20/06/2023

The RDS synthetic data strategy for Scotland: one year on

Dr Lynne Adair (Forrest), Data Curation Manager

Research Data Scotland

About Research Data Scotland

Research Data Scotland

We aim to help researchers make the most of existing data about people, places and businesses in Scotland.

- RDS is not a research organisation or a data collector
- We simplify and speed up the use of excellent data that already exists
- Our focus is on sensitive case-level data, not aggregated data

Our partners and stakeholders

Research Data Scotland

The Scottish Government
Riaghaltas na h-Alba

Public Health Scotland

National Records of Scotland

epcc

1495 University of Aberdeen

University of Dundee

The University of Edinburgh

University of Glasgow

NHS Greater Glasgow and Clyde

DataLoch

DaSH
Grampian Data Safe Haven
University of Aberdeen • NHS Grampian

Health Informatics Centre

HDRUK
Health Data Research UK

Office for National Statistics

Scottish Centre for Administrative Data Research

Unlocking the power of public sector data to make it quicker and simpler to do research and improve lives.

Public sector data use: issues

Research Data Scotland

Messy, difficult to use

Restricted access – held in safe haven

- Long process
- Have to apply in advance for data that may not be suitable for the project

How can synthetic data help improve access and reduce disclosure risk?

Research Data Scotland

Synthetic Data Definition

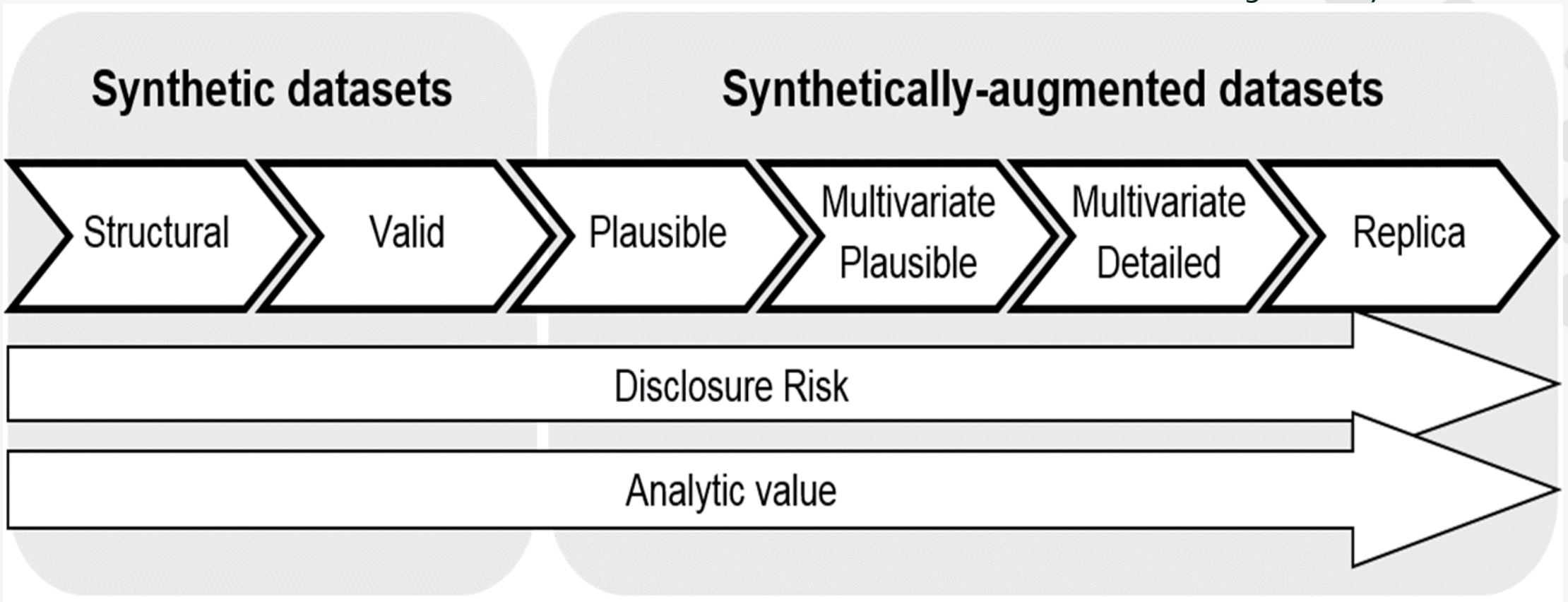
‘A new copy of a data set that is generated at random but made to follow the structure and some of the patterns of the original data set. Each piece of information in the data set is meant to be plausible, but it is chosen randomly from the range of possible values’

‘Accelerating public policy research with synthetic data’: ADR-UK Report Dec 2021

Fidelity Spectrum

Low fidelity

High fidelity



[ONS methodology working paper series number 16 - Synthetic data pilot - Office for National Statistics]

Uses and benefits of synthetic data

Research Data Scotland

Support researchers in advance of them accessing data

Training

- Use synthetic data as a training resource

Data discovery

- Support metadata catalogue (see structure of dataset, patterns of missing data)

Code development

- Writing and testing code before full data access is available
- To limit access to real data

RDS Synthetic data work

Scoping the synthetic data landscape

- Produced a discussion document and synthetic data strategy
- Set up a Scottish working group

Workstreams:

- Disclosure risk and Information governance
- Synthesis
- Access, promotion and engagement

Disclosure risk and IG

Research Data Scotland

Workstream	Sub Streams	Key questions	Existing/planned /proposed Activities	Priority/Timescale
1: Disclosure risk and IG	1. Disclosure risk	How do we measure the utility and assess the disclosure risk of synthetic datasets?	1. Create advice for data controllers on how to evaluate the disclosure risks from synthetic data	Underway - outcomes expected June 2023
			2. Develop a standardised synthetic data classification system to help understand fidelity and risk	From June 2023
			3. Summary statistics for presumptive disclosure	
	2. IG/Data controller engagement	How do we ensure that IG arrangements balance utility and privacy? What needs to be in place for data controllers to feel comfortable with the creation and use of synthetic data?	1. Create an IG group to hold regular meetings around synthetic data IG issues (to be done as part of RDS IG steering group)	Set up in Apr 2023
			2. Hold a workshop on SD for IG people and data controllers	2024
			3. Conversations with data controllers	2023

Synthesis

Research Data Scotland

Workstream	Sub Streams	Key questions	Existing/planned /proposed Activities	Priority/Timescale
2: Synthesis	1. Tools	Which synthetic data tools are most appropriate for synthesis of data for different uses and levels of fidelity?	1. Investigation of synthetic data tools and synthetic data requirements	Underway
			2. Develop Synthpop further (enable CART based methods to be more scalable and build in an SDC element)	Autumn/Winter 2023?
			3. Synthpop maintenance	Ongoing from 2023
	2. Synthesis projects	What demonstrator projects should we consider? Think about purpose, which datasets, fidelity, access, timescales, tools. This would include prototyping including data controller engagement, IG	1. Test synthesis included with 'synthetic data tools' project in Tools 2.1.1 above	2024?
			2. SMR01 synthesis	Underway
			3. RDS demonstrator project - low fidelity pupil census	Late 2023
			4. Development of high-fidelity(?) education data training datasets	Late 2023
			5. Imaging data synthesis	2024/2025
	3. Automation	How do we automate synthetic data production?	1. Investigate feasibility and means of automation, and scaling up production	2024/2025

Access, promotion and engagement

Research Data Scotland

Workstream	Sub Streams	Key questions	Existing/planned /proposed Activities	Priority/Timescale
3: Access, promotion and engagement	1. Access and promotion	How might we make synthetic data that is already synthesised more widely available?	1. Compile list of all known synthetic datasets - either created or planned	From Feb 2023
			2. Investigate whether we can make more widely available the SLS/SCADR Admin data training synthetic dataset	Feb-June 2023
			3. What other datasets can we make more widely available?	2024
			4. How do we promote available synthetic datasets?	2024/2025
	2. Synthetic data community engagement	What other work around synthetic data is going on in the UK and beyond, that we can learn from?	1. Set up a UK SD group with HDRUK, ADRUK and DARE	2023 onwards
			2. Use SDWG to circulate papers and invite speakers. Create library of documents and contacts	2023 onwards
			3. User workshop for those who have started synthesis	Autumn 2023
	3. Public engagement	How can we ensure that public trust is maintained in the production and use of synthetic data?	1. Discussion with SCADR Public Panel - early stage discussion	2nd May 2023
			2. ADRUK public consultation	2023-2024
			3. RDS public engagement strategy	2024/2025
	4. User engagement	What do researchers want?	1. RDS User workshop	Completed Oct 2022
			2. RDS User testing panel	As required
			3. Commission external user engagement support	As required
			4. Consolidate findings from different sources	From June 2023
			5.Interview researchers who used high fidelity synthetic data in the SLS and beyond	From June 2023

User engagement: workshop Nov 2022

Research Data Scotland

Short presentation describing what synthetic data is and RDS plans

Breakout sessions for participants to discuss:

- Benefits of and issues around synthetic data
- What sort of data (eg for training, data discovery or code development) and level of fidelity would be the most useful for researchers

Blog:

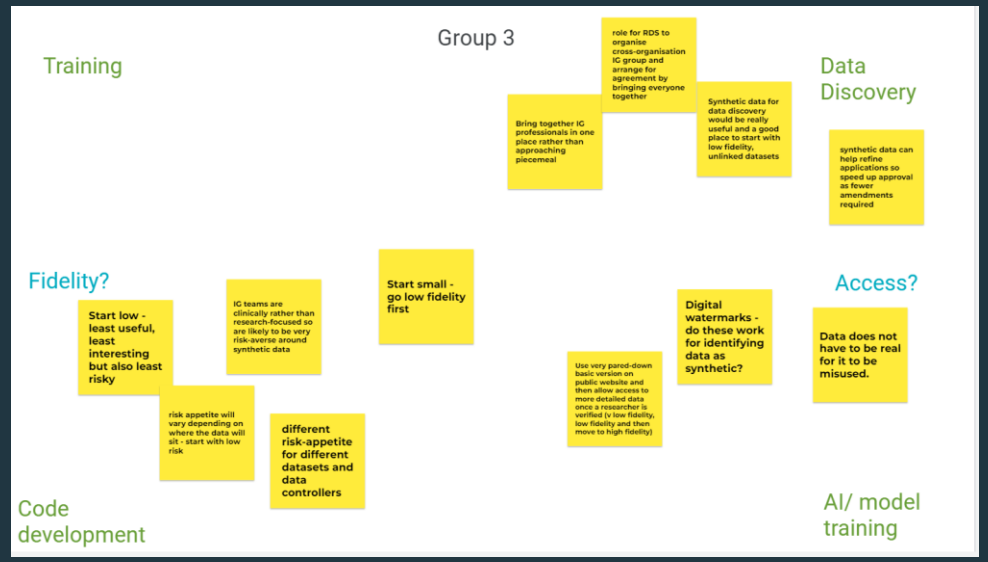
Unlocking the potential of synthetic data (researchdata.scot)

User engagement: workshop Nov 2022

Research Data Scotland

Outcomes

- Start by generating low fidelity datasets
  - More acceptable to data controllers
  - Good enough for researchers
- Manage IG challenges
  - Bring together IG people to discuss synthetic data management



Public engagement

Research Data Scotland

Presentation to RDS/SCADR public panel (May 2023)

- ‘What do you think of when you hear the term ‘synthetic data’?’
  - Mentimeter responses: ‘artificial’ ‘made up’ ‘not real’
- Described synthetic data, data access issues, how SD might help, use case (training/data discovery/code development) spectrum of generation (low to high fidelity synthesis)
- Use it to support researchers in advance of them accessing data, NOT for drawing conclusions from

Questions:

- What are your concerns around synthetic data?
- Are you comfortable with the uses described?
- Would you be comfortable with synthetic data being freely accessed online?

Themes

- Lively discussion, some strong views and mixed opinions
  - one strongly against, others on a spectrum between supportive and concerned
- Concerns:
  - that even with caveats people may still use the data to draw conclusions from, if data is publicly-available
  - code based on synthetic data may skew the research
- Unclear how synthetic data will benefit researcher and speed up research process

Public panel outcomes

Research Data Scotland

Better describe the benefits of synthetic data, how it differs from real data in ease and speed of accessibility

Lot of worries around mis-use and drawing conclusions from it – control access to synthetic data and don't make publicly-available

Worries about introducing bias into the research if synthetic data used for code development

Is there any evidence for this? Speak to researchers

Synthpop management

Research Data Scotland

- Synthpop is an open-source R package that allows users to create synthetic versions of confidential individual-level data
- RDS to take over synthpop management
- Future:
  - Update documentation
  - Produce a simple user guide and FAQs
  - Hold user workshops and promote synthpop use
  - Consider offering a chargeable synthesis service to others

Funding

Research Data Scotland

Funded work to:

- Investigate synthetic data tools
- Provide guidance and advice for data holders on how to evaluate the disclosure risks from synthetic data
- Develop example synthetic datasets

Set up a further synthetic data funding stream – autumn 2023

Research Data Scotland

Lynne.Adair@researchdata.scot