Dr Lynne Adair (Forrest)
Data Curation Manager

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

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

Founder Members of RDS

The Scottish Government
Riaghaltas na h-Alba

Public Health Scotland

The University of Edinburgh

National Records of Scotland

RDS

Joining Members of RDS (2022)

University of Glasgow

1495
University of Aberdeen

University of Dundee

RDS principles

- Only enable access to data for research for public good
- Only access data once an individual's personal identity has been removed
- All data is always kept in a controlled and secured environment
- All income that RDS generates will be re-invested into service
- RDS will be transparent
- Firms that access public data for the public good through RDS will share any commercial benefits back to improve public services

LA Background

- Admin data researcher for 10 years
- Research Support for 4 years
- Worked at ADRC-S, SCADR and Scottish Longitudinal Study (SLS)
- Developed synthetic datasets for SLS researchers using Synthpop package
- Worked at RDS for 11 weeks…
- Remit: develop a synthetic data strategy

Synthetic Data Definition

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

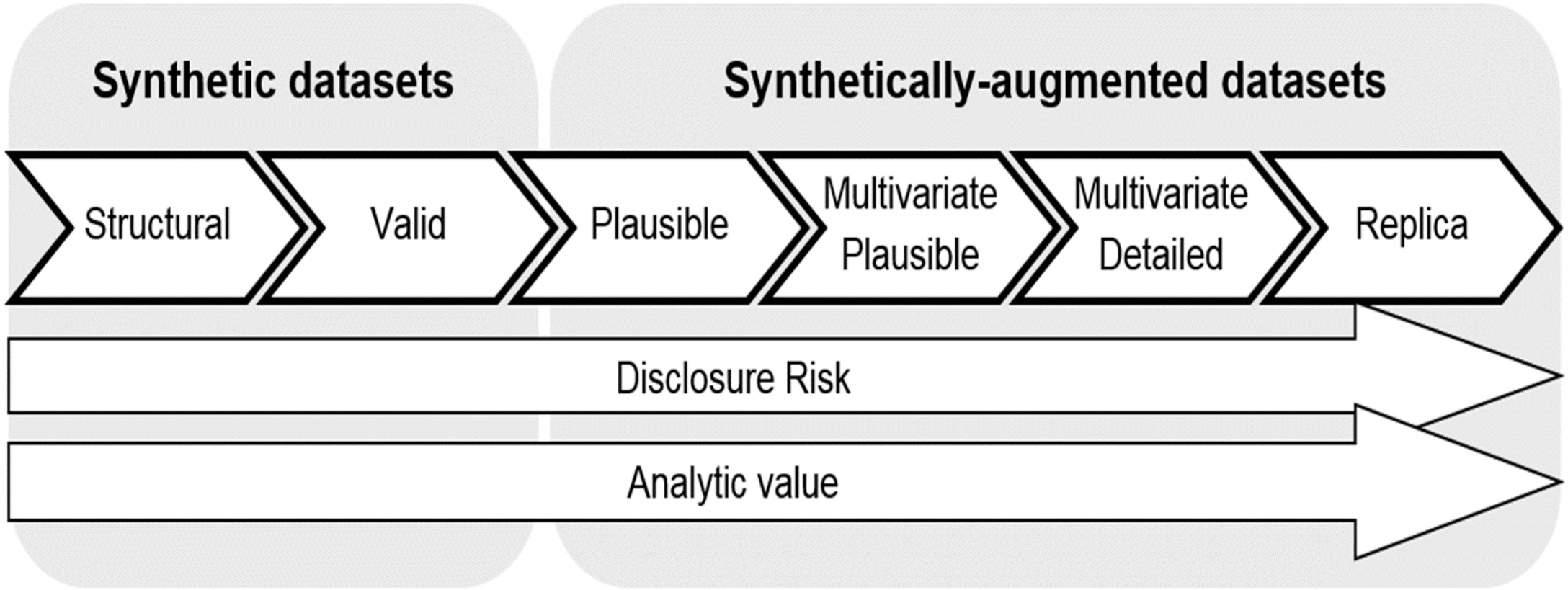
US Census Bureau

‘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 to 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, NSS, HDRUK, ADR-UK

- What experience do you have in using synthetic data?
- What purposes 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 you've experienced?

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

Synthetic data landscape

- Several organisations have created ad-hoc, low-fidelity synthetic data using bespoke code in R/python/other, and this is fairly easy to do
- 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

- Addressing 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 setting access to real data
- AI/Model training

Synthetic data benefits

- **To researchers and data controllers:**
  - Upskill users in use of admin data by using synthetic data as a training resource
  - Better data discovery (can augment meta data catalogue)
  - Reduce time needed in secure settings and with access to real data
- **Public/Generally:**
  - Greater use of data

Access methods/location

- Safe haven
- Data in safe haven but accessed via VPN from elsewhere
- Analytical Workbench
- Dataset released to researcher after signing an undertaking form
- Published on website

Training/Accreditation Requirements

- None
- Researcher at approved institution
- RDS Approved researcher accreditation (RDS ARA)
- RDS ARA + ONS Safe Researcher Training

Plans

- Set up a Scottish working group to identify similarities and differences in 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 and legal expertise
- Public engagement

Plans

- Fund work to :
  - Investigate synthetic data tools and synthetic data requirements
    - Evaluate and compare commercial and open-source tools
    - Cost-benefit analysis
    - Different solutions for low- and high-fidelity synthesis?
  - Develop a standardised synthetic data classification system to help understand the fidelity and risk of the synthetic data
  - Develop an example synthetic dataset (low- and high-fidelity versions)

Proposed Outcomes

- A test **synthetic dataset**
- 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
- For the future: **Synthetic datasets** for training, data discovery and code development

Questions

- What experience do you have in using synthetic data?
- What purposes 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 you've experienced?
- Anything we should consider that hasn't been covered?

Thank You

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

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