PUBLIC VIEWS ON SYNTHETIC DATA:

Key takeaways from the 'Discussing Data' Project

DISCUSSING DATA

UKRI Economic and Social Research Council

ADR UK

Cardiff University Prifysgol Caerdydd
Centre for Trials Research
Canolfan Ymchwil Treialon

Introduction

Public sector organisations, like the NHS and government departments, are starting to create something called synthetic data.

This is data that mimics real administrative data, but doesn't include real people's information. It's used to help researchers practice or prepare for working with real data.

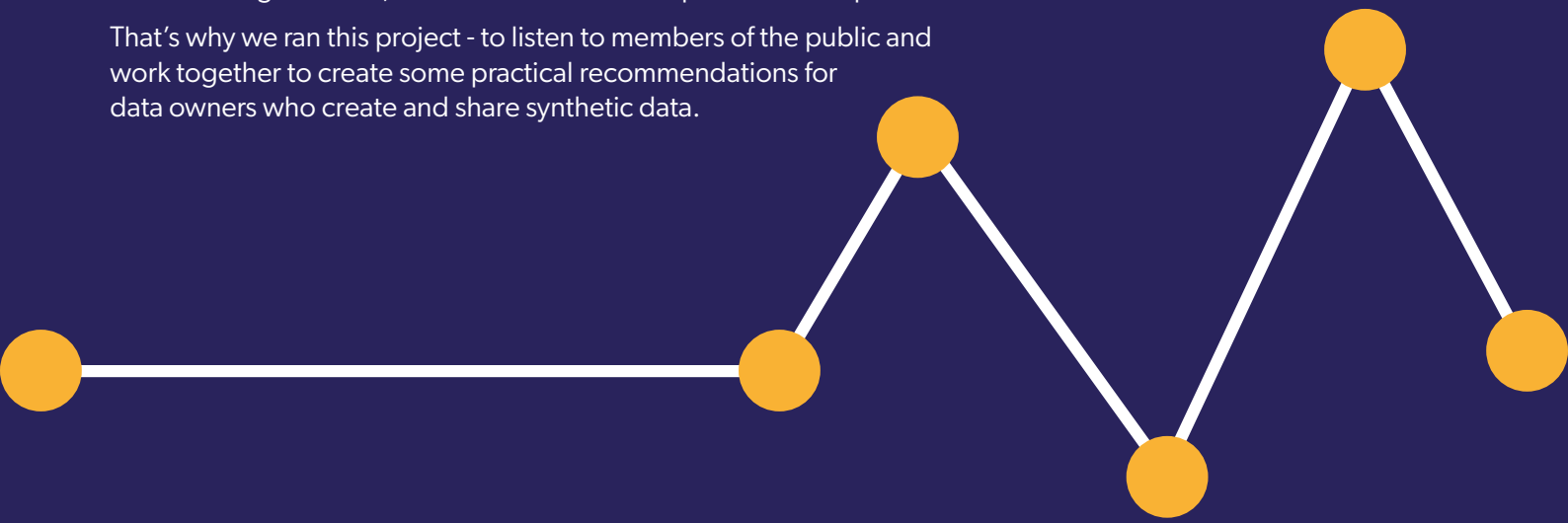
Until now, the public haven't really been asked what they think about this - whether they're happy for synthetic data to be shared with researchers, who should get access, and how it should be explained to the public.

That's why we ran this project - to listen to members of the public and work together to create some practical recommendations for data owners who create and share synthetic data.

What is Administrative Data?

Administrative data is real information collected about people by public organisations, like the government or the NHS, as part of their everyday work.

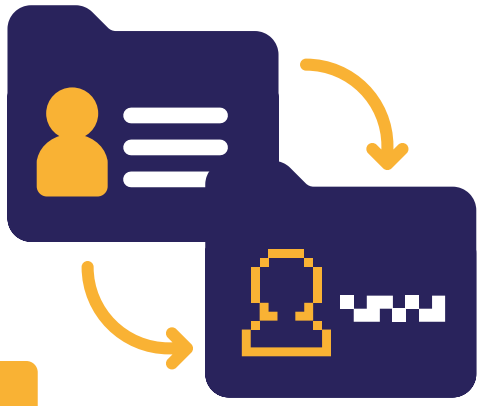
This could include data about our health, education, or employment. These organisations are responsible for looking after the data - they are the **data owners**.



What is synthetic data and why is it useful?

Synthetic data is created to reflect real data without using real people's information.

It is made by analysing the patterns in the real data and creating new data that follows the same trends – but without including any actual individuals.



For example, in healthcare, synthetic patient datasets will *look* like a set of medical records, but none of the records will belong to real patients.

It can take time for researchers to gain access to real data as data owners need to carry out checks to make sure data is used safely and appropriately.

While waiting for access, researchers can benefit from using synthetic data to:

- test research ideas and computer codes
- train students and junior researchers, without the -need to handle confidential information

Some synthetic data is very detailed (high-fidelity) and designed to be as close as possible to real data.

In this project, we focused on low-fidelity synthetic data. This only reflects broad patterns and isn't generally used for detailed analysis.

What did the ‘Discussing Data’ project do?

We ran a public engagement project to explore how people feel about synthetic data. This involved four online workshops with 39 members of the public from different backgrounds across the UK.

The workshops included:

Expert talks
– so people could learn about what synthetic data is and how it’s used.

Group discussions
– where attendees spent time in smaller groups to share their thoughts, concerns, and ideas.

Interactive activities
– helping people explore what synthetic data might be used for and how it might be shared.

During the final workshop, attendees worked with us to develop **recommendations for data owners** - the organisations that create and share synthetic data.

To read about this in more detail, you can see our paper here.

Being inclusive

We worked with *Egality Health* to involve members of the UK public from a wide range of backgrounds to get a diverse mix of views. This included people of different ages, cultures and locations.

We asked people to contact us if they were interested in joining the workshops.

138 applications received

44 people invited to participate

39 people took part

What we found

Here are some of the key things we learned from the discussions:

1 People knew very little about synthetic data at the start

Most attendees had never heard of synthetic data before, and some found the name confusing or misleading. The term 'synthetic' made some people think of fake or unreliable data.

2 Transparency is key

People wanted clear, easy-to-understand explanations of what synthetic data is, why it's being used, and what the benefits and risks are. They felt that data owners need to communicate this more clearly.

> **"Accessibility needs to be at the forefront of any comms about synthetic data."**

3 Controlled access is preferred

There was concern that if synthetic data was made freely available to anyone, it could be misunderstood or misused. Most attendees felt there should be some level of control over who can access it and for what purposes.

4 Ethical oversight matters

Many people assumed synthetic data was generated automatically by AI, and they were concerned about a lack of human oversight. They wanted to know that real people are making decisions about how synthetic data is created, checked, and used.

5 Public trust depends on clarity and fairness

People wanted to know that synthetic data was being used for **public benefit**, such as for health research. However, they were less comfortable if they felt it could be used in ways that were unfair, unclear, or not properly regulated.

We asked the workshop attendees to suggest alternative words to synthetic data. They suggested many options but there was no single word that everybody agreed upon.

Our recommendations

Based on everything we heard in the workshops, we worked with attendees to create practical recommendations for data owners - the organisations responsible for creating and sharing synthetic data.

We hope these recommendations will help organisations share synthetic data responsibly, in ways that align with public attitudes and expectations.

Introducing synthetic data

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

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?

RECOMMENDATION 4

Explain the personal privacy benefits that using synthetic data offers.

Accountability for creating your synthetic data

RECOMMENDATION 5

Provide a simple explanation for how your synthetic data is created. In particular, you should explain the role of humans vs 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

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

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 creative 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.

Feedback from our public members

After each workshop we invited our public members to complete an online feedback survey.

We asked them what worked well and what we could improve.

What worked well:

- The sessions were well-structured, easy to follow and held at flexible times
- The use of Zoom worked well, especially for people using screen readers
- Breakout rooms gave everyone a chance to contribute

Attendees also suggested some improvements for future workshops:

- Sharing presentation slides in advance
- Mixing up the breakout groups so attendees could meet different people
- Allowing more time for small group discussions
- Consideration of simpler language, for people whose first language is not English

"The workshops were really well led to supply information and context from a very neutral stance. Learning was built very well layered session to session, leaving me feeling that I had understanding of a concept I originally had zero knowledge of."

"It is an amazing piece of coproduction."

"I would like to see public participants empowered to summarise the discussions from breakout sessions."

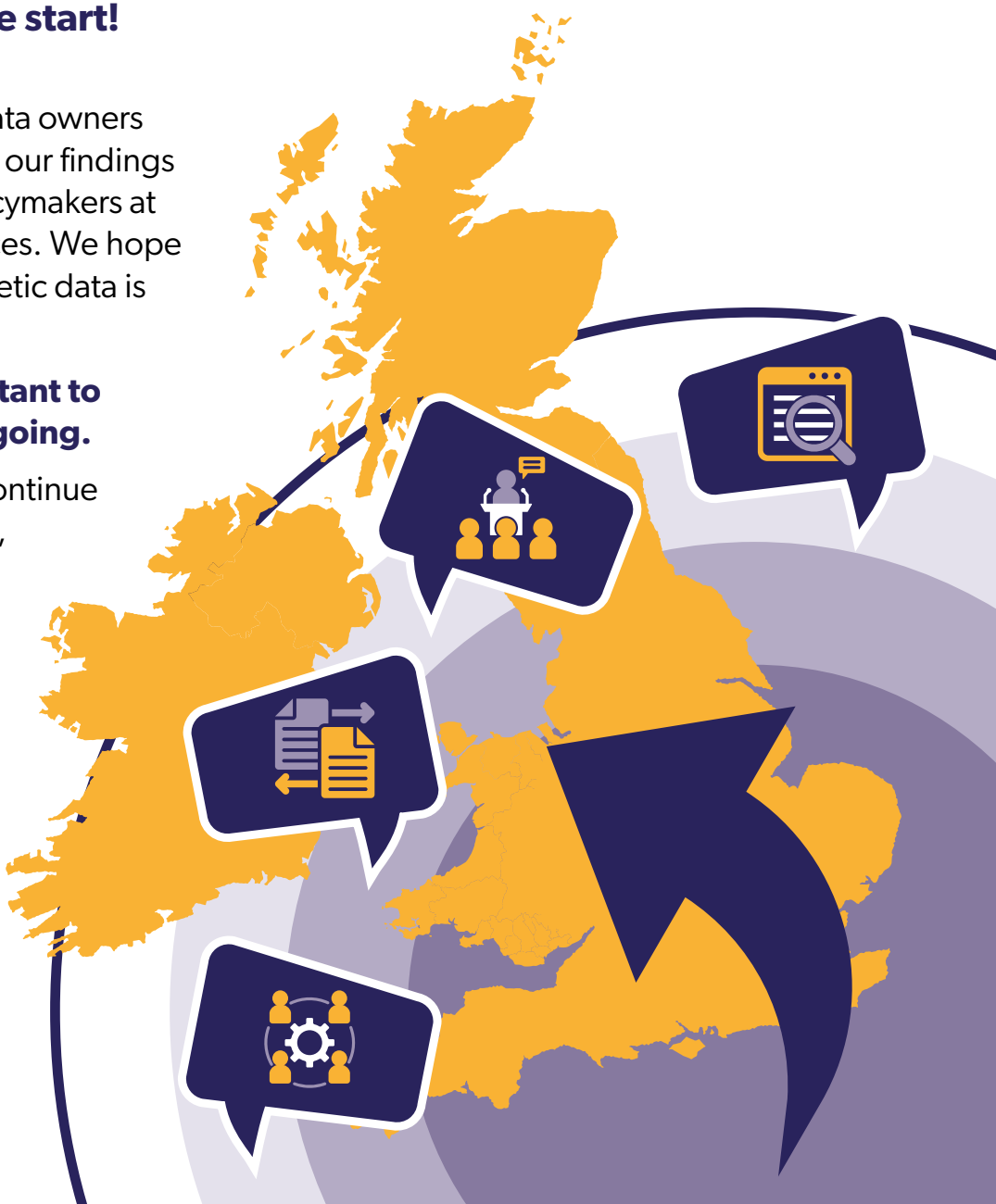
What happens next?

This project is just the start!

We will be sharing these recommendations with data owners across the UK and sharing our findings with researchers and policymakers at workshops and conferences. We hope they will shape how synthetic data is managed in the future.

We also think it's important to keep the conversation going.

We will look for ways to continue to engage with the public, data owners, researchers, and policymakers as synthetic data technology develops.



Acknowledgements

We sincerely thank all of the members of the public who took part in the workshops for their time, insights and feedback, those happy to be named include (Alphabetical):

Ann Muir
Araya Gautam
Clara MdB
Emily Lam
Hasmukh
Jane
Justin Greenwood
Linda Farey
Moira Auchterlonie
Patrica Jamal
Rebecca Hargreaves
Dr Sarah Markham
Sienna-Mae
Steve Moore
Mr Zul Hussain

We are also incredibly grateful to our two public contributors Farheen Yameen and Jim Fitzgibbon for their invaluable contributions to designing and managing the project and supporting the workshops.

We also benefited from input and advice from several expert groups and members of the public. With thanks to our Management Group and Steering Committee, the ADR UK Public Insight Panel and all of our expert speakers: Emily Oliver, Lora Frayling, Mhairi Aitken and Joshua Stock. Finally, thank you to Egality Health for their help in working with community groups to promote and recruit to the project.

The project was funded by ADR UK and URKI, grant number ES/Z502479/1

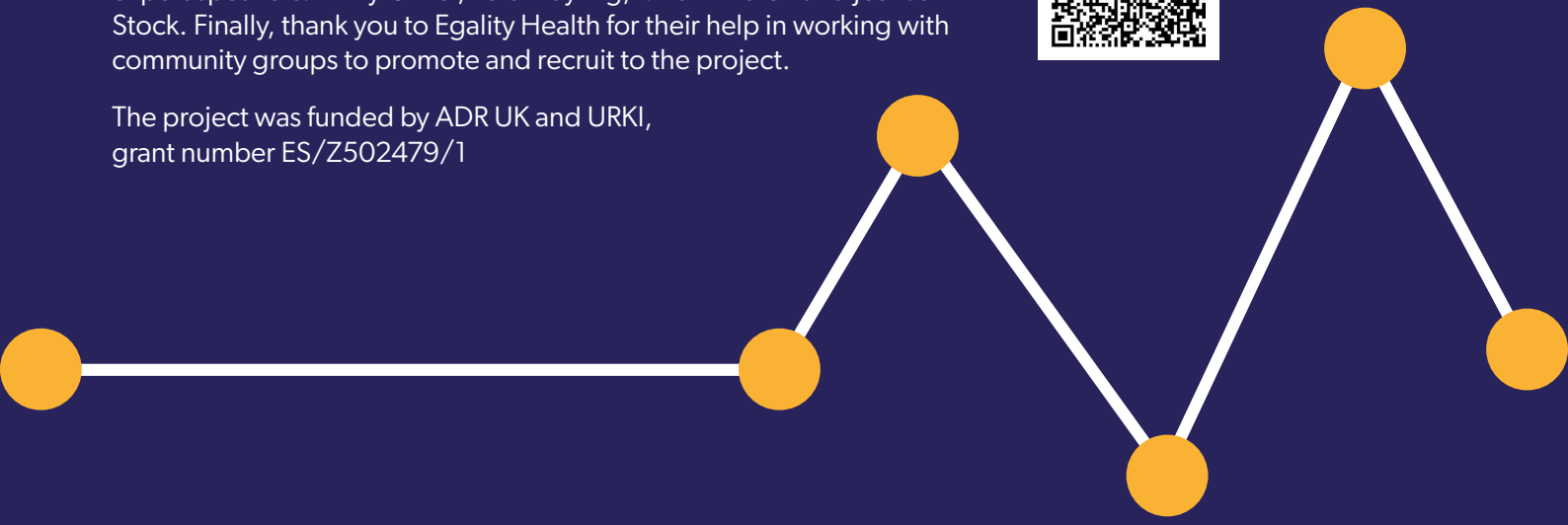
Contact the researchers

Dr Fiona Lugg-Widger
LuggFV@cardiff.ac.uk

Dr Rob Trubey
TrubeyRJ@cardiff.ac.uk

Dr Claire Nollett
NollettCL@cardiff.ac.uk

For further information please see our project website:



© 2025 by Discussing Data Project at Cardiff University is licensed under CC BY-NC-ND 4.0