HDRUK
Health Data Research UK

HDR UK Synthetic Data Special Interest Group

Workshop Outputs – June 2022

HDRUK
Health Data Research UK

HDR UK workshop on synthetic data generation

Professor Katie Harron, Child Health Informatics Group, UCL Great Ormond Street Institute of Child Health
Dr Alison Elderfield, Data Strategy Manager, HDR UK

Research using administrative or routinely collected data has the potential to generate huge benefits to healthcare services and patients. However, gaining access to potentially sensitive, individual-level data can be an extremely lengthy process. Once data have been received, other time-consuming elements of studies using administrative data include understanding the structure of the data and developing data cleaning, management and analysis plans. If there was a way to do these tasks in parallel to applying for access to the data, research using these data sources could be streamlined. This is where synthetic data holds huge potential.

Synthetic data is artificially generated data designed to mimic real datasets, without containing any personally identifiable information. Synthetic data is a field that is gaining momentum. In 2020, HDR UK formed a Synthetic Data Special Interest Group, and hosted a first meeting in December 2020. The webinar aimed for an attendance of around 20 people and attracted 150 attendees. In parallel, there has been interest from Administrative Data Research UK (ADRUK), whose mission is to transform the way researchers use government data, in the use of synthetic data to make research more efficient.

By June 2022, Synthetic Data has become an increasingly notable topic of discussion. The HDR UK Synthetic Data Special Interest Group, jointly led by colleagues at UCL, brought together those interested in generating or accessing synthetic data for research, to share latest developments and advances in the field, and to form a community. There was a need to map the landscape, encourage networking, and spark discussions to understand the barriers to wider use of synthetic data. The Group hosted 55 people to learn about a range of activities in this space from speakers who are using or generating synthetic data, to support the community, and to agree on some top priorities to focus on during the next two-three years. The project outcome was to better understand the challenges and the opportunities for accelerating the uptake of synthetic data, by bringing together researchers from diverse backgrounds.

HDR UK is a conduit to facilitate the community to push forward with progress and developments in the use of synthetic data. While HDR UK does not itself generate synthetic data, the data is important to HDR UK because it offers the potential to speed up access to UK healthcare datasets.

Synthetic data: the story so far

There is a spectrum of fidelity in synthetic data:

- High fidelity synthetic data preserve statistical relationships
- Low fidelity synthetic data might only preserve the structure of the data

Depending on the fidelity of the data, synthetic datasets could potentially:

- Facilitate easier access to data for those who are generating hypotheses and developing tools
- Prepare and train researchers for the practical challenges of working with national clinical datasets
- Be used as pilot data (instead of real data) to strengthen researchers' applications when they apply for access to real clinical datasets

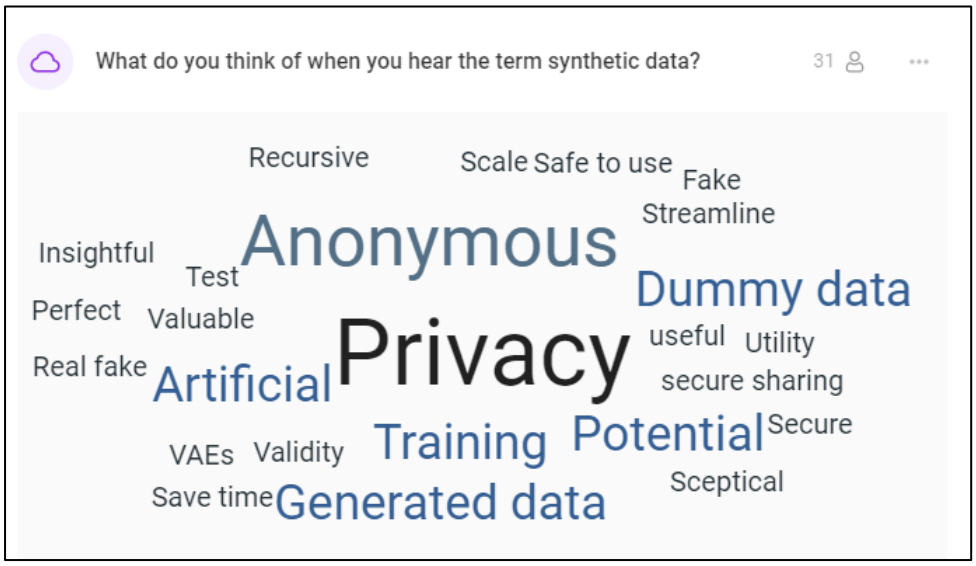
However, synthetic data are not yet widely used in medical and health research and several challenges remain:

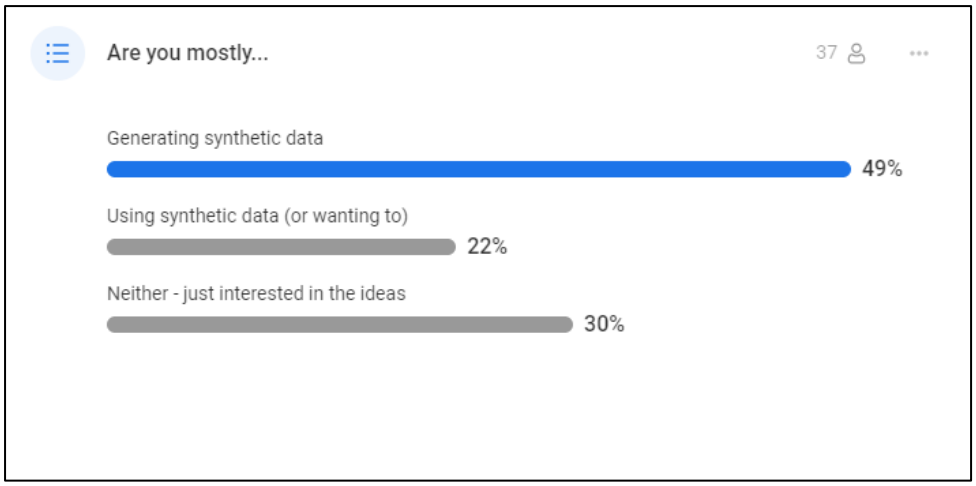
- Generating high fidelity data is a complex process
- Evaluating the utility of synthetic data is a challenge
- Understanding implications for privacy and disclosure is also complex
- Communicating about synthetic data is difficult
- There is some controversy around whether synthetic data could ever be used for real analyses or to support decision making
- There is a lack of evidence about the cost-effectiveness of producing synthetic data

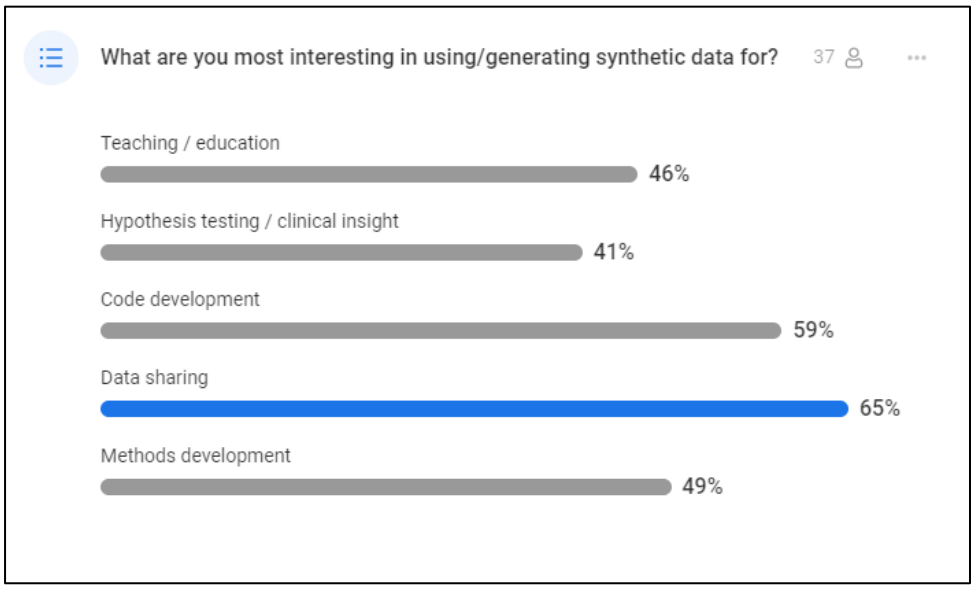
Synthetic data workshop: summary of discussions

We had engaging presentations from a variety of backgrounds, and we aimed to represent the broad range of entities who have an interest in generating these data, including health and other administrative data (see programme below). Many issues were discussed with some key priorities and challenges identified.

The audience were asked several introductory questions:


What do you think of when you hear the term synthetic data?
31
Recursive
Scale
Safe to use
Fake
Streamline
Insightful
Anonymous
Test
Dummy data
Perfect
Valuable
Privacy
useful
Utility
Real fake
Artificial
secure sharing
Training
Potential
Secure
VAEs
Validity
Sceptical
Save time
Generated data


Are you mostly...
37
Generating synthetic data
49%
Using synthetic data (or wanting to)
22%
Neither - just interested in the ideas
30%


What are you most interesting in using/generating synthetic data for?
37
Teaching / education
46%
Hypothesis testing / clinical insight
41%
Code development
59%
Data sharing
65%
Methods development
49%

Key priorities identified

Availability of synthetic datasets at low fidelity levels

One of the most consistent themes in discussions throughout our workshop was that whilst there has traditionally been an emphasis on developing methods for generating high fidelity datasets, an important first step in accelerating the use of synthetic data would be to facilitate a roll out of low fidelity datasets from a range of data providers. These low fidelity datasets could be used for training, to consolidate and validate methods for generating and evaluating synthetic data, and to demonstrate use cases. This will pave the way for high fidelity datasets by showing the value of synthetic data in a low risk and low resource setting and demonstrating what more could be done with more sophisticated methods. Such a roll out would require incentivisation to data providers to generate synthetic data and to ensure it is made available to a broad range of users.

Education for the public and researchers

The terminology used to describe synthetic data is varied and there has been an evolution of terms. The language used to describe synthetic data is not public-facing, and more will need to be done to explain what synthetic data are and to gain public trust in the use of their data for this purpose. A standardised taxonomy is required, so that a shared understanding of the importance and use of synthetic data amongst data scientists and the public can be developed. Organisations should improve communications about the use of data, and the levels of risk involved.

Evaluating privacy risk

Although many methods exist for evaluating utility of synthetic data, there is still more to be done in terms of evaluating privacy risk. These two aspects – privacy and utility – need to be tackled jointly, as there is an inherent trade-off between the two.

HDRUK
Health Data Research UK

Workshop programme: Thursday 16th June 2022

Purpose of the meeting:

1. Bring together those who are interested in generating or accessing synthetic data for research
2. Share latest developments and advances in the field
3. Understand the barriers to wider use of synthetic data

Total participants (inc. panellists): 55

Time	Item	Speakers	Chair
09:30	*Registration and Welcome Tea and Coffee*		
10:00	**Welcome and overview**	Katie Harron (UCL)	
10:15	**Accelerating public policy research with easier & safer synthetic data**	Paul Calcraft (BIT)	Katie Harron
10:40	**Synthetic data: the ONS perspective**	Iain Dove (ONS)	
11.05	*Coffee Break*		
11:20	**How private is synthetic data?**	Gillian Raab (Edinburgh)	Chris Rowe
11:45	**Real world applications of synthetic data – leveraging national cancer data to accelerate insights for patient benefit**	Julia Levy (IQVIA)	
12:35	*Lunch Break*		
13:25	**Data needs for product development: Can synthetic data help?**	Suzanne Weller (Privitar)	Geoff Hall
13:50	**Synthetic data applications**	Puja Myles (CPRD)	

14:30	*Coffee break*		
14:45	**Panel session (remaining challenges/next steps)**	Tony Calland (CAG) Lora Frayling (Simulacrum) Chris Rowe (Innovate UK) Robin Mitra (Cardiff University) Paul Berg (IQVIA)	Katie Harron
15:45	**Closing Remarks**	Katie Harron (UCL)	

Session 1 - Paul Calcraft

Title: Accelerating public policy research with easier & safer synthetic data

Summary: Critical research projects are frequently delayed or abandoned once the *theory* of the research plan meets the "reality" of the data itself, with substantial time and energy wasted along the way. We can prevent this by sharing - early and often - some synthetic data that looks and feels like the real data, but contains no real information or even any correlations. We're trying to make this as easy as possible to do quickly and safely, initially across UK government. We'll talk about the approach, what we've achieved, and where we're hoping to go with it next.

Session 2 – Iain Dove

Title: Synthetic data: the ONS perspective

Summary: Synthetic data presents an opportunity for data owners to improve data access and enable more research, with the associated challenges of proving that the data are non-disclosive and suitable for sharing. This talk will provide an overview of the ONS perspective on synthetic data, how synthetic data is used in ONS, and current synthetic data projects.

Session 3 - Gillian Raab

Title: How private is synthetic data?

Summary: In developing the R package synthpop for data synthesis, the major focus has been to improve data utility so as to ensure that the results will reproduce what would be found from the real data. Differential Privacy (DP) is considered by some to be the gold standard for protecting privacy. I will discuss how DP may relate to another measure of disclosure control for synthetic data.

Session 4 – Julia Levy

Title: Real world applications of synthetic data – leveraging national cancer data to accelerate insights for patient benefit

Summary: NHS Digital's Cancer Analysis System (CAS) is one of the most detailed cancer databases in the world containing robust data for all cancer patients in England. IQVIA supported an initiative led by Health Data Insight (HDI), to develop the Simulacrum, a synthetic oncology dataset structured to model properties of CAS, but which contains no actual patient data. In this presentation, we will describe how this synthetic data set is being leveraged to accelerate real-world insights into patient care and outcomes.

Session 5 – Suzanne Weller

Title: Data needs for product development: Can synthetic data help?

Summary: We will explore some of the scenarios where having access to realistic data is crucial for the development of new products and services. By looking at the strengths and limitations of different types of synthetic data we will discuss whether it can meet these requirements.

Session 6 – Puja Myles

Title: Synthetic data applications

Summary: This presentation will focus on potential synthetic data applications including for validation of machine learning algorithms, product development, education/training, privacy preservation, sample boosting and bias correction.

Session 7 – Suzy Gallier - Absent

Title: Adoption and development of methodologies for the use of synthetic data within health data hubs (PIONEER and BREATHE).

Summary: This talk will explain the background to the creation of synthetic datasets in the HDR UK hubs PIONEER and BREATHE, comparing the different approaches each hub has taken. Both hubs have been leading the way with their licenced access to synthetic data and this talk will include an overview of the data offering, approach and an overview of a recent joint hub event on Synthetic data in AI for SMEs.

Discussions following the panel session

Polls

1. **Given what we've heard today, what do you think the main priorities are for this field going forward?**

- Public education is required on synthetic data. This would involve developing a common lexicon and agreement on how to engage with data providers, researchers, and the public. This could help to achieve public understanding and buy-in regarding the use of synthetic data
- Evolving a common vocab, measures to discuss fidelity and utility and developing techniques to create high-fidelity data. Clear distinctions are needed between low fidelity and high fidelity. It would be useful to develop metrics for measuring synthetic data.
- Specific governance requirements. This is for use cases and to establish a process/protocol for releasing synthetic datasets to users. This would involve working together across organisations to standardise policy
- Demonstrate various applications of synthetic data to data controllers and the utility for the range of putative use cases. This would involve reaching clarity on limitations and advances.

2. **Are there any topics you would have like to have seen, which weren't discussed today? Are there any other events you'd like to see in this field?**

- Increased learning for others: Public/patient success stories, experiences from more data users, e.g. students, learn more about what is happening in other countries, and regular opportunities to re-engage with this community so we can see how our thinking evolves and keep learning from each other.
- Advances in synthetic data, for example, synthetic data of free text fields, unstructured synthetic data approaches and a discussion of concerns.
- Evaluation standards and more validation studies, together with data quality issues, how it is measured, its impact on analytics, BI, etc.
- Comparisons with other privacy protection approaches.