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**