HDRUK
Health Data Research UK

Synthetic Data Special Interest Group

Workshop Outputs – December 2020

HDR UK workshop on synthetic data generation

Aiden Doherty, National Implementation Project Lead on reproducible machine learning, HDR UK;
Neil Sebire, Chief Clinical Data Officer, HDR UK;
Clara Fennessy, DIH Programme Officer, HDR UK

Health Data Research UK held a workshop on 9th December 2020 to understand the current UK Health Data landscape regarding synthetic data activity. Synthetic data is artificially generated data designed to mimic real datasets, but not containing personally identifiable information. We had engaging presentations from data custodians, industry specialists, and academics working in both medicine and computer science (see programme just below). Many issues were discussed (see appendices at bottom) with two key challenges identified in this workshop:

What privacy risks are associated with synthetic data?

Our workshop highlighted successful research projects that have made significant progress towards understanding these issues. However, there is some disparity in currently utilised methods that can include a number of time intensive, and sometimes ad-hoc, hand-crafted tests. It is also important to note key differences between synthetic data generation methods (for e.g. perturbation-based approaches and model-based approaches) when considering privacy risks as the risks can be very different. A partial solution might be the introduction of synthetic data competitions, where some groups would be incentivised to reidentify participants in synthetic datasets to inform the development of more robust privacy-preservation methods. However, at present there is no consensus on which evaluation methods should be used to determine whether individual participants are likely to be re-identified or not.

How to evaluate synthetic data generation methods?

It is difficult to mathematically evaluate how 'good' (or not) a method is for the generation of a synthetic healthcare datasets. A number of methods have been used including Bayesian network analysis, data perturbation, conditional generation, generative adversarial networks, perturbation graphs, Markov models, etc. However, at present there is no consensus on which method is most suitable for particular tasks when generating synthetic healthcare datasets.

To address these challenges, it will be important to bring together a diverse group of stakeholders to collaborate on a driver project for a synthetically generated national healthcare dataset. In particular this should include data architects, methods developers, and clinical scientists; all supported by an active public and patient involvement group. It is possible that such a dataset could be hosted on Health Data Research UK's Innovation Gateway, which can then support a variety of uses including academic use, teaching & training, benchmarking of methods, and commercial use (algorithm development).

If you would like to contribute to future efforts in this space, please do register your interest with us. Visit our webpage for more information.

HDRUK
Health Data Research UK

Workshop programme: Wednesday 9th December 2020

Purpose of the meeting:

1. To understand the current UK Health Data landscape regarding synthetic data activity
2. To propose suggestions for development of an initial 'position paper', regarding HDR UK strategy and activity in the space
3. To identify potential areas for targeted activity to support further collaborative research applications.

Link to recording: https://www.youtube.com/watch?v=V0b254unqSM&feature=youtu.be

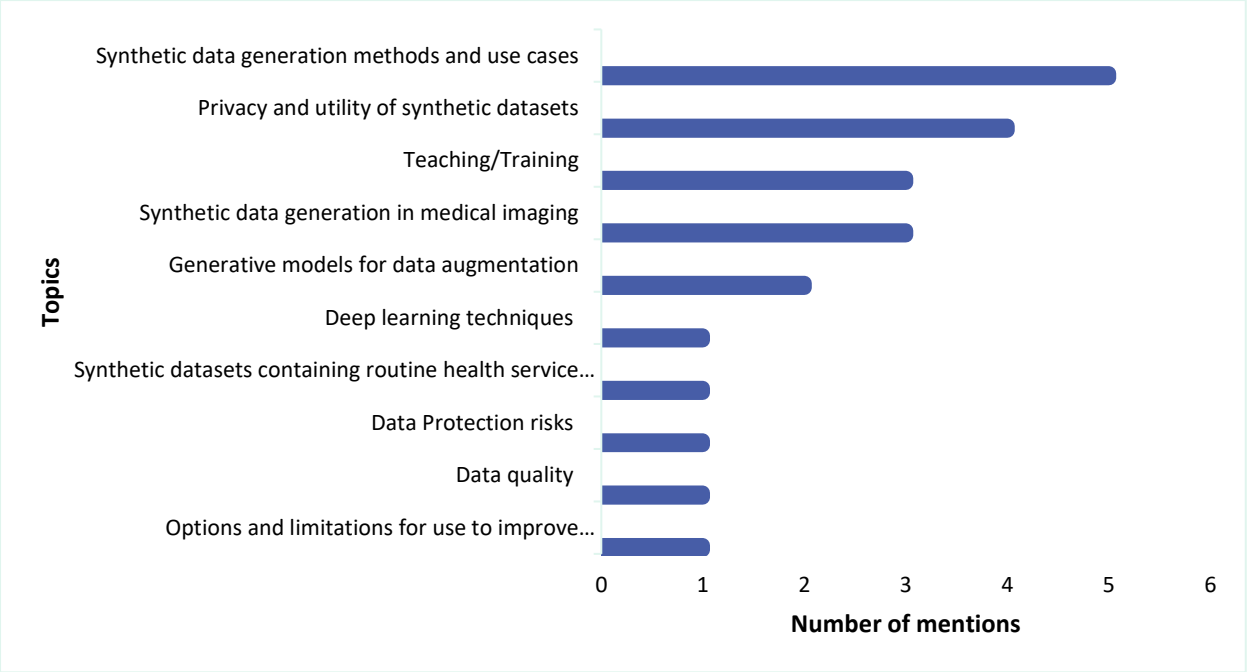
Total participants (inc. panellists): 149

Time	Item	Speakers
14:00	**Welcome & overview**	Aiden Doherty, Neil Sebire *Health Data Research UK*
14.10	**A data custodian's view of synthetic healthcare datasets**	Puja Myles *MHRA*
14:35	**How technically close are we to a vision for synthetic healthcare datasets?**	Mihaela van der Schaar *University of Cambridge*
15:00	**Real world challenges in sharing synthetic healthcare datasets: A case study of HES A&E and next steps**	Jonny Pearson *NHSX*
15.25	**Synthetic generation of complex healthcare data types (sensor data)**	Peter Charlton *University of Cambridge*
15.50	**Sharing synthetic healthcare datasets to advance cancer research**	Lora Frayling *Health Data Insight*
16.15	**How can we ensure patient privacy?**	Allan Tucker *Brunel University*
16.40	**Group discussion around key challenges to realise a vision for synthetic healthcare datasets**	All
16:55	**Meeting close**	
	• AOB • Next steps	Aiden Doherty, Neil Sebire *Health Data Research UK*
17:00	**Meeting Ends**	

Information collected prior to the workshop

It is important to note that data presented to all responses below are from a subset of the delegates (n<60) and may thus be subject to certain biases.

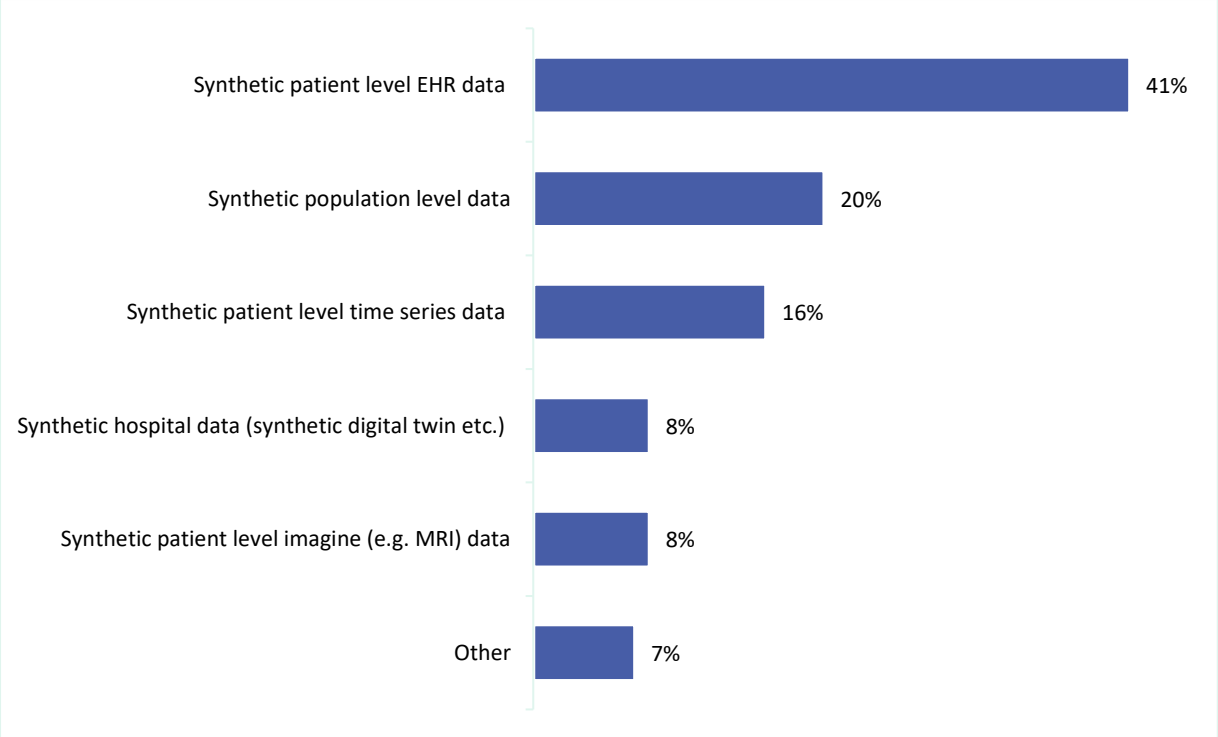
Desired learnings ahead of the workshop:



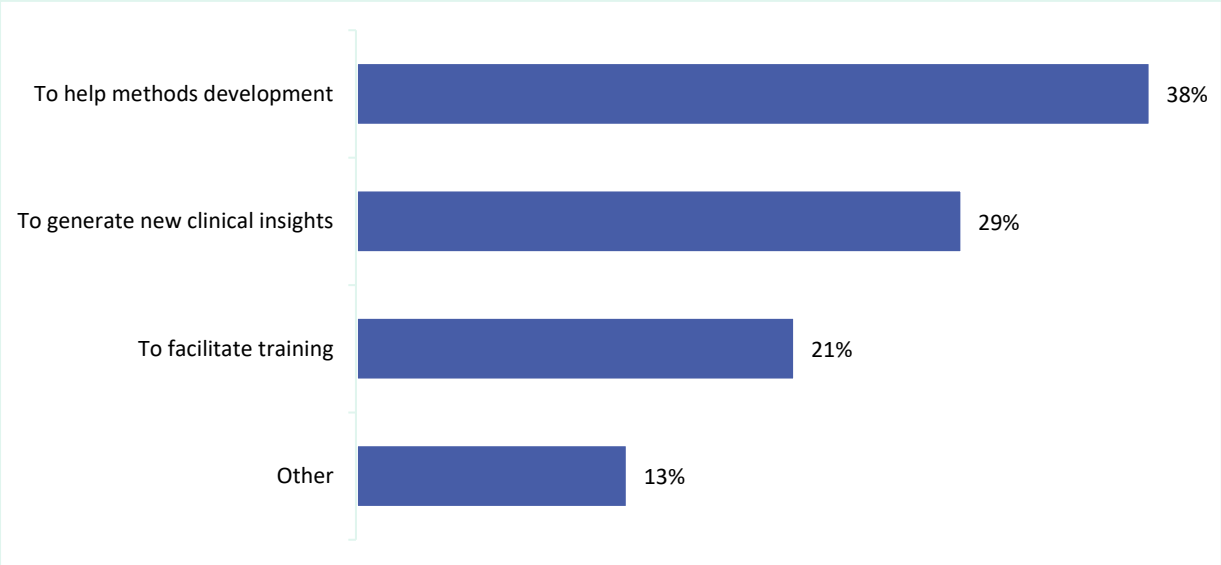
Discussions during workshop

Poll Questions

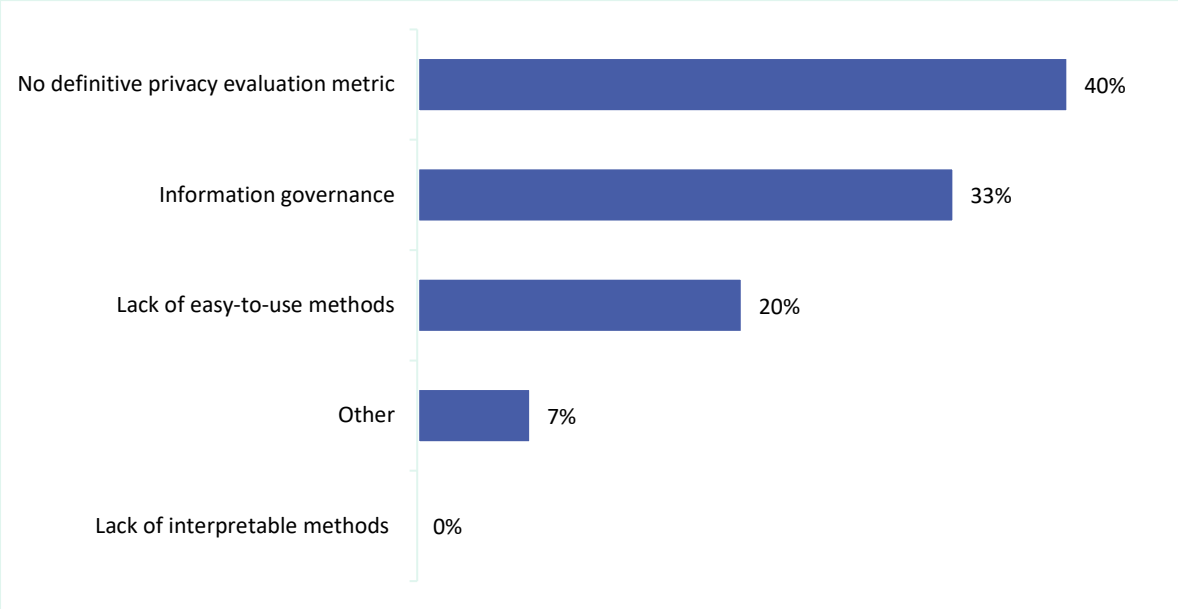
Which of the following broad areas (recognising potential overlap) would be your priority area of interest regarding synthetic data: (61 responses)



What do you view as the main reason for generating synthetic healthcare datasets? (24 responses)



What do you view as the major challenge towards realising a vision of commonly available synthetic healthcare datasets? (30 responses)



Free text questions to webinar attendees:

Question	Responses (where provided)
Please share any other major challenges towards realising a vision of commonly available synthetic healthcare datasets?	Epistemic risks when synthetic data are based on large proportions of original datasets with high importance for primary research (NHS, research studies, etc.)
Please share your other reasons for generating synthetic healthcare datasets	To allow for development and analysis prior to receiving real data (timelines for access can be long, and synthetic data can give a head start).
	To release registry to pharma funders. They could interrogate & send models to run on the real data. We have protection from overfitting and multi comparisons.
Please enter any additional ideas on priority interest areas for synthetic data	Multi modal data
	Synthetic (high-dimensional) molecular data.
What do you view as the big challenges in generating synthetic healthcare datasets?	Fidelity - preserving the same relationships in the data, when learning those relationships is the objective of using that data in the first place.
	Neural Networks models tend to memorise training data in various ways.
Are there methods used in other fields that we should consider in the healthcare domain?	Normalising flow