RDS synthetic data strategy draft 1.2: Summary

Introduction

Research Data Scotland (RDS) is working to improve the economic, social and environmental wellbeing in Scotland by enabling access to and linkage of data about people, places and businesses for research in the public good. Within this remit, RDS, working together with external partners and other data organisations, aims to develop a coordinated strategy for the production and use of synthetic data in Scotland. RDS will provide system leadership, facilitation, information governance support, resource funding and other support as required.

Synthetic data is '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, not by pointing to any original individual in the data set' (1).

The initial task was to investigate what other, similar, administrative data organisations were doing around synthetic data, both in Scotland and beyond. Conversations were held with all four Regional Safe Havens (RSHs), Public Health Scotland (PHS), Office for National Statistics (ONS), Health Data Research UK (HDRUK), and NHS National Services Scotland (NSS) and others. These discussions were used to identify what the RDS synthetic data strategy might include.

Issues raised in conversations

Several organisations have created ad-hoc, low-fidelity synthetic data using bespoke code in R/python/other, and this is fairly easy to do. However, to produce high fidelity data and scale up synthetic data production, some sort of synthetic data tool would be useful. Work needs to be done to determine the most suitable tools, in terms of their ability to deal with different data types and relationships, handle large numbers of variable categories and deal with temporal data. However, some tools from commercial companies are very expensive and so expense has to be weighed against the utility. Open-source tools are available that can create high fidelity synthetic data. User requirements also need to be scoped.

The issue of Information Governance (IG) challenges on trying to release high fidelity data was raised. IG expertise would be useful to help with this. Privacy protection, and how useful this was as a technique, was discussed, as well as how to assess the fidelity and disclosure risk of synthetic data. One suggestion was to produce a framework for validating the synthetic data against the real data, to make the decision on what is safe to release much easier.

Different uses and types of synthetic data

Synthetic data can be utilised for a variety of purposes but the main uses for RDS are likely to be:

- Training in using linked administrative data
- Data discovery (augmenting the metadata catalogue)
- Code development – writing and testing code:
    - Before full data access is available
    - To use outwith the safe setting to limit the requirement for safe setting access until running final models
- AI/Model training

There is a synthetic data spectrum based on the fidelity of the synthesis. The higher the fidelity the more like the real data the synthetic data is. But with the increase in analytical value for the user comes a greater disclosure risk and so a balance has to be found. We need to consider what level of fidelity is necessary and this may differ depending on the use of the data.

We also need to consider how the synthetic data is accessed and where it will be held. This again may vary depending on the fidelity of the data and what it is being used for, as may the training and accreditation requirements of the users.

RDS proposal

Work for RDS to lead on:

- 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
- Consult with data controllers and the public re their understanding and concerns around synthetic data
- Explore options with data controllers for a synthetic data demonstrator project
- Supply IG expertise
- Investigate whether we can ingest/hold and make more widely available the SLS/SCADR Admin data training synthetic dataset that is already developed

Work for RDS to fund partner organisations to conduct:

- Investigate different synthetic data tools and synthetic data requirements (led by RSHs)
    - Evaluate and compare commercial and open-source tools
        - Conduct a cost-benefit analysis
        - Consider different solutions for low and high-fidelity synthesis?
- Provide guidance and advice for data holders on how to evaluate the disclosure risks from synthetic data (led by Professor Gillian Raab)
    - Develop a standardised synthetic data classification system to help understand fidelity and risk
- Investigate ways of automating synthetic data production
- Develop an example synthetic dataset (potentially low, medium, and high-fidelity versions) as one of the RDS demonstrator projects

Deliverables

- Develop a coordinated strategy for the production and research use of synthetic data for Scotland (potentially aligning with ADR-UK, HDRUK and ONS)
- Produce a trial synthetic dataset
- Produce an evaluation of synthetic data tools
- Produce 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
- Produce 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
- Produce a public facing explanation of synthetic data (with SCADR?)

Future work

- Work with data controllers to produce synthetic datasets for training, data discovery and code development on an ongoing basis

1. Accelerating public policy research with synthetic data: a report from the Behavioural Insights Team: Dr. Paul Calcraft, Dr. Iorwerth Thomas, Martina Maglicic, Dr. Alex Sutherland Accelerating public policy research with synthetic data - ADR UK