RDS Synthetic Data User Workshop

22/11/2022

Research Data Scotland

Workshop Agenda

2.00: Introduction and housekeeping

2.05: Synthetic data – intro presentation

2:10: Breakout rooms session A: synthetic data interest, benefits and issues

2.25 RDS synthetic data plans - presentation

2:35 Breakout rooms session B: synthetic data use cases

2:50: Group feedback and discussion

3:00 Finish

Dr Lynne Adair (Forrest)
Data Curation Manager

Introduction to synthetic data

Research Data Scotland

Data Access

Research Data Scotland

RDS Mission Statement:

To promote and advance health and social wellbeing in Scotland by enabling access to public sector data about people, places and businesses for research in the public good

Data access issues:

- Long timescales
- Information Governance hurdles
- Restricted access – safe haven
- Data availability

How can synthetic data help with this?

Synthetic Data Definition

‘Synthetic data are modelled statistical outputs released in a format that closely resembles the confidential data format’

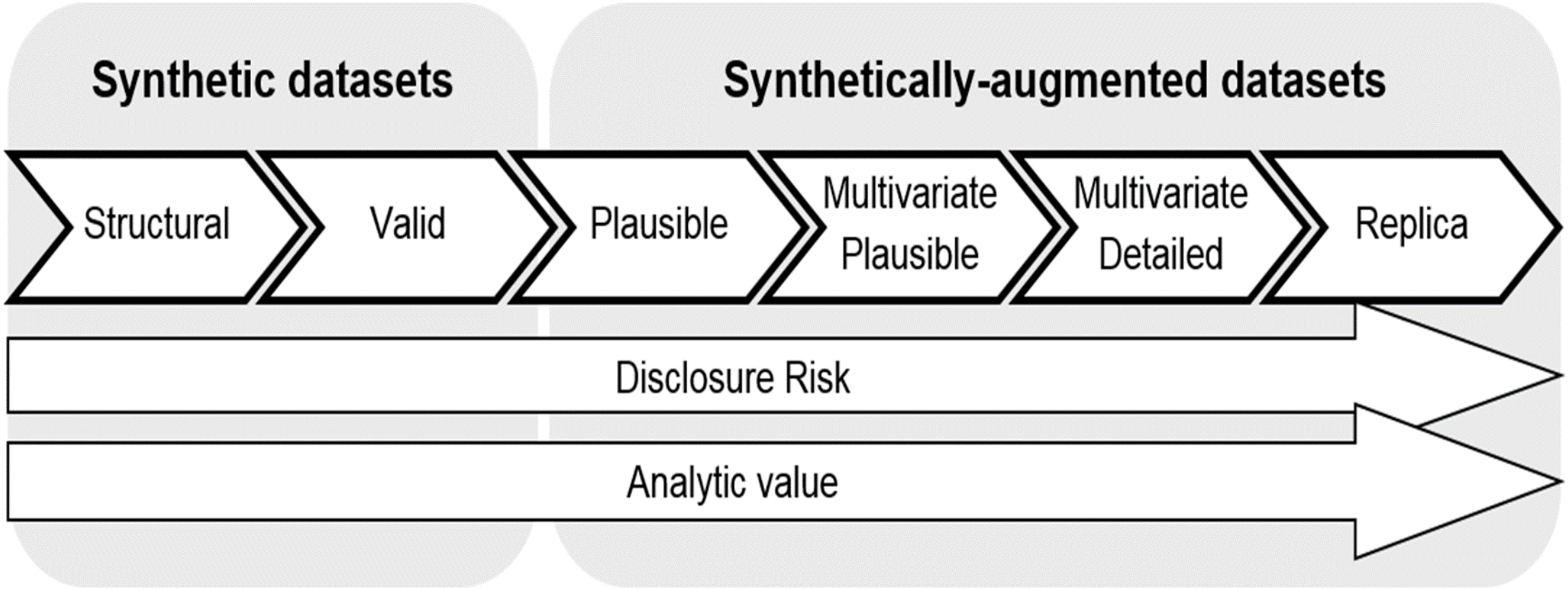
US Census Bureau

‘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’

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

Spectrum

- Low fidelity, high fidelity



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

Scoping

Investigate what other data organisations are doing around synthetic data: RSHs, PHS, ONS, NHS-NSS, HDRUK, ADR-UK

- What experience do you have in using synthetic data?
- What do you use synthetic data for? What would you like to use it for?
- What tools are you using/planning to use?
- What are the issues/benefits?

Discussions used to identify what the RDS synthetic data strategy might include

Questions

- Who are you and what is your interest in synthetic data?
- What is your experience with/understanding of synthetic data?
- What do you think are the benefits and issues?
- What can RDS do to help?

Breakout rooms – session A

Research Data Scotland

Dr Lynne Adair (Forrest)
Data Curation Manager

RDS strategy for the production and research use of synthetic data for Scotland

Research Data Scotland

Synthetic data landscape

- Several organisations have created synthetic data:
  - Ad-hoc, low-fidelity synthetic data using bespoke code in R/python/other
  - Synthpop or other tools
- For production of high-fidelity data, and scaling up of synthetic data production, synthetic data tool would be useful
- Work needs to be done to determine the most suitable tool(s):
  - Ability to deal with different data types and relationships, handle large numbers of variable categories and deal with temporal data
  - Different tools for different fidelity requirements?
  - Commercial v open-source
  - Commercial tools – expense v utility

Synthetic data landscape

- Need to address understanding and privacy concerns – public and data controllers
- Information Governance (IG) challenges on trying to release high fidelity data
- How to measure how closely the synthetic data resembles the real thing and thus is a disclosure risk
  - Standardisation of the different types of synthetic data and the terminology around this is required

Uses of synthetic data

- Training in using linked administrative data
- Data discovery (augment metadata catalogue)
- Code development:
  - Writing and testing code before full data access is available
  - To limit safe haven access to real data
- AI/Model training

Synthetic data benefits

- **To researchers and data controllers:**
  - Upskill users in use of messy admin data by using synthetic data as a training resource
  - Better data discovery (can augment meta data catalogue)
  - Speed up time to begin analysis before full permissions granted
  - Reduce time needed in safe haven and with access to real data
- **Generally:**
  - Improve access to data for research

Issues to address

- Different levels of fidelity for different use cases?
- Access methods/location?
  - Safe haven/safePod
  - Data in safe haven but accessed via VPN from home
  - Dataset released to researcher after signing an undertaking form
  - Published on website or accessed after registering
- Level of training/accreditation required?
  - None, accredited researcher, Safe Researcher Training
  - Depends on fidelity/location/purpose

RDS Plans

- Set up a Scottish working group to look at synthetic data needs, governance, and access for different organisations
- Survey researchers/users on their synthetic data requirements
- Speak to data controllers re their understanding and concerns around synthetic data
- Supply IG expertise
- Public engagement

RDS Plans

- Fund 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
- Investigate automating synthetic data production
- Set up a UK-wide synthetic data group (with HDRUK, ADR-UK and DARE UK) for collaboration and co-ordination of work

Proposed Outcomes

- Example **synthetic datasets**
- A **data/tools matrix** to allow selection of the most appropriate tool and level of fidelity required for production of different types of synthetic data
- A clear **synthetic data classification system** in terms of fidelity and risk that can be used when speaking to data controllers
  - Specify training and IG requirements, access methods and location for each synthetic data classification
- Potentially, for the future: Work with data controllers to **produce Synthetic datasets** for training, data discovery and code development **on an ongoing basis**

Questions

Synthetic data use cases – what is most useful SD purpose?

- Training in using linked administrative data
- Data discovery (augment metadata catalogue)
- Code development:
    - Writing and testing code before full data access is available
    - To limit safe setting access to real data
- AI/Model training

High fidelity or low fidelity synthesis?

Access methods?

Breakout rooms – session B

Research Data Scotland

Feedback

Research Data Scotland

Thank You

Dr Lynne Adair (Forrest)
Data Curation Manager
lynne.adair@researchdata.scot

Twitter:
@DrLynneAdair
@RDS_Scotland
Website: https://researchdata.scot