibility.

Recommendation 8: Real life examples are particularly helpful to the public. Use accessible case studies from researchers to: - demonstrate what synthetic data is - report the outcomes from researchers using your synthetic data - emphasise the positive impact for the public
Recommendation 9: Use a range of communication methods including infographics and engaging videos to convey information about synthetic data to the public.
Recommendation 10: Work with the public to ensure all of this information is accessible to people with a diverse range of needs.

Public member experience and feedback

Twenty-two members completed the online survey after the first workshop, 16 after both workshops 2 and 3, and 12 after the final session. Feedback after workshop 4 was that small discussions in breakout rooms were seen as a particular positive, alongside the time slot for the workshops, the smooth running, clear explanations and use of Zoom which is compatible with screen readers. Points for improvement were mainly around timings and language, with more time for small group discussion and less feedback to the main room, and simple language for people whose first language is not English. One person also asked for the agenda with timings in advance.

The emoji exercise at the end of each workshop provided useful feedback on where members 'were at' with their understanding of and acceptance towards synthetic data. The general trend was from 'exploding heads' at the start, through to 'thinking' and more smiley emojis at the end of the series. This helped the project team to gauge the impact of the workshops and facilitated a discussion with all members.

Discussion

This public consultation examined attitudes toward the use of low-fidelity synthetic data for research, generating recommendations based on knowledge gaps, concerns and preferred communication strategies. Our findings reveal that synthetic data is not widely understood by the public; most public members initially had limited knowledge of its purpose, how it is created and how it has been used by researchers.

Our recommendations emphasise that public-facing communication on synthetic data should begin with clear, accessible explanations that are tailored to a basic knowledge level. These should differentiate synthetic data from real data, highlight existing privacy safeguards, and outline its practical benefits for data owners, researchers and the public. As public members became more informed, they showed cautious support for synthetic data use but consistently sought transparency from data owners on data creation processes, access and intended use cases. This reinforces the idea that clear and accessible communication is essential for building public trust in the use of synthetic data.

The recommendations highlight the importance to data owners of creating public-facing information that extends beyond technical guidance for researchers. This includes providing accessible information that explains the implications and potential value of synthetic data for the general public.

Throughout the workshops, members repeatedly emphasised the importance of building public trust in the creation, use and monitoring of synthetic data. They highlighted the need for clear and transparent communication, particularly through the use of simple language that ensures the public can fully understand the topic. Members also sought reassurance of a transparent access process, to help build trust in researchers and others using synthetic data properly, in line with agreements made.

Although there is little literature on the general public's view on synthetic data, what we discussed in our workshops do align with professional perspectives on this topic. For example, a recent survey of biomedical scientists in the United States reported a lack of knowledge of the term "synthetic data", and a preference for strong oversight in its use, including human oversight. Additionally, 65% of respondents also indicated concerns of synthetic data exacerbating data inequities [19].

Similarities can also be seen with previous work exploring public perspectives on data sharing and the use of routine data in research. For instance, views across the population are diverse, and context plays a critical role in establishing and maintaining trust [20]. As observed in our workshops, concerns about the motivations for accessing and using data are also evident in discussions regarding the use of routine administrative data in research [21, 22]. Muller et al. have discussed the concept of a social licence for data-intensive health research, which appears to extend into the realm of synthetic data. This social licence reflects an additional moral responsibility that goes 'beyond compliance with laws and regulations' [23]. It calls for responsible practices that avoid exploitation and address both actual and perceived risks—an idea that strongly resonates with the discussions in our workshops.

Strengths and limitations

By employing a deliberative workshop approach, we progressively enhanced public members' understanding, fostering more informed discussions and constructive feedback on key issues such as privacy, data security, benefits, and potential risks. This approach highlights the effectiveness of interactive public engagement in capturing public attitudes toward complex topics like synthetic data.

By incorporating feedback from our PI partners, we also made the workshops more accessible to the public. For example, we offered evening sessions, switched from Teams to Zoom for ease of use, and renamed the project from "DELIMIT" to the more relatable "Discussing Data." Following their advice, we included presentations explaining how synthetic data is created and provided real-world examples of its benefits for researchers. They also guided us in addressing sensitive issues, such as sending reminder emails, accommodating religious considerations for workshop timings, and helping us to design a feedback survey combining multiple-choice and open-ended questions.

While some suggestions increased workload for the project team—such as adapting to Zoom and revising materials—the overall impact on the project's success was significant.

Both PI partners had largely positive experiences. One noted feeling valued and supported, while the other described it as one of their best PI experiences. However, as feedback was collected by the project team rather than an independent researcher, this should be acknowledged as a limitation. Key positives included time for workshop rehearsals, breakout room facilitation support, open discussions at PMG meetings, and feeling heard. Suggested improvements included allowing more time for workshop discussions and tailoring expert presentations further for a lay audience.

We actively sought to recruit a diverse group of public members, collaborating with Egality Health and community organisations to include individuals from varied demographic and socioeconomic backgrounds. While prior knowledge of synthetic data was not required (or desired), we observed that some public members brought prior knowledge of the related topic of administrative data to the workshops, and workshops attracted a relatively high proportion of attendees with at least higher educational attainment. The goal of this deliberative workshop approach was not to achieve a fully "representative" sample but rather to gather a breadth of perspectives, facilitating in-depth exploration of synthetic data issues. As such, caution should be exercised in interpreting these findings as representative of the wider UK public and we recommend further tailored engagement by data owners to ensure trust and address specific contexts.

The online-only format, while enhancing accessibility across geographic regions, may have unintentionally limited participation from individuals with restricted digital access or lower technological proficiency. To improve inclusivity, future consultations might consider a hybrid approach that combines online and in-person formats, accommodating public members with varied levels of digital literacy. The series of three workshops across five weeks may have also been a barrier for participation and may have contributed to the drop-out rate of 25%. The team did not have capacity to bring in additional new members once the workshops had begun (i.e. by running the workshop a third time to bring them up to speed) which has been done in other recent public consultations [24].

It is important to highlight that the focus of these workshops was on low-fidelity synthetic data. We cannot assume the same perspectives for high-fidelity synthetic data where the risk to privacy is greater, and the uses are more varied. The applications and use-cases of high-fidelity synthetic data will likely raise different ethical challenges, and further work is needed to examine public attitudes towards a spectrum of synthetic data types. This would provide more nuanced insights into how fidelity levels influence perceptions of privacy and utility.

This project was one of two funded projects in response to a call for complementary projects to explore the use of synthetic data from the perspectives of data owners and trusted research environments and the perspectives of the public [25]. The recommendations presented here are particularly relevant to data owners and trusted research environments. However, the underlying principles—such as clear language and accessible information about synthetic data—extend to researchers, funders, and regulators as well.

We acknowledge that the successful implementation of these recommendations by data owners will of course depend on several barriers and enablers. A key barrier is the potential resource requirement: producing accessible communication materials (e.g. animations, infographics, case studies) and maintaining registration systems for synthetic data access may be challenging for some data-owning organisations with limited capacity. Another barrier is the lack of harmonised standards across sectors, which may complicate consistent messaging and public engagement. Conversely, strong enablers include the growing policy emphasis on transparency and public involvement in data use, existing infrastructure in trusted research environments, and the availability of established guidance on public involvement and communication from organisations such as ADR UK and NIHR. The clear appetite expressed by public members for more transparent, accessible communication is itself an enabling factor, as it signals a pathway for building public trust and a social licence for the use of synthetic data. Future work should continue to examine these barriers and enablers in practice, including how recommendations can be adapted for different organisational and sectoral contexts.

With synthetic data likely to become increasingly important to UK and international data policy, we hope that the approach taken here provides a practical framework for ongoing public engagement in this area. Future research is likely required to assess evolving public attitudes to the use of synthetic data for research (both low- and high-fidelity) as it becomes more commonplace. Particular attention will be needed to manage concerns raised through these workshops, including the risks of potential re-identification, bias amplification, and accountability for misuse, which could undermine confidence if left unaddressed. Ensuring transparency around how synthetic data is generated and applied will be critical to sustaining both regulatory compliance and public trust. Monitoring public perception over time will be essential to ensure sustained trust. Insights and the recommendations from this project can serve as a starting point for data owners looking to responsibly create and share synthetic versions of their health and administrative datasets.

Conclusion

This is the first UK-wide conversation that has been undertaken with members of the public to understand their perspectives on how data owners should create and share synthetic data for research. Whilst we did not recruit a fully representative sample, we used strategies to ensure the inclusion of individuals from diverse backgrounds to inform recommendations. The recommendations are directed towards data owners but our findings are relevant for researchers and policy makers - it is a collective issue to build and maintain public trust on this matter. We hope this work can serve as a foundation for future synthetic data policy.

Acknowledgments

We would like to express our sincere thanks to the members of the public who took part in the workshops for their commitment, enthusiasm and contributions. We are also grateful to our expert speakers Emily Oliver, Lora Frayling, Mhairi Aitken and additional steering group members Shayda

Kashef, Harriet Baird, Sophie McCall, Katie Oldfield who helped to shape the project and our recommendations. We also thank Sarah Rawlinson for her help in facilitating the workshops.

Funding

This study was funded by ADR UK and UKRI, grant number ES/Z502479/1. The Centre for Trials Research receives infrastructure funding from Health & Care Research Wales.

Statement on conflicts of interest

None declared

Ethics statement

This project was a public consultation/engagement activity and not a research project. As such, Cardiff University agreed that research ethics approvals were not required. Rather, a public engagement risk assessment was completed which considered how to conduct the project in an ethical manner. Public members were provided with a welcome pack containing information on the project and they provided email consent to take part in the workshops, and for the project team to use anonymous quotes in written reports.

Data availability statement

As this project was not research, no data were generated or used.

References

1. Kokosi T, Harron K. Synthetic data in medical research. BMJMed. 2022;1:e000167. Available from: https://doi.org/10.1136/bmjmed-2022-000167.

2. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digit Med. 2024;6(186). Available from: https://doi.org/10.1038/s41746-023-00927-3.

3. Drechsler J, Haensch AC. 30 years of synthetic data. Statist Sci. 2024;32(2):221-42. Available from: https://doi.org/10.1214/24-STS927.

4. Research data Scotland. About Synthetic Data. 2024 [accessed 18 December 2024]. Available here: https://www.researchdata.scot/work-and-impact/synthetic-data/.

5. NHS England. Artificial data pilot. 2024 [accessed 10 December 2024]. Available from: https://digital.nhs.uk/services/artificial-data.

6. Clinical research practice datalink. Synthetic data. 2024 [Accessed 10 December 2024]. Available from: https://www.cprd.com/synthetic-data.

7. Savage N. Synthetic data could be better than real data. Nature. 2023 Available from: https://doi.org/10.1038/d41586-023-01445-8.

8. ADR UK. An interim ADR UK position statement on synthetic data. 2023 [Accessed 10 December 2024]. Available from: https://www.adruk.org/news-publications/publications-reports/an-interim-position-statement-from-adr-uk-on-synthetic-data-new645d0cdb19d98062623548/.

9. Beierle TC. The quality of stakeholder-based decisions. Risk Analysis. 2002;22(4):739-49. Available from: https://doi.org/10.1111/0272-4332.00065.

10. Campbell JW. Public participation and trust in government: Results From a vignette experiment. JPS. 2023;38(2):23-31. Available from: https://doi.org/10.52372/jps38203.

11. National consumer council. Deliberative public engagement: nine principles. 2008 [accessed 12 December 2024]. Available from: https://www.involve.org.uk/sites/default/files/uploads/document/Deliberative-public-engagement-nine-principles_0.pdf.

12. Egality health. Egality health. 2024 [accessed 10 December 2024]. Available from: https://egality.health/.

13. Health data research UK. Your voice, your story. 2024 [accessed 12 December 2024]. Available from: https://www.hdruk.ac.uk/about-us/involving-and-engaging-patients-and-the-public/get-involved/hdr-uk-stories/

14. National institute for health and care research. People in research. 2024 [accessed 12 December 2024]. Available from: https://www.peopleinresearch.org/.

15. Engage. Personal & public involvement. 2024 [accessed 12 December 2024]. Available from: https://engage.hscni.net/.

16. National institute for health and care research. Payment guidance for researchers and professionals. 2024 [accessed 10 December 2024]. Available from: https://www.nihr.ac.uk/payment-guidance-researchers-and-professionals.

17. Research Data Scotland. Synthetic Data. 2024 [accessed 02 September 2025]. Available from: https://www.researchdata.scot/work-and-impact/current-projects/synthetic-data/.

18. Staniszewska S, Seers K, Altman DG, Denegri S, Littlejohns P, Thomas V. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ. 2017;358:j3453. Available from: https://doi.org/10.1136/bmj.j3453.

19. Wagner JK, Cabrera LY, Gerke S, Susser D. Synthetic data and ELSI-focused computational checklists—A survey of biomedical professionals' views. PLOS Digital Health. 2024 Nov 20;3(11):e0000666. Available from: https://doi.org/10.1371/journal.pdig.0000666.

20. Kennedy H, Oman S, Taylor M, Bates J, Steedman R. Public understanding and perceptions of data practices: a review of existing research. Living With Data, University of Sheffield. 2020. https://livingwithdata.org/project/wp-content/uploads/2020/05/living-with-data-2020-review-of-existing-research.pdf.

21. Stockdale J, Cassell J, Ford E. "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. Welcome Open Res. 2019;3(6). Available here: https://doi.org/10.12688/wellcomeopenres.13531.2.

22. Aitken M, de St. Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics.2016:17(73). Available here: https://doi.org/10.1186/s12910-016-0153-x.

23. Muller SHA., Kalkman S, van Thiel, GJMW, Mostert M, Delden JJMVD. The social licence for data-intensive health research: towards co-creation, public value and trust. BMC Med Ethics. 2021:22(110). Available here: https://doi.org/10.1186/s12910-021-00677-5.

24. Pearse H, Evans J, Radicati A, O'Keeffe E. Using health data to identify and approach people about health and care research: A Public Dialogue. National Centre for Social Research. 2024 [accessed 17 December 2024]. Available here: https://healthandcareresearchwales.org/health-and-care-professionals-policy-and-guidance/data-research-programme.

25. ADR UK. Funding awarded to two projects to explore the use of synthetic data. 2024 [accessed 17 December 2024]. Available here: https://www.adruk.org/news-publications/news-blogs/funding-awarded-to-two-projects-to-explore-the-use-of-synthetic-data/.

Abbreviations

Q&A:	Question and Answer
PI:	Public Involvement
PMG:	Project Management Group
PSG:	Project Steering Group