Synthetic data position statement

An interim ADR UK position statement on synthetic data

Updated July 2023

Economic and Social Research Council

1 Purpose

The purpose of this statement is to communicate the current position of ADR UK (Administrative Data Research UK) on the issues of synthetic data production, access, governance and use. These issues are evolving, so ADR UK's position will be updated in line with new findings and developments.

2 Background

Synthetic data, also known by other names such as artificial, dummy, simulated or fake data, has been emerging as a key area of development for supporting administrative data for research. ADR UK has identified the need for such datasets for:

- training purposes
- exploratory analysis to determine if the data is going to be helpful for a particular research project
- instances where researchers need to progress with developing their code, understanding the structure of the data, and testing different statistical methods before they can get access to the realdata.

Where health data is held securely, there is a similar demand for this resource.

This statement uses the term 'low-fidelity' synthetic data which we define as a version of the data that resembles the real data but does not include any information about real individuals. It also does not preserve any relationships between different pieces of information meaning it is a powerful tool in supporting data security. High-fidelity data, on the other hand, preserves the relationships between information, but no points correspond to real individuals.

In 2021, ADR UK commissioned the Behavioural Insights Team to engage government departments in a discussion about synthetic data and their level of comfort in and barriers to producing it. The report documents various concerns related to:

- data quality and how well it reproduces the relationships and characteristics of the real data
- tensions between protecting privacy and retaining a level of utility of synthetic data
- systemic or technical barriers such as a lack of knowledge and understanding, ethical and legal barriers, and inconsistent technological support being available for users.

The report recommended that ADR UK should:

- encourage the use and sharing of low-fidelity synthetic data across government and with researchers
- expand the use of synthetic data for training and improve the efficiency of live projects
- develop a cross-government repository of synthetic data, accessible to government analysts and accredited researchers without a specific project proposal.

3 ADR UK approaches across the partnership

There are various approaches to the creation and use of synthetic data across the ADR UK partnership. Issues of governance and access also vary, as does the level of fidelity of different datasets. We are also aware that different levels of fidelity of any one synthetic dataset might require different levels of access and/or governance. Below, we present summaries from each national partnership, reflecting their approach through use cases and supporting the value of debate and difference.

3.1 ADR England

The ADR UK Strategic Hub (which manages the ADR England portfolio) is keen to make low-fidelity synthetic data as widely accessible as possible in England so that researchers and analysts might develop their awareness and understanding of the potential of real data and their ability to use it. Synthetic datasets can be used for training and for preparatory work by researchers and analysts. As well as reducing data-owner concerns, having low-fidelity synthetic datasets openly accessible in the public domain could expedite access to real data. Researchers may also be able to use synthetic data to develop better-informed project proposals with less requirement for protracted discussions with data owners about the viability of their research.

Currently, the ADR England portfolio has low-fidelity synthetic datasets available for magistrates' and Crown Court data, and for the Grading and Admissions Data (GRADE) dataset. Access is obtained by emailing the relevant data owner. The Office for National Statistics (ONS) has created a synthetic version of the Annual Survey of Hours and Earnings (ASHE) dataset that was made accessible via the UK Data Service to test the concept of an external repository.

3.2 ADR Northern Ireland

Having considered and evaluated different synthetic data types, ADR Ireland decided that the model of 'univariate synthetic data' (low-fidelity) struck the appropriate balance between analytical value and disclosure risk. Univariate synthetic data is created by working through each variable completely separately using the percentage distributions from the real data to randomly generate fabricated data.

The main intended purpose of this approach is to facilitate remote writing of code and general preparatory work without needing access to the real data and a secure room. It can also serve as a familiarisation and training aid for newer researchers. Currently, univariate synthetic data has been created for the Northern Ireland Longitudinal Study (NILS), the Earnings and Employees Study, and the Education Outcomes Linkage project. Access is obtained by submission of a univariate synthetic data user agreement form.

3.3 ADR Scotland

The Scottish Centre for Administrative Data Research has created the R package '*synthpop*' for creating and evaluating synthetic data. Since it was first made available as an open-source package in 2014 it has been widely used by a variety of groups including the ONS. As well as supporting a variety of methods of creating synthetic data, the package provides tools for evaluating the utility of synthetic data and allows various methods to minimise disclosure risk. The popularity of the package has increased in recent years and now consistently has more than 2,000 downloads per month.

Staff at the Scottish Longitudinal Study (SLS) use the package to provide datasets for preliminary analysis to users of the SLS and synthetic data sets have been created to use in training courses.

Although the use cases above require high-fidelity synthetic data, the package can also be used with minimum effort to create low-fidelity synthetic data. It has been used in this way to create low-fidelity synthetic data for the Annual Survey of Hours and Earnings which has been released as a pilot project for a few users (separately from the ONS project mentioned above).

3.4 ADR Wales

ADR Wales is currently exploring how it will facilitate the use of low-fidelity synthetic data to benefit administrative data researchers and those in the early stages of their careers while maintaining appropriate governance and security. They view synthetic data as an integral part of their offering that will enhance the user experience and will help to develop research skills in the wider data-intensive science community. As a new and developing field, ADR Wales is committed to offering synthetic data wherever possible as an alternative means of producing research outputs. The partnership proposes to offer access to synthetic resources via a sandbox training environment with an integrated and appropriate governance wrapper.

SAIL Databank has a variety of synthetic datasets available, initially covering Welsh health data derived from primary and secondary care records and has committed to creating synthetic versions of all core datasets which SAIL Databank hosts. This will include wider records across:

- the care pathway
- administrative data such as education records
- other population-level datasets representing Wales.

ADR Wales is currently in consultation as to the exact policies and processes of sharing and using this synthetic data, but envisages initial use cases will be:

- technological and analytical methodology development
- a major training platform to support administrative data researchers in using and understanding large-scale longitudinal health and administrative data.

The partnership will produce synthetic data based on the real datasets held in the SAIL Databank both in terms of the complexity of the variables and the scale of the datasets. To generate a rich training environment, they aspire to have synthetic data holdings which match the real-world data. This means that the training can address issues related to the real-world complications of dealing with administrative data.

SAIL Databank plans to create synthetic versions of all its data holdings so that it can offer linked synthetic data covering all the areas currently served by SAIL Databank. SAIL Databank will focus initially on low-fidelity approaches to demonstrate the utility of synthetic data, prior to investing in more advanced approaches.

Plans for synthetic medical imaging data are under development – higher-fidelity synthetic data will be developed for cohort datasets that require multi-variant synthetic data to have any utility.

4 Public attitudes on synthetic data

Public engagement is vital to the work we undertake and we are committed to engaging the public around synthetic data and its use. The topic of synthetic data has been briefly explored in a small number of public engagement exercises, most prominently in May 2023 with ADR Scotland's public panel. The panel explored what synthetic data is, its use, and Research Data Scotland's approach to synthetic data. The results of the recent engagement with the ADR Scotland public panel highlight the need to further explore public understanding and acceptability of synthetic data. This reiterates the results of other ADR UK public engagement on the creation and use of synthetic data, which has thus far been met with mixed feelings and caution. We are actively exploring further public consultation in this area and will announce future plans around this in due course.

5 Acknowledgements

This statement is provided by ADR UK (Administrative Data Research UK). ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation).

6 Contact

Name: Emily Oliver, Head of Research and Capacity Building, ADR UK Strategic Hub
Email: Emily.Oliver@esrc.ukri.org

Visit the ADR UK website

@ADR_UK