THE BEHAVIOURAL INSIGHTS TEAM

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

14th December 2021

Funded by:

ADR UK Data-driven change

UKRI Economic and Social Research Council

Executive summary

Governments need to understand how citizens really behave, and how their policies affect people’s lives. At the same time, citizens demand better policy and government services. Data, used well, can dramatically accelerate both of these aims. But expanding the use of data in government also increases concern about individual privacy. What if society-level patterns in behaviour and outcomes could be easily analysed by researchers to inform policy and services, without risking the privacy of any individual citizen? An idea from a Harvard professor in 1993[1] may provide exactly that: synthetic data.

What is synthetic data?

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 (e.g., an athlete’s height will usually be between 1.5 and 2.2 meters, and would never be 1 kilometer), but it is chosen randomly from the range of possible values, not by pointing to any original individual in the data set.

Data that is generated in this way reveals very little, if anything, about any individual in the original data set, but still represents the data well as a whole. It can therefore be used:

- To do some types of exploratory analysis on data without requiring a fully secure environment and information governance procedures.
- To train researchers on how to handle particular data sets in unique or challenging formats (common in administrative data) without requiring full access to secure environments for all trainees.
- To improve the efficiency and safety of analysing personal or confidential information, by allowing researchers to write and test their analysis code on low-risk synthetic data before getting access to the real data.
    - It may also be desirable for researchers never to access the real data, but to send their code to the data holder, for the data holder to run securely, sending back only the aggregate results.

Approach

If synthetic data is to fully realise its potential, we should first understand how it is perceived by practitioners and other stakeholders in government. We therefore:

1. Interviewed researchers at the Northern Ireland Statistics and Research Agency, Administrative Data Research Scotland, the Office for National Statistics (ONS) and the ONS Secure Research Service, the Ministry of Justice, the Department for Education, and the UK Data Service.
2. Had further conversations with No. 10 Data Science, the Ministry of Housing, Communities & Local Government, and the Open Data Infrastructure for Social Science and Economic Innovations (Netherlands).

During this engagement, we proposed an approach to unlocking more data value across government that received generally positive support. This approach is to encourage the

[1] "Discussion: Statistical Disclosure Limitation". *Journal of Official Statistics*. **9**: 461–468. 1993.

generation of low-fidelity synthetic data to be stored in a cross-government repository, accessible to government researchers and accredited external and academic researchers. This has the following benefits:

- It would help researchers understand what data is available, to inspire and hone better research questions.
- It would help researchers *find* the right data for existing research questions and understand how easy the data will be to work with.
- It would allow researchers to write and test their analysis code outside of secure settings, and therefore:
    - Reduce the amount of time researchers need to spend in secure settings, or with direct access to confidential or personal information.
    - Prevent delays in data access from stalling analysis projects.

Given there is no script or package available that generates only this safer kind of low-fidelity synthetic data, we decided with Administrative Data Research UK (ADR UK) to:

3. Develop a prototype Python notebook[2] that makes it easy for a researcher to generate low-fidelity synthetic data.

This report describes the potential uses of synthetic data in more detail, findings from our short engagement exercise, our prototype development work, and suggested next steps.

Recommendations

- ADR UK should encourage the use and sharing of low-fidelity synthetic data across government and with researchers for:
    - enabling researchers to rapidly discover if the data for a given policy challenge is available and usable
    - writing and testing code for a project before full access is available, reducing the impact of data delays
    - reducing time needed in secure settings and with access to sensitive data.
- ADR UK should expand the use of synthetic data for training, so researchers can be exposed to relevant idiosyncratic data sets earlier, improving researcher efficiency on live projects.
- A cross-government repository of synthetic data should be developed, accessible to government analysts and accredited researchers without a specific project proposal.
    - This will help researchers discover available data, design more informed project plans and help refine answerable research questions with policy colleagues.
    - A semi-automated pipeline could be established to generate low-fidelity synthetic data as a routine end-of-project activity, or upon depositing data with ONS.

The release of our prototype Python notebook[2] supports all of these recommendations.

[2] https://bit.ly/synthetic-generator-colab

Table of contents

1. Introduction

In 2020, the Behavioural Insights Team (BIT) worked with Administrative Data Research UK (ADR UK) to understand behavioural barriers to data linkage (connecting up data about the same person from multiple different sources, for example to understand the effect of new education programmes on long-term health and wellbeing) and data sharing for government research[3]. While overcoming many barriers will require a complex combination of changes to policy and culture, privacy technologies, for which the UK is globally competitive[4], may provide a way out of the bind. In this report, we look at one particularly promising privacy technology called synthetic data.

What is synthetic data?

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. Please see the summary table on the following page. Each piece of information in the data set is meant to be plausible (e.g., an athlete's height will usually be between 1.5 and 2.2 meters, and would never be 1 kilometer), but it is chosen randomly from the range of possible values, not by pointing to any original individual in the data set.

Data that is generated in this way reveals very little, if anything, about any individual in the original data set, but still represents the data well as a whole. But using synthetic data is not entirely without risk, and it is not currently well understood by policy makers and the civil service. This report explains what synthetic data is and the different ways it can be used to drive value in government.

Outline of the document

The remainder of this section provides a brief overview of what synthetic data is, why we might want to make use of it, and how different forms of it have been classified as having different levels of risk for individual privacy or confidentiality.

In the subsequent section we then summarise the results of our engagement exercise, focusing on how synthetic data is already being used, ideas for future applications, as well as the doubts and concerns practitioners have about it.

We then discuss the prototyping work we have done towards simplifying the technical process of synthetic data generation.

[3] Gibbons, D. *et al*. (2021) 'Applying Behavioural Insights to Cross-government Data Sharing' BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021).

[4] Safer technology, safer users: The UK as a world-leader in Safety Tech. (2020) Independent report for the Department for Digital, Culture, Media & Sport. Available at https://www.gov.uk/government/publications/safer-technology-safer-users-the-uk-as-a-world-leader-in-safety-tech/safer-technology-safer-users-the-uk-as-a-world-leader-in-safety-tech (Retrieved 24 Feb. 2021).

In the final section we discuss recommendations based on our recommendations for encouraging the use of synthetic data for administrative data in government to benefit public services and policy.

Simplifying terminology: High- and low-fidelity synthetic data

Existing terminology for synthetic data is complex and often unintuitive to those new to the concept: it may be subcategorised as 'structural', 'univariate', 'multivariate', 'multivariate plausible' and more (see Table 1). Fundamentally, synthetic data can be produced at different levels of fidelity, that is, to more or less accurately reflect the original data set. Making more accurate synthetic data is more complicated and computationally intensive, and more likely to disclose personal or confidential information. However, it offers more value to the user of the data set because they can learn more from it.

In this paper we advocate for a simple, yet critical, distinction between high-fidelity and low-fidelity synthetic data. We define this below and illustrate it on the following page.

Low-fidelity synthetic data

At the lowest levels of fidelity, synthetic data may reflect the original data only in its format: how it is laid out, and what types of information it contains. For example, let us take a data set containing the height and weight of every athlete at the 2016 Olympics. From a low-fidelity synthetic version of this data set, you would be able to tell that the original data set contains information about weight (in kilograms) and height (in meters). But low-fidelity synthetic data does not preserve any relationships between different pieces of information, so you would not be able to see that, on average, weight increases with height.

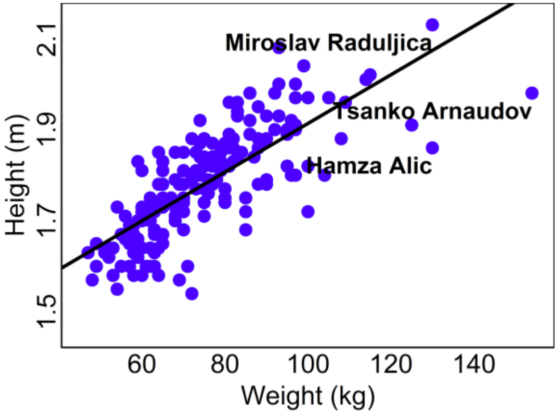
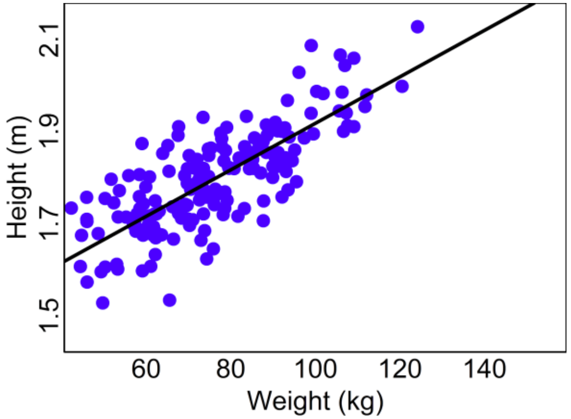
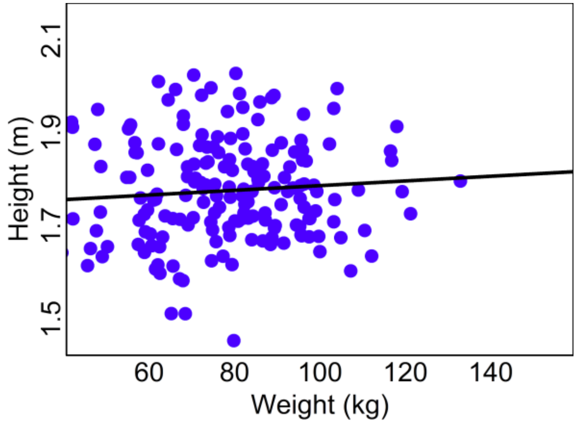
Historically, low-fidelity synthetic data has been referred to as 'dummy' or 'fake' data. This usually refers to the lowest possible fidelity, where values may not even be plausible (e.g., a height of 1,000m), just of the same type (a number of meters).

High-fidelity synthetic data

At the highest level of fidelity, many complex relationships between information in the data set may still be present, despite no individual row of data referring explicitly to any original individual from the real data set. In this scenario, you would be able to get a reasonably accurate estimate of how much, on average, height tends to increase with weight, while knowing nothing of any individual's height or weight. Relationships between variables, such as between height and weight, are detected and conserved automatically when generating high-fidelity synthetic data.

Synthetic data and disclosing information

Because the records in a synthetic data set are artificial and do not (except by rare coincidence) correspond to records in the real data, in principle there are fewer concerns regarding disclosure risk when handling synthetic data. As we shall discuss, in reality the true picture is less clear cut: high-fidelity synthetic data may still be able to disclose personal or other sensitive information under certain circumstances, but in principle these risks can be mitigated through the application of advanced privacy preserving techniques.

Original Data	High-Fidelity Synthetic Data	Low-Fidelity Synthetic Data
Height and weight of every athlete at the 2016 Olympics.	Each point is randomly generated from the general relationship between height and weight at the 2016 Olympics.	Each point is randomly generated from the general statistics of height and the general statistics of weight at the 2016 Olympics.
● **Every point corresponds to a real person**. ● Weight **increases** with height, on average.	● **No points correspond to a real person**. ● Weight **increases** with height, on average.	● **No points correspond to a real person**. ● Weight is **not visibly related** to height.
		

What the data can tell us

● Data format: names and layout of columns, including units (meters, kilograms) of any information. ● **Every athlete's** weight and height. ● **Exact** minimum, average and maximum weight/height of the athletes. ● Weight increases with height, with a **correlation of ~60%**.	● Data format: names and layout of columns in the original data, including units (meters, kilograms) of any information. ● **Approximate** minimum, average and maximum weight/height of the athletes. ● Weight increases with height, with a **correlation of ~60%**.	● Data format: names and layout of columns in the original data, including units (meters, kilograms) of any information. ● **Approximate** minimum, average and maximum weight/height of the athletes.

What the data is useful for

● Performing all types of analysis. ● **Requires**: Secure environment, full governance and data sharing agreements.	● Understanding what types of information is held in a data set and how usable it is. ● Writing analysis code that can later be run on the real data. ● Identifying possible relationships. ● **Requires**: Less stringent governance; data sharing agreements and secure environment may be unnecessary.	● Understanding what types of information is held in a data set and how usable it is. ● Writing analysis code that can later be run on the real data. ● **Requires**: Minimal governance; data sharing agreements and secure environments very unlikely to be required.

Why make use of synthetic data?

We consider three proposed uses for synthetic data in government:

- For doing some types of exploratory analysis on data without requiring a fully secure environment and information governance procedures.
- For training researchers on how to handle particular data sets in unique or challenging formats (common in administrative data) without requiring full access to secure environments for all trainees.
- To improve the efficiency and safety of analysing personal or confidential information, by allowing researchers to write and test their analysis code on low-risk synthetic data before getting access to the real data.
    - It may also be desirable for researchers never to access the real data, but to send their code to the data holder, for the data holder to run securely, sending back only the aggregate results.

Synthetic data can be an accelerator for better use of data across government. Primarily, this is achieved by its ability to reduce barriers and bottlenecks.

Synthetic data removes barriers to accessing data

To write analysis code or perform exploratory research, researchers and analysts require access to data that is in the correct format and (in the latter case) resembles the true data in some or all of its relationships. While the real data set will fulfil these requirements by definition, access to it requires navigating barriers:

- External researchers must often undergo a stringent access process that leads to delays and that may require that a researcher be physically present at a secure location to access data.
- Government analysts may find it useful to be able to explore datasets from other departments as part of their work, but legal or operational constraints may prevent this.

These impediments are costly. If preliminary testing or exploratory analysis could be carried out by external researchers prior to having access to the real data, more studies could be carried out and more research value can be realised. The ability to perform these steps without needing to be physically present in a location (as is sometimes required) is also vital if a researcher’s ability to access that location is limited (for instance, during a COVID-19 lockdown).

Data access bottlenecks impede policy formulation

Similarly, if government analysts cannot access data that provides important information available to other departments in a timely fashion, or at all, this can lead to delays and policy mistakes, at best wasting taxpayer’s money on badly conceived plans and at worst costing lives.

If data was easily available across government that allowed for preliminary testing and exploratory analysis, it would be quicker to determine what data is available that is relevant to policy formulation. As it is, it may be quite challenging to determine what data is available that is relevant to a given policy problem, and whether that data will be easy to work with, without gaining access to it directly. Gaining access usually requires a specific request, clearance, and an extended information governance process, but it is hard to know if that effort will be worth it. Greater visibility of data held across government could support more responsive, refined and targeted policy making.

Data access bottlenecks impede crisis response

During a crisis situation such as the Covid-19 pandemic, the unimpeded flow of relevant information between government departments and ministries is vital to the formulation of a viable response. If it was possible to perform exploratory analysis of data held by other departments, it could be quicker to identify which data is relevant to the crisis at hand and what research questions can and should be answered.

Examples of synthetic data use and research in the US and UK

US Census Bureau

The US Census Bureau provides high-fidelity synthetic data built on a linked underlying data set combining the census with administrative tax and benefit data (Survey of Income and Program Participation data)[5], covering the years from 1984 to 2008. They also plan to release high-fidelity synthetic data for the 2020 Census that incorporates a form of differential privacy[6] in order to reduce the risk of being able to identify individuals from the synthetic data. Differential privacy is an advanced technique that aims to reduce and precisely quantify the probability that high-fidelity synthetic data contains data identifying an individual by introducing a controlled level of error (small random variations).

NIST Differential Privacy Synthetic Data Challenge

In 2018 the US National Institute of Standards and Technology issued a challenge aiming to determine the best algorithms for generating differentially private synthetic data[7]. The aim was to spur the development of these algorithms, and the data gathered resulted in the production of a benchmark of existing approaches[8].

ONC Synthetic Health Data Generation

The US Office of the National Coordinator for Health Information Technology is developing modules for the Synthea application that uses publicly available health data to generate synthetic data representing plausible patient histories[9]. This data is intended to be used for testing and refining analyses prior to access being granted to real data.

SYLLS

The Synthetic Data Generation for UK Longitudinal Studies service provides synthetic data showing longitudinal transitions derived from UK longitudinal studies data. It is part of a case study discussed later on in this report.

NILS and SLS

Synthetic data from both the Northern Ireland Longitudinal Survey and the Scottish Longitudinal Survey are available to researchers as ‘bespoke’ data sets. NILS synthetic data is low-fidelity, whereas high-fidelity SLS synthetic data is generated using the `synthpop` R package[10]. `synthpop` has also been used to generate high-fidelity synthetic data sets for training purposes. These uses are discussed in a case study later on in this report.

MHCLG Troubled Families Data

Low-fidelity synthetic data was produced for this project and made available to researchers on request for the purposes of understanding the contents of the data set, and developing and testing analysis code prior to analysing the source data. It was also suggested that analysis code could be supplied to MHCLG to be run in their secure environment, releasing only aggregate results to external parties.

Open Data Institute Tutorial

The Open Data Institute have produced a tutorial based on previous work with NHS England explaining how to use the DataSynthesiser application to produce both high- and low-fidelity synthetic data[11].

National Cancer Registry

The Simulacrum[12] data set is a multivariate synthetic data set based on real patient data taken from Public Health England's National Cancer Registration and Analysis Service, intended to allow exploratory analysis and code testing by researchers. As with many high-fidelity synthetic data sets, it is not a full substitute for the final data because not all properties of the underlying data are reproduced.

ONS Data Campus

Researchers at the ONS Data Campus have begun to investigate the feasibility of generating high-fidelity synthetic data sets using a variety of techniques[13].

Ministry of Defence

The Ministry of Defence has also been investigating the potential of producing realistic synthetic data suitable for sharing with external researchers who wish to perform analyses of data in cases where the real data would reveal sensitive information, such as location or the performance of particular kinds of equipment[14].

[5] United States Census Bureau. (2019) *Synthetic SIPP Data.* Retrieved from https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html on 22 Feb. 2021.

[6] United States Census Bureau. (2021) *Disclosure Avoidance and the 2020 Census - Census Bureau.* Retrieved from https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html on 27 Jan. 2021.

[7] NIST. (2021) *2018 Differential Privacy Synthetic Data Challenge.* Retrieved from https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic on 23 Feb. 2021.

[8] Bowen, C., and Snoke, J., (2020) 'Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge', arXiv:1911.12704. Available at: https://arxiv.org/abs/1911.12704 (Accessed: 23 Feb. 2021).

[9] healthit.gov. (2021). *Synthetic Health Data Generation to Accelerate Patient-Centered Outcomes Research.* Retrieved from https://www.healthit.gov/topic/scientific-initiatives/pcor/synthetic-health-data-generation-accelerate-patient-centered-outcomes on 22 Feb. 2021.

[10] synthpop. (2019) *synthpop - Welcome.* Retrieved from https://synthpop.org.uk/index.html on 18 Jan. 2021.

[11] theodi. (2019). *Anonymisation with Synthetic Data Tutorial.* Retrieved from https://github.com/theodi/synthetic-data-tutorial on 22 Feb. 2021.

[12] simulacrum. (2021). *Home page*. Retrieved from https://simulacrum.healthdatainsight.org.uk/ on 22 Feb. 2021.

[13] Data Science Campus, ONS. (2020) *DSC-50 Synthetic data using generative models*. Retrieved from https://datasciencecampus.github.io/projects/DSC-50-Synthetic-data-using-generative-models/ on 22 Feb. 2021.

[14] Walker, K., (2020) *Synthetic data: Unlocking the power of data and skills for machine learning.* Retrieved from https://dataingovernment.blog.gov.uk/2020/08/20/synthetic-data-unlocking-the-power-of-data-and-skills-for-machine-learning/ on 23 Feb. 21.

Synthetic data and data linkage

Linking data involves taking two or more data sets containing records belonging to the same individual (a person, or a vehicle, or a house, and so on) and matching records corresponding to that individual together. Data linkage, as described in our earlier report[15], is critical for answering many important policy questions because the way people behave and interact with government services and policy in one area of life is likely to affect them in myriad ways, many of which may only be possible to detect by linking, for example, health, education, and employment data together.

Because synthetic data is artificially generated from a single data set, synthetic data sets generated from different data sets will not be linkable even if the original data sets are. Identifiers in the synthetic data sets will no longer match up because they are generated independently and randomly. To generate synthetic data that preserves relationships between two or more real data sets that are linked, the synthetic data must be generated from the real data *following* linkage. The preliminary linkage step should make use of GUILD guidelines[16] in order to minimise any potential errors in the linked data set from which synthetic data will be generated.

Office for National Statistics (ONS) classifications of synthetic data

A preliminary study of the use of synthetic data by ONS suggested that synthetic data could be classified into the categories shown in Table 1. While this classification does not account for the ways that methods such as *differential privacy* (detailed later) might permit more accurate synthetic data to be produced with greater privacy protections, it provides a useful framework for thinking in more detail about classes of synthetic data and the question of utility versus privacy for given data sets and applications.

In general, methods towards the bottom of the table are more faithful to the original data (higher fidelity; preserving more of its attributes and relationships) and allow a greater variety of use cases but pose a correspondingly higher risk of disclosing personal or confidential information.

What do the categories mean?

The ‘Synthetic’ category refers to synthetic data that is typically wholly artificial with little to no relationship to the real data, which at most reproduces the format and structure of real data. It may also rule out implausible or impossible relationships between data fields (so that certain structural rules are followed, e.g., all infants are unmarried).

The ‘Synthetic-Augmented’ category refers to data that is generated using techniques that reproduce some to many of the underlying statistical properties of the data. This includes (at the Synthetic-Augmented Plausible level) the distribution of each individual data field but not

[15] Gibbons, D. *et al*. (2021) ‘Applying Behavioural Insights to Cross-government Data Sharing’ BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021)

[16] Gilbert, R. *et al.* (2018) *GUILD: GUidance for Information about Linking Data sets* Journal of Public Health 40(1), 191-198. Available from: https://doi.org/10.1093/pubmed/fdx037 (Accessed 24 Feb. 2021).

the relationships between them, that is, it preserves the ‘univariate’ properties of the real data.

As more of these relationships (‘multivariate properties’) are preserved, the synthetic data becomes more realistic, but the disclosure risk begins to become equivalent to that of the real data (hence disclosure risk categorisations of ‘High’ to ‘Extremely High’; it is assumed here that the real data under discussion is of sufficient breadth and/or contains identifying information such that it carries a disclosure risk).

Table 1: Proposed ONS categorisation of synthetic data (based on Bates *et al.*[17])

	Category	Sub-category	Source	Relationships Preserved	Disclosure Risk and Data Set Value
Increasingly faithful to original source	Synthetic	Structural	• Available metadata • Values derived from open sources and implausible distributions; do not match true distributions	• Data types • Format	• No risk of disclosing information • Will produce worthless statistical analyses • Useful for basic testing of statistical analysis calculations (i.e., does the calculation understand the format of the data file correctly)
	Synthetic	Valid	• Available metadata • Values use plausible distributions and open sources; do not match true distributions	• Data types • Format • Missing values • No impossible values	• Minimal risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Sometimes plausible ranges etc can be disclosive • Useful for advanced testing of statistical analysis calculations (e.g., does the calculation correctly process variable types and account for missing values) • Will produce worthless statistical analyses
	Synthetically-Augmented	Plausible	• Real data set • Values generated from true distributions	• Data types • Format • Univariate distributions • Missing values	• Non-negligible risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Care must be taken with some kinds of data (e.g., names) • Useful for extended testing of statistical analysis calculations (e.g., does the calculation give plausible results for univariate statistics) • Will produce minimally useful statistical analyses
	Synthetically-Augmented	Multivariate Plausible	• Real data set • Values generated from true distributions, preserving some relationships between them	• Data types • Format • Univariate distributions • Missing values • Some relationships between variables	• High risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Useful for teaching and testing experimental approaches to statistical analysis • Will produce somewhat useful statistical analyses
	Synthetically-Augmented	Multivariate Detailed	• Real data set • Values generated from true distributions, more effort made to match joint distributions	• Data types • Format • Univariate distributions • Missing values • Joint distributions	• Very high risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Useful for teaching and testing experimental approaches to statistical analysis • Will produce somewhat useful statistical analyses
	Synthetically-Augmented	Replica	• Real data set • Values generated from true joint or conditional distributions • Deidentification techniques are applied	• Data types • Format • Univariate distributions • Missing values • Joint distributions • Lower-level patterns	• Extremely high risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis and is **critically important** • Can use in place of real data in statistical analyses; the results should be equivalent • Should be available only in secure research facilities

[17] Bates, A. G., Špakulová, I., Dove, I., and Mealor, A., (2019) 'Synthetic data pilot' ONS methodology working papers 16. Available at: http://bit.ly/ons-synthetic-spectrum (Accessed: 27 Jan. 2021).

What we mean by the ‘fidelity’ of synthetic data

When we speak of low-fidelity synthetic data in what follows, we mean data towards the low-utility end of the spectrum (Category and Subcategory in blue text in Table 1). As described on pages 7 and 8, these might require relatively light disclosure processes and so be subject to less onerous access requirements than either real data or high-fidelity synthetic data. High-fidelity synthetic data requires more stringent disclosure processes as it resembles the real data more closely than low-fidelity synthetic data.

2. Emerging themes on use of synthetic data

In this section we summarise attitudes towards and assessments of synthetic data gathered during our six engagement interviews, follow up conversations with policy and analytical colleagues, and desk research.

We have generally found that attitudes to high-fidelity forms of synthetic data (that pose non-negligible disclosure risk) vary considerably, including among the synthetic data specialists we spoke to.

However, those interviewed were generally positive about some proposed uses of low-fidelity data that we will detail later.

Finally, some scepticism about the viability of the advanced privacy protecting approach known as differential privacy has been expressed by analysts who have attempted to implement it.

Our engagement work also flagged a number of structural barriers to the widespread adoption of synthetic data, typically focused around departmental technical capabilities, lack of knowledge and legal concerns.

Finally, we summarise some discussions of the kinds of tools and pipelines that might be helpful in normalising or scaling the use of synthetic data. These are a proposal for the semi-automated generation of synthetic data, and a list of characteristics that a good synthetic data generation tool should have.

We also provide a case study of how synthetic data is already utilised by the UK Longitudinal studies as an illustration of what can and has been done to support its use. Features of how they make use of synthetic data also illustrate some of the points raised in our discussions of the above points.

Case study: SYLLS, the Scottish Longitudinal Study and `synthpop`

The Synthetic Data Generation for UK Longitudinal Studies (SYLLS) project[18, 19] provides synthetic data for longitudinal transitions derived from the base Longitudinal Study data, containing data fields of common interest to social science researchers (health, marital status, social grade, religion, birth and death rate estimates) in three separate sets: ONS Longitudinal Study (England and Wales), Scottish Longitudinal Study (SLS), and the Northern Ireland Longitudinal Study (NILS). All three are generally available with intent of familiarising any interested party with longitudinal data and the scope of questions it can answer.

Bespoke high-fidelity synthetic data service (Scottish Longitudinal Study)

A further bespoke synthetic data service is provided by the Scottish Longitudinal Study[20]. (As of 22 Jan. 2021, a similar service is available for NILS with respect to low-fidelity synthetic data only[21].) This uses the R package `synthpop`[22, 23] in order to create a synthetic data set for selected data fields that accurately reflects the underlying data by including data and relationships between variables that are absent from the SYLLS data set, and is available to approved researchers who may find it difficult to access the location of the SLS Safe Setting.

Due to concerns over the ultimate validity of the data set (it is not clear that `synthpop` will reproduce all features of the underlying source data) this synthetic data is intended only for testing analysis code; analyses intended for publication must ultimately be performed on the source data set. It is intended as a tool, not a substitute for the real data. Furthermore, researchers must liaise with SLS support staff in order to compare the results of their analyses on both source and synthetic data sets, in order to identify discrepancies that can be used to refine the approach used in future interactions with the service.[24]

There are several notable features of this approach to high-fidelity synthetic data:

- Unlike the SYLLS data, it is not generally available to the public, despite it in principle possessing a degree of privacy protection effectively that of state of the art techniques[25] (but there are some caveats[26]).
- Suspicion of potential introduction of error by the synthesis process entails that it should not be used for final analyses. (A similar view regarding exploratory versus final analyses is seen in the online centralised differential privacy[27] literature, although the use case is very different[28].)

These features also reflect concerns we have seen among the researchers that we interviewed, which will be discussed later in Data quality concerns.

[18] CALLS-HUB. (2021) *Synthetic LS data.* Retrieved from https://calls.ac.uk/guides-resources/synthetic-ls-data/ on 18 Jan. 2021.

[19] Dennett, A., Norman, P., Shelton, N. and Stuchbury, R. (2016) 'A synthetic Longitudinal Study data set for England and Wales"' *Data in Brief* 9(December 2016). Available at: https://doi.org/10.1016/j.dib.2016.08.036 (Accessed: 18 Jan. 2021).

[20] SLS-DSU. (2021) *How are synthetic data created?* Retrieved from https://sls.lscs.ac.uk/guides-resources/synthetic-data/how-is-synthetic-data-created/ on 18 Jan. 2021.

[21] NILS-RSU. (2021) *NILS Univariate Synthetic Data.* Retrieved from https://www.nils-rsu.co.uk/nils-univariate-synthetic-data/ on 15 Feb. 2021

Uses for synthetic data

Synthetic data is typically envisaged as being useful in three broad applications which we summarise below, then go on to discuss in more detail:

- **Training:** Providing training data that allows researchers to familiarise themselves with administrative data sets and what can be done with them, without the need for extensive data access requirements.
- **Testing:** Providing realistic data that allow researchers to test analysis code or (for high-fidelity synthetic data) conduct exploratory research prior to making final runs on the real data in a secure environment.
    - Providing early access to a test data set allows research to proceed more quickly, as code can be tested prior to being granted access to the real data, and the time spent in secure access environments is reduced.
- **Sharing:** Allowing easier sharing of data within or between government departments, making it clearer what data sets are available and making it easier for analysts to test and refine analysis code prior to or instead of having access to the real data, much as discussed for external researchers.

These categories overlap to some extent: sharing often implies testing, and testing if performed on an unfamiliar data set might take on aspects of training. As discussed in the Introduction, the chief benefit of these uses is that they allow parts of the research process to be performed earlier than they otherwise would be if only the real data requiring secure

[22] synthpop. (2019) *synthpop - Welcome.* Retrieved from https://synthpop.org.uk/index.html on 18 Jan. 2021.

[23] Nowok, B., Raab G. M., and Dibben, C., (2016) 'synthpop: Bespoke Creation of Synthetic Data in R' *Journal of Statistical Software* 74(11). Available at: http://dx.doi.org/10.18637/jss.v074.i11 (Accessed 18 Jan. 2021).

[24] SLS-DSU. (2021) *How reliable are results from synthetic data?* Retrieved from https://sls.lscs.ac.uk/guides-resources/synthetic-data/how-reliable-are-results-from-synthetic-data/ on 18 Jan. 2021.

[25] Elliot, M., (2014). 'Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team'. Available at: http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf (Accessed 15 Feb. 2021).

[26] It is possible that a disclosure risk still exists for small enough demographic groups and in cases where the underlying data is sufficiently closely replicated. See Bates, A. G., Špakulová, I., Dove, I., and Mealor, A., (2019) 'Synthetic data pilot' ONS methodology working papers 16. Available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot#synthetic-data set-spectrum (Accessed: 27 Jan. 2021).

[27] Online centralised differential privacy involves enacting differential privacy processes as a query is made on a real data set rather than generating a differentially private synthetic data set from real data. See Dwork, C. and Roth, A (2014) 'The Algorithmic Foundations of Differential Privacy' Foundations and Trends® in Theoretical Computer Science 9(3-4), 211-407. Available at: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf (Accessed: 26 May 2021).

[28] See for instance Gaboardi, M., Honaker, J., King, G., Nissim, K., Ullman, J., Vadhan, S., and Murtagh, J., (2016) 'PSI (Ψ): a Private data Sharing Interface'. Available at: https://privacytools.seas.harvard.edu/publications/psipaper (Accessed: 15 Feb 2021).

access privileges is available. It also potentially *scales access*, in that its very existence and availability may encourage many more people to make use of it.

Behind these potential uses lies an assumption: because it is not the original data (and thus does not contain information pertaining to real people), the access requirements for synthetic data should be less stringent than those for accessing real data. How strongly those we spoke to agreed with this underlying assumption varied. It was generally held to be true for the lower fidelity forms of synthetic data, but a number were much more cautious with respect to high-fidelity forms of synthetic data; this will be discussed in detail in a later section.

Training

Researchers working with administrative data need to understand both how to analyse different data sets and what is in them. This is difficult if access to data sets of interest is conditional on passing through a slow and/or intensive secure access application process, particularly if the aim is to train a large number of researchers efficiently[29].

Training on specific data sets can be more useful than training on generic ones

Familiarity with the formats and contents of specific data sets held in a given repository provides an advantage over the use of 'generic' training sets or open data held elsewhere. If I am a researcher who intends to research demographic data in the UK, it is much more efficient to learn how to analyse data from the UK Longitudinal Studies than it is to learn how to analyse administrative data sets taken from an open data repository run by New York City and *then* have to familiarise myself with any unique features of the Longitudinal Studies data. If I am a newly hired government analyst who must familiarise myself with how best to perform analyses on data sets relevant to (for example) COVID-19, it is probably best that I have data that resembles *these specific* data sets rather than any others. Synthetic data can provide those data sets.

Synthetic data for training does not have to be high-fidelity

Since this data is being used for teaching rather than research, it is not necessary for it to have full fidelity to a real data source. Those we spoke to observe that it is possible to learn a good deal from a correctly structured and formatted data set (including comprehensive descriptions of its contents) where data in individual columns follows the correct distributions, even if none of the correlations in the real data set are reproduced, although one noted that having some correlations still present can be useful.

Synthetic data for training can have less stringent access requirements

An example of a high-fidelity training synthetic data set is one that has been generated from an extract of Scottish Longitudinal Study data using `synthpop` that has been used in five day training courses by the Scottish Centre for Administrative Data Research[30]. This data is

[29] The new Cabinet Office/Treasury Evaluation Task Force is a case in point, as are wider civil service reforms on empirical methods.

[30] synthpop. (2019) *Examples of uses - synthpop* Retrieved from https://www.synthpop.org.uk/examples-of-uses.html on 22 Jan. 2021.

typically held on secure cloud storage, accessible by attendees, for the duration of the course.

If low-fidelity synthetic data is sufficient for most training needs, a less onerous access process for these training sets would be justified, particularly as data generated for this purpose should not include personal identifying information in the first place.

Testing

Data access requirements cause delays

Researchers often need to develop and test code for the analysis of data sets. This can entail a considerable delay before a research project can be started following the initial application for access, and analyses can only be performed on some data sets in an entirely secure environment (using an encrypted VPN and/or at a physically secure location).

Requiring access to data in order to know why you need access to the data

Furthermore, access to secure research settings typically requires that the researcher provide a specific research question as a condition of access to the data; however, exploratory research on the data is often necessary to clearly define such questions and identify whether the data is capable of answering it in the first place[31].

Synthetic data can help by providing properly structured data on which researchers can test their code as in the case of training, and (for high-fidelity synthetic data) uncover relationships between variables that are of interest. More controversially, it could be a complete substitute for real data, although this would require extreme confidence that the methods used for its generation minimised disclosure risks. (Even synthpop, which is noted by one of those we spoke to as preserving previously undiscovered relationships that are present in the real data, is not fully trusted at present in practise, as can be seen from our case study.)

The value of good metadata and of low-fidelity synthetic data

Our discussions indicate that questions of 'which data are present' can be answered by the publication of comprehensive descriptions of the data set (metadata), including e.g., the types and distributions of values in various fields[32]. This would function as detailed documentation so that researchers could be aware of this information prior to making an application. Others observed that metadata plus a structural synthetic data set, the lowest-fidelity version from Table 1, is probably much more useful than metadata alone. This is because analysts can get a better feel for the data and what it can be used for by browsing and manipulating it in their analysis package of choice. Metadata is not always easy to browse, especially if there are a large number of columns. Further, a synthetic data set

[31] Gibbons, D., *et al*. (2021) 'Applying Behavioural Insights to Cross-government Data Sharing' BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021) pp 39.

[32] The metadata available on the UK Data Service's Nesstar Catalogue provides a good example of what is meant by 'comprehensive' in this context. See http://nesstar.ukdataservice.ac.uk/webview/, accessed on 26 Feb. 2021.

allows an analyst to start writing code immediately; the synthetic data set can help them check and debug their analysis before needing access to the real data.

As with the data used for training, practitioners indicated that relatively low-fidelity synthetic data is still of use for testing, despite not reproducing all features of the data set.

Lack of uptake

It was noted during our discussions that while proposals for this use of high- or low-fidelity synthetic data can generate positive responses from researchers, there have been cases where the uptake of a synthetic data service is lower than expected. In the case of synthetic data based on Scottish Longitudinal Survey data, it was suggested that this might be because Scottish researchers found secure research facilities relatively easy to access and so did not feel the need to make use of synthetic data. In a second case relating to low-fidelity data, it was suggested that low quality documentation and poorly named data columns may have been a factor in why few of the requests for synthetic data resulted in analysis code being shared for use on the real data. If this is the case then it highlights the importance of including good metadata and documentation as part of a roll-out of synthetic data. However, it must be stressed that we are not in a position to make a fair comparison with the uptake of real data, and it is possible that many of those who requested this synthetic data might have had other reasons for not going ahead with their analyses.

Sharing

The value of synthetic data for sharing

Some analysts that we spoke to expressed doubt that synthetic data would be of much interest for sharing between departments since the ONS Secure Research Service should provide a reasonable means of doing so. However, synthetic data could be accessible between departments far more quickly than the original data, due to more relaxed access requirements. There may also be cases where the data cannot leave the data holder, in which case synthetic data can support the external development of analysis code that can then be run safely by the data holder.

Reluctance to sign off on real data

Reluctance by individual data owners (who can lack technical expertise and can have other responsibilities) to sign off on the sharing of real data was flagged as an issue by one person we spoke to (and is consistent with the findings of our earlier report[33]). Some forms of synthetic data might reduce that barrier by reducing the risk, though as we discuss subsequently this may not be as straightforward as it first appears.

Inter- versus intra-departmental sharing

One department noted an 'extreme demand' amongst colleagues for synthetic versions of data held by their department that could be accessed in order to develop and test their analysis code outside of secure settings. This suggests that synthetic data is certainly envisaged as a means of sharing data within individual departments by some practitioners.

[33] Gibbons, D. *et al*. (2021) ‘Applying Behavioural Insights to Cross-government Data Sharing’ BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021) pp 14-19.

The desire to make use of synthetic data for testing ‘outside of secure settings’ could also generalise to inter-departmental data sharing, since it is not intended as a replacement for the real data, but as a means of ensuring the robustness of analyses that would later be carried out on the true data.

An appetite for lower risk *intra*-departmental use of synthetic data, if served, could build confidence and demand for using the approach between departments.

Doubts and concerns regarding synthetic data focus on data quality and privacy

Here we discuss concerns regarding synthetic data itself. If synthetic data is to fulfil its promise, then potential risks and other barriers to its use must be addressed. We firstly discuss concerns around the quality of synthetic data and its implications for potential breaches of privacy or institutional embarrassment, as well as the difficulty of distinguishing it from real data. We then discuss specific concerns raised by the notion of differentially private synthetic data, that is, synthetic data generated in such a way as to be consistent with a rigorous mathematical definition of privacy.

Data quality concerns

The quality of synthetic data — how well it reproduces the relationships and characteristics of the real data — was of concern to those we spoke to in several ways. We summarise some points of interest below:

What if the data are ‘too good’?

While low-fidelity synthetic data were generally viewed positively, there were concerns over synthetic data that more accurately reproduces the underlying data being more widely available, since that potentially allows individuals with unusual characteristics to be distinguished (or have their data accidentally reproduced); this is still disclosive, since that data could be argued to be equivalent to a deidentified record[34].

- The ONS Synthetic Data Spectrum reflects this concern by classifying synthetic data according to its quality, which is proportional to the disclosure risk of the synthetic data[35].
- Some of those we spoke to were sufficiently confident in the low disclosure risk of a given synthetic data generation method, but thought that decision makers would still be over-cautious.

Risk of being mistaken for real data

Related to the above, synthetic data will often reproduce the format of the original data and the presence of errors in the data set - in fact, that it does so is often a reason for generating it in the first place. There is therefore a risk that synthetic data if released unlabelled into the

[34] Beatty, R. (2020) ‘Synthetic Data’, Report for NISRA, paras 35-37 and 46-48.

[35] Bates, A. G., Špakulová, I., Dove, I., and Mealor, A. (2019) 'Synthetic data pilot' ONS methodology working papers 16. Available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot#synthetic-data set-spectrum (Accessed: 27 Jan. 2021).

public domain might be mistaken for genuine data, with consequent risk of embarrassment. Those we spoke to noted that as a consequence of this, some civil servants may be more reluctant to sign off on the release of more realistic synthetic data.

A partial solution could be to use digital watermarking techniques[36]. These encode a 'watermark' signal identifying the data source into the data itself while leaving the overall statistical properties unchanged. However, the data is still not obviously synthetic to anyone accessing the data set, as specific tools would be needed to show that a watermark is present. Clearly labelling the data set as being synthetic would make its nature obvious, but these labels could still potentially be removed inadvertently or by malicious actors. We are not currently aware of any tamper resistant methods that can identify data as synthetic in an immediately obvious fashion.

Differences from real data

On the other hand, synthetic data may also be different from real data in crucial ways.

- An additional risk to mistaking synthetic data for real data is that an analysis on synthetic data may give very different results to an analysis on the source data set, leading to erroneous conclusions. This can mean a certain degree of conservatism with respect to the publication of analyses based on synthetic data is warranted (as seen with respect to the SLS discussed above), although one respondent's view was that they would be happy with published analyses being based on synthetic data, provided that it is sufficiently high-fidelity.
- It can be argued that there are risks with conclusions being drawn from synthetic data generated with e.g., demographic data, particularly if it has policy implications; in these cases, real data should be used as analysts can be certain that it is correct.

Misrecorded information and undesirable values

Privacy is not always the underlying concern: one of those we spoke to noted that even publicly available data sets can contain misrecorded information that could be potentially undesirable for the department. The prevalence of missing data may disclose unwanted information about the quality of the data set maintained by the owner.

Another possibility is that the presence or prevalence of certain undesirable categorical values in a data set (e.g., a label that may indicate a mistake made by a civil servant), if they are reproduced, may reflect badly on the department.

In low-fidelity synthetic data, where consistency checks between data fields (e.g., marital status and age) are not always enforced, there is also potential for confusion or accusations of poor data quality if synthetic data is mistaken for real data, since undesirable combinations of fields (for example, a married four-year old) might be generated. These types of errors would not be introduced in high-fidelity data, but as greater accuracy also increases (or is often perceived as increasing) the risk of privacy leakage, there is a tension between these considerations.

[36] Panah, A. S., van Schyndel, R., Sellis, T., and Bertino, E., (2016) 'On the Properties of Non-Media Digital Watermarking: A Review of State of the Art Technique' IEEE Access 4: 2670-2704. Available from: https://ieeexplore.ieee.org/document/7473843. (Accessed 27 May 2021.)

A suggested solution was to maintain some level of control, to ensure a level of understanding, when accessing synthetic data, although one that is relatively light touch when compared to that for accessing real data.

To ensure the utility of the data for writing analysis code, it is important that, to the furthest extent possible, errors in the original data set are similarly reproduced in the synthetic data so the analyst can devise a strategy for handling them. It is important therefore to communicate to analysts the type of error that is easily introduced by the synthesis process itself (e.g., in low-fidelity synthetic data, unusual combinations of traits seen in one individual) so they can be differentiated from genuine data quality issues to be addressed.

Differential privacy and related techniques

Differential privacy[37] is an approach to privacy protection that can be combined with synthetic data generation to ensure that it conforms to a rigorous mathematical definition of privacy limiting the ability of interested parties to infer whether a given individual is a member of the data set or not.

How does differential privacy work?

While there are some variations depending on the precise method used, typically when differential privacy is used to create synthetic data, some additional randomness (calibrated noise) is injected into the data. This is done in such a way that aggregate statistical properties are largely unaffected while individual records are perturbed, so that the probability of reproducing an individual's data or inferring their membership of the data set is reduced.

In the original differential privacy scheme[38] the degree of privacy granted is controlled by a parameter ε (epsilon). There is a trade-off between privacy and the utility of the data set: as ε is made smaller it will be less likely that an individual record can be inferred to be a member of the data set, but the noise introduced will increase and may render the synthetic data useless for analysis.

Views of differential privacy

Differentially private synthetic data approaches are being used to prepare data from the 2020 US Census for public release[39]. In spite of this large-scale government led implementation, many of those we spoke to were sceptical about the usefulness of differentially private synthetic data; we summarise some of their key observations here.

[37] Page, H., Cabot, C., and Nissim, K., (2018) 'Differential privacy: an introduction for statistical agencies', Government Statistical Service.

[38] Dwork, C., McSherry, F., Nissim, K., and Smith, A., (2017) 'Calibrating Noise to Sensitivity in Private Data Analysis' *Privacy and Confidentiality* 7 (3):17-5. Available from: https://doi.org/10.29012/jpc.v7i3.405 (Accessed 10 Feb. 2021).

[39] Census Bureau. (2021) *Disclosure Avoidance and the 2020 Census - Census Bureau* Retrieved from https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html on 27 Jan. 2021.

What is ‘privacy’?

Differentially private synthetic data generation is dependent on the choice of parameters that select a trade-off between the privacy of an individual’s data and the utility of the data set. These parameters must be chosen as a policy decision by an institution.

- One of those we talked to was sceptical that the strictly mathematical definition of privacy that forms the basis of differential privacy sufficiently captures all the ways in which a synthetic data set could pose a legal or ethical risk of disclosing information.
- They also observed that the meaning of the privacy parameters are opaque to most non-experts - for example, simply stating that a data set has a particular ε value is unlikely to convince many of its privacy preserving properties, whether non-specialist data holders or the general public.
- We also note that differential privacy assumes a worst-possible case scenario where the individual attempting to gain knowledge of a target’s information having access to resources that in reality would be very difficult to access. For practical purposes, a given synthetic data set can be much more secure than is suggested by its differential privacy classification[40], which further confuses the issue. It is also very different from other approaches towards assessing disclosure risk.

Differential privacy and data quality

One of those we spoke to noted that in their experience generating good quality synthetic data for microdata using differential privacy is difficult; the introduction of noise tends to overwhelm any useful information present.

- It is worth observing that even in large-scale attempts to create synthetic data with differential privacy such as the US Census, concerns remain about the ways in which the introduction of noise and consistency conditions may bias the synthetic data in ways that might result in some erroneous transfer of resources from diverse urban populations to segregated rural populations[41].
- A related concern is that the introduction of noise into data to reduce disclosure risk for high-fidelity synthetic data would render modelling performed on that data invalid[42]. While granting that there are reasonable concerns here, in general this may be overcautious as the noise injection in these techniques is controlled so as to reproduce the aggregate properties of the original data. Techniques used to compensate for measurement error might be used to perform statistical inference on differentially private synthetic data in order to overcome these problems[43].

The somewhat lukewarm response of some researchers suggests that the use of advanced privacy protecting techniques needs further consideration. We discuss our recommendations in the section on Advanced privacy protection techniques.

[40] McClure, D., and Reiter, J.P., (2012). ‘Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data’ *Transactions on Data Privacy* 5 535–552.

[41] Petti, S. and Flaxman, A. D., (2020). ‘Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff’, *Gates Open Research* 3 1722.

[42] Beatty, R. (2020) ‘Synthetic Data’, Report for NISRA , para 46.

[43] Evans, G. and King, G. (2021). ‘Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs data set’, Working Paper. Available at: https://gking.harvard.edu/dpd (Accessed 27th Jan. 2021).

Systemic or technical barriers to the use of synthetic data

Those we spoke to indicated that there were a number of systemic and technical barriers facing the widespread introduction of synthetic data. If the use of synthetic data is to be widespread within government, it will be necessary to address these factors that are not inherent to the notion of synthetic data itself, but reflect the wider legal and institutional environment.

- Lack of knowledge regarding synthetic data.
- Ethical and legal barriers.
- Inconsistent technological support.

Lack of knowledge regarding synthetic data in government and amongst the general public

The need for technical knowledge amongst officials

As was also found in our earlier report on data linkage[44], practitioners expressed concerns about how well non-practitioners understand the issues involved. Even when synthetic data could be seen to have a reasonably strong guarantee of being privacy protected (training data lacking identifying information and generated using `synthpop`, hence carrying a strong privacy guarantee) one of those we spoke to, identified a potential excess of caution with regards to accessing it.

Accurate judgements regarding the safety of synthetic data require technical and specialised knowledge, and it did not seem clear to some of those we spoke to that this knowledge was widespread. They suggested that in general, lack of knowledge amongst data providers can lead to two negative outcomes: excessive conservatism, or the belief that synthetic data is entirely risk-free. We discuss recommendations with respect to concerns within government in the section Risk aversion and lack of knowledge.

The public does not know what synthetic data is

Similarly, lack of knowledge about synthetic data among the general public and privacy campaigners, particularly for sensitive information (e.g., concerning children) was noted as a potential risk, especially given the concerns noted below regarding the difficulties of distinguishing an unlabelled synthetic data set from real data.

On the other hand, for other kinds of data (e.g., concerning rare diseases), it was also observed that public attitudes towards the use of real data for beneficial research may be more lenient than is often allowed for by data providers. This suggests that public opinion of the use of high-fidelity synthetic data in those cases might be quite positive.

One of those we spoke to stated that despite having a desire to make widespread use of high-fidelity synthetic data in their department and a willingness to defend doing so, a prerequisite of doing this, given the risks, would be a public information campaign. We note

[44] Gibbons, D., *et al*. (2021). ‘Applying Behavioural Insights to Cross-government Data Sharing’ BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021) pp. 17-19.

that this could be preceded by (e.g.,) experimental trials that test public preferences for different approaches to the production of synthetic data.

We discuss recommendations with respect to informing the general public about synthetic data in the section The general public.

Ethical and legal barriers

Legal and ethical considerations regarding how data may be used[45] have been flagged by some of those we spoke to as barriers to the use of medium to high-fidelity synthetic data with certain kinds of sensitive information (e.g., mental health data). Our previous report[46] flagged a number of problems and issues relating to similar concerns with data linkage, for example:

- Broad definitions of 'personal data' that have not been narrowed by legal precedent
- Changes in data protection law (e.g., introduction of the GDPR).
- A lack of knowledge among policy staff as to what is and isn't legal (sometimes leading to legally permitted data uses being misconstrued as illegal).
- Uneven development of ethics and consent procedures across government (both overdevelopment and underdevelopment may hinder data linkage and sharing).

It is unlikely that there can be a technological solution to these issues; while synthetic data can reduce the risk of privacy breaches, for all but the lowest quality category it does not completely eliminate them. There is a need for clear guidance on specific problems (e.g., what if data relating to an actual individual is accidentally reproduced in a synthetic data set) and what data disclosure procedures might be needed to reduce any risks that might arise from this. We discuss recommendations relating to these barriers in the section on Risk aversion and lack of knowledge.

Inconsistent technological support

The gap between aspiration and available technology

An interviewee noted that a barrier to the use of synthetic data within government is the extreme variability of information technology (IT) support available from department to department. This can lead to some analysts being restricted to very basic computational tool sets while others have access to the latest software and approaches (cloud-based computing, access to open source tools, use of containerisation[47] to account for future changes in favoured tools) while maintaining the required level of security. They also noted that many departments wish to do their own data science as opposed to contracting it out to third parties.

[45] Such as consent and reasonable expectations regarding data usage, as well as possible constraints relating to the GDPR.

[46] Gibbons, D. *et al*. (2021). 'Applying Behavioural Insights to Cross-government Data Sharing' BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021) pp 15-17.

[47] A method ensuring that everything required for a computer program to work - including its operating system - is included in one package so that the program can be shared between different computers without having to install additional software or change other aspects of the computing environment.

If this is so, then in some cases there would be a gap between a department's desire to make use of data science and its capacity to do so. It is consistent with our earlier report's conclusions[48] that uneven access to the required technological resources is a barrier to data sharing and linkage. It is likely that factors identified therein (e.g., dependency on legacy or proprietary formats and software) might also arise in this context.

All these factors would also impact negatively on their ability to generate and make use of synthetic data, particularly in cases where analysts are restricted to using out of date tools, approaches to computing, and programming languages.

Other resources may also be a constraint

In addition, some departments may not have access to the resources needed for the large-scale generation of synthetic data. Even restricting generation of high-fidelity synthetic data to bespoke requests may stretch the resources of smaller departments and also larger departments who have a heavy load of urgent tasks to deliver.

We discuss recommendations relating to these problems in the section Technological considerations.

Views on approaches to generating synthetic data at scale

During some meetings we discussed technical approaches that might make the generation of synthetic data easier, particularly at scale. Two aspects of these discussions are summarised in this section, suggesting possible directions and prerequisites for the technologies involved.

Semi-automated generation of synthetic data

Some of those we spoke to were asked how they would view a tool or pipeline for the semi-automated generation of synthetic data. It was generally viewed favourably: it was observed that automation would be most appropriate for low-fidelity synthetic data, which is less risky from a privacy perspective and yet still of use to researchers and analysts in terms of data exploration and code development. It might also help build familiarity with synthetic data amongst researchers, data holders and the general public.

Ideal characteristics of a synthetic data generation tool

After a number of these discussions, we summarise requirements that any synthetic data generation tool should meet as follows:

1. It should be capable of being run on a broad range of departmental IT infrastructure.
2. It should have undergone a solid development process, and undergone a strong, independent Quality Assurance process.
3. It should give clear guidance on what privacy protections it provides and does not provide. For example, these could include indications of how the level of synthetic data fidelity a tool provides corresponds to a given level of disclosure control or the

[48] Gibbons, D. *et al*., (2021). 'Applying Behavioural Insights to Cross-government Data Sharing' BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021) p 13.

use of automated tests to ensure that individual records from the original data are not included in the output synthetic data set.

Given the concerns identified regarding the consistency of technological support between departments and the degree to which these protections are understood by non-experts, the first and third points seem especially salient.

3. Synthetic data prototype code

In response to the following considerations:

- We identified support for a low-fidelity synthetic data approach to encourage cross-department data sharing.
- Existing tutorials for synthetic data generation focus on high-fidelity synthetic data.
- Writing code that is *functionally incapable* of generating higher fidelity synthetic data would be safer to use for this purpose.
    - Using existing libraries, small tweaks to parameters can result in generation of high-fidelity synthetic data.
- Low-fidelity synthetic data generation is much more amenable to semi-automation.
- Low-fidelity synthetic data generation requires much simpler code than the fully. functional synthetic data generation libraries that currently exist, and so would be easier for a department or researcher to vet or QA themselves before use, if desired.

We developed a simple Python script that could be used by researchers to generate low-fidelity synthetic data sets, corresponding roughly to 'Synthetic-Augmented Plausible' in the ONS classification. This generates univariate synthetic data (that is, it does not preserve correlations between different variables present in the real data), and has been tested on three publicly available administrative data sets:

- NHS Accident and Emergency data used in the Open Data Institute synthetic data tutorial[49].
- Data taken from the City of New York Citywide Payroll data set[50].
- Data taken from the City of New York Open Parking and Camera Violations data set[51].

The code is provided here in a Python notebook format[52] in order to make the process of synthetic data generation clear and easy to follow. It also ensures that relevant information on the nature of synthetic data is provided in a readable format.

Possible extensions

- At present the program will generate univariate synthetic data for a given data set. However, that does not guarantee that the resulting data set will not disclose sensitive information. The addition of automated tests of the synthetic data set (e.g., of whether records from the original data set are reproduced or whether correlations between different data fields still exist) might help build confidence in the safety of its output.

[49] theodi. (2019). Retrieved from https://github.com/theodi/synthetic-data-tutorial/blob/master/data/hospital_ae_data.csv on 23 Feb. 2021.
[50] NYC Open Data. (2019). *Citywide Payroll Data (Fiscal Year).* Retrieved from https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e on 23 Feb. 2021.
[51] NYC Open Data. (2021). *Open Parking and Camera Violations.* Retrieved from https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89 on 23 Feb. 2021.
[52] https://bit.ly/synthetic-generator-colab

- The above could be combined with some degree of interactivity so that the user is able to confirm assumptions about e.g., the variable type that the program is making.
- A full set of tests checking that all subroutines work both individually (unit tests) and together (integration tests) as advertised will also help confirm that the code is working as intended.
- More sophisticated separation of the step obtaining the distributions of the real data from the step used to generate the synthetic data, so that it can be made clear that any correlations between data fields are destroyed.

4. Recommendations for further use

In general, we have found from our discussions that there is scope for the expansion of the use of synthetic data, but that the levels of both knowledge and enthusiasm are inconsistent. We have discussed one major attempt to make use of synthetic data for both research and training purposes (the Scottish Longitudinal Study) and noted that the research component has been inconsistent in uptake, and that in the eyes of many practitioners, most uses of synthetic data will reduce but not eliminate the need for disclosure procedures or some level of security surrounding data.

Those we spoke to have also noted that lack of knowledge among policy makers, data holders and practitioners, as well as technical constraints and potential legal and ethical questions, are all factors that might frustrate the widespread use of synthetic data.

Since, as outlined in the Introduction, an increase in the use of synthetic data might prove to be of great benefit to research and analysis within and outside government, it is worth discussing how some of these problems and barriers might be addressed.

We will first make our principal recommendation with respect to synthetic data, before discussing more domain specific recommendations that will also support its uptake.

Principal recommendation

ADR UK should encourage the use and sharing of low-fidelity synthetic data across government and with researchers for the purpose of data discovery and testing.

Low-fidelity synthetic has a lower risk of disclosing information

Low-fidelity synthetic data poses minimal disclosure risks compared to higher fidelity forms, since it does not provide realistic relationships between variables. However, it is still of great use to researchers since it provides realistic distributions of values for individual variables and the correct data formats for each data file.

This is valuable information for researchers and analysts, both in terms of data discovery (learning what data is held by each department or group), and for reducing the amount of time spent idle while waiting for access to the real data to be granted.

Analysts and researchers can be given access to low-fidelity synthetic data without stringent access requirements and could then develop and test their analysis code while waiting for access to the real, more secure data. Analysts could also share the code directly back with the data owner, for them to run securely, returning only the aggregate results. This would be similar to the bespoke high-quality synthetic data service currently provided by the Scottish Longitudinal Study.

As we have described, this means that less time is needed with access to sensitive data and/or in secure settings, and that delays to data release and information governance (which are common) need not entirely stall the writing of data cleaning and analysis code, which is often a time consuming part of any analysis project.

Low-fidelity synthetic data can be used in training

Low-fidelity synthetic data is sufficient for most training needs, justifying a less onerous access process for such synthetic data sets. To keep disclosure risk low, but improve the training experience, **a sensible route may also be to reflect the original data in a low-fidelity, univariate way, but to add in some randomly generated correlations between variables so that researchers have patterns to observe and model during their training.**

Large scale production will produce familiarity with synthetic data

Large scale production of low-fidelity synthetic data may also help familiarise both officials and practitioners with the concept. Normalising the production of low-fidelity synthetic data versions and its release under less stringent conditions than real data could give a better appreciation of the techniques and risks involved, setting the stage for discussions of whether or not to release higher-fidelity synthetic versions of data sets in due course. **The development of tools such as the one prototyped as part of this study and of automated pipelines for synthetic data production will play a vital role in enabling this production at scale.**

We note that regardless of whether or not higher-fidelity synthetic data is eventually deemed suitable for widespread use, production of low-fidelity data at scale for use by researchers and analysts has clear advantages over the present state of affairs**.**

Further work will be needed to produce high-fidelity synthetic data at scale

Further work will be required to enable the large-scale production of high-fidelity synthetic data. This is partly since we feel that the greater disclosure risk will render many data holders reluctant to grant access to this data, particularly as many may be risk averse and/or feel they lack sufficient knowledge of synthetic data to make a sensible decision regarding this risk.

It is also the case that high-fidelity synthetic data requires more effort to produce than low-fidelity data: given the large amounts of administrative data for which synthetic versions could be of use the production of many low-fidelity synthetic data sets is a better return on investment than the production of fewer high-fidelity synthetic data sets. This is especially true if, as many have suggested, the final analysis would still need to be run against the real data for verification, even if the synthetic data was very high fidelity.

Bespoke production of high-fidelity synthetic data sets still has value

However, 'bespoke' production of high-fidelity synthetic data sets on request (as currently implemented for the Scottish Longitudinal Survey) should be encouraged where resources permit. Research into the production of high-fidelity synthetic data sets and building use cases for them should be supported[53].

Similarly, the availability of high-fidelity synthetic data for certain data sets, although on a much smaller scale, will also contribute to this process of familiarisation. It will also help build

[53] e.g. Katie Harron at UCL has an upcoming project to provide robust synthetic data from a new linkage between the National Survey of Sexual Attitudes and Lifestyles (Natsal 3) and two administrative data sources: Hospital Episode Statistics (HES) and the National Pupil Database (NPD).

and maintain a base of expertise in the production of high-fidelity synthetic data that could serve as the basis for future work on scaling up its production or the assessment of new and developing technologies and techniques in the field, such as differential privacy.

Recommendations within government

We have identified three broad problems that need addressing in order to produce synthetic data at scale for purposes of training, testing and sharing. These are

- Technological considerations
- Risk aversion and lack of knowledge
- Use of advanced privacy preserving technologies

Technological considerations

Overcoming technological fragmentation

Our engagement has identified that a barrier to the sharing and generation of synthetic data between departments is technological fragmentation: different departments will often have widely different philosophies regarding technology and information security. This can either promote the kind of approaches needed to practise data science in a secure environment or inhibit them by prohibiting access to state of the art tools. A contributing factor to this may be lack of resources.

Ways of overcoming this barrier could include:

- **Pooling of technological resources** might mitigate resource constraints to some extent.
- **Consistent minimum technological standards** would allow the generation of synthetic data ‘in house’ with the appropriate standardised tools, which might then be shared between departments with fewer hurdles, or might make the sharing of data possible in the first place where legal restrictions would normally prevent it.)

Centralised synthetic data storage

The sharing of synthetic data, clearly marked as such, between government departments would appear to avoid many of the worries associated with the release of data to the public, such as being mistaken for real data. While even high-fidelity synthetic data could arguably pose less of a risk in such a context, due care should be taken to ensure that this is in fact the case. **A centralised store of synthetic data should be considered that is easier for analysts to access relative to the true data.**

Tools for generating synthetic data

Related problems are the sheer number of data sets from which it might be desirable to generate synthetic data, and the tools used to do so. In light of this, we suggest:

- **The creation of a standard tool that meets the three requirements described previously** (portability between departmental machines; solid development and QA; clarity on privacy protection) that is maintained by a single entity in government in order to ensure a consistent standard, with each version released subject to QA by external experts. Our prototype tool is a step in this direction.

- This tool should reflect a common philosophy underpinning the technology available to departments in order that it is widely usable.
- Ideally it should be able to generate a class or classes of synthetic data that correspond to particular ONS categorisations in order to make clear how disclosure risk should be handled.

Automated pipelines for synthetic data generation

Given the level of effort involved in generating bespoke high-fidelity synthetic data sets, it is unlikely that this will be performed outside of a small number of data sets of high importance at present. However, low-fidelity synthetic data is relatively easy to generate, and software to generate it is relatively easy to write, as the code and tutorial accompanying this report demonstrate.

We recommend **investigating the feasibility of an automated pipeline for synthetic data generation:**

- Ideally, upon being deposited in secure storage a data set is passed into an automated pipeline which generates a corresponding set of synthetic data and metadata, which is passed into provisional storage.
- Upon receiving a request for synthetic data corresponding to that data, the synthetic data set is removed from provisional storage, checked for errors and for disclosure risk, and (following any necessary modifications for these stages) passed into a general secure store for synthetic data, where it can be subsequently requested as needed.
- This ready availability of synthetic data would not only aid in testing analysis software (both for government analysts and external researchers) but would help contribute to a process of familiarisation with synthetic data and its potential uses, potentially mitigating the problem to which we turn in the next section.

Risk aversion and lack of knowledge

Lack of knowledge regarding synthetic data as well as potential legal and ethical quandaries surrounding it have been identified as potential barriers to increasing its use.

Explain what synthetic data is and its risks and benefits

In general, **increasing awareness of synthetic data and its potential (through tutorials, training and ubiquity)** might alleviate this. If officials and policy makers are fully aware of the benefits such as scaling access and increasing the efficiency of research and policy formation, they may be more likely to allow the production of synthetic data for given data sets. Being aware of potential risks will allow them to assess what kind of synthetic data is appropriate for a particular data set. Furthermore, we would recommend:

- Using consistent terminology. The distinction between low-fidelity and high-fidelity synthetic data we have outlined should help clarify conversations early, without unnecessarily specific breakdowns that are hard for non-experts to keep track of. Moving on from historical uses of ‘dummy’ and ‘fake’ data to the more general low-fidelity synthetic data reminds users that these options lie on a spectrum.
- **Clear and consistent advice on what tools or approaches are recommended to generate synthetic data.** This should involve the Information Commissioner's Office

at an early stage in order to ensure that this advice is compliant with the law concerning personal identifying information.

- **Requesting a formal opinion from the ICO as to what, if any, synthetic data might constitute personal identifying information and hence be subject to the GDPR.** This could be incorporated into our earlier report's suggestions regarding consistency between departmental legal gateways regarding data sharing in general[54].

We also recommend **the use of concrete examples** when discussing the disclosure risks of different levels of synthetic data. For example, consider the following data entry:

UID	County	Age	Income	Planet of Origin
00000001	Essex	34	£100,000	Mars

Depending on the quality of synthetic data generated from a UK-wide data set that includes this record, it may be possible to draw any of the following conclusions:

1. It is possible to earn £100,000 in the UK.
2. Someone in the UK has earned at least £100,000.
3. Someone in the UK has earned exactly £100,000.
4. Someone in Essex has earned at least or exactly £100,000.
5. Someone from Mars living in Essex has earned £100,000, and if we know that there is precisely one Martian living in Essex, we will know exactly what they have earned, even if the data is synthetic. This is because high-fidelity synthetic data may preserve this kind of correlation for individual records with rare characteristics.

Concrete examples such as the above can convey the disclosure risks associated with different classes of synthetic data more clearly than more abstract, technical discussions such as those in Table 1, particularly to non-experts.

Automatic synthetic data generation and non-practitioners

Another barrier identified by one of those we spoke to is that in some departments, responsibility for a data set may be held by an individual whose primary responsibilities are non-technical and who may be averse to the potential risk of releasing data due to the consequences of doing so inappropriately. This might also apply to the generation of synthetic data from that data.

- If the automatic generation of synthetic data on deposition is normalised, this might no longer be an issue.
- However, this could also be mitigated in the case of low-fidelity synthetic data by **emphasising the lower risks of releasing such synthetic data.**

[54] Gibbons, D. *et al*. (2021) 'Applying Behavioural Insights to Cross-government Data Sharing' BIT report. Available at: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf (Retrieved 24 Feb. 2021) pp 38-39.

Advanced privacy protection techniques

Given the scepticism expressed by practitioners towards the value of approaches such as differential privacy at this point in time, **further work is needed before they are considered for adoption.**

However, in the longer term it is perhaps worth keeping track of developments in this field, given the potential it might have for allowing the safe release of high-fidelity synthetic data to researchers or even the general public.

- **Attention should be paid to the success or otherwise of the use of differential privacy in the 2020 US Census and in other administrative data applications elsewhere in the world.** Particularly important issues that should be considered include: the guarantees of protection given by a given differential private algorithm - does it work as advertised? - and whether a given approach introduces biases into the data that might influence the direction of policy, as discussed in an earlier section of this report.
- **Any tools built should perhaps be implemented with an eye towards incorporating differential privacy or related techniques at a later date.**

Promoting knowledge of differential privacy

A particular barrier identified is the lack of transparency due to differential privacy's technical nature, and the problems involved in conveying the meaning of the privacy parameters used by the technique, which may lead to over- or under-confidence with respect to the privacy or utility provided by a given differentially private synthetic data set.

This is particularly important to keep in mind since the approach itself gives no guidance as to how these parameters should be set: this is explicitly framed as a matter of policy, not mathematics or science. We recommend:

- **Building awareness or knowledge of these approaches amongst analysts and civil servants**. This may help practitioners and policymakers assess the risks and trade-offs involved, although this might not in the end lead to the widespread adoption of a given approach as this awareness should also include knowledge of its weaknesses.
- **Some measure of public engagement** (subject to the above caveats) may also be of use here, especially as there is likely to be more concern regarding the disclosure risks of some data sets than others.

Recommendations addressing stakeholders outside of government

Researchers

Factors affecting uptake of synthetic data should be investigated

In general, it seems that the major use of synthetic data for researchers is for exploratory analysis, testing statistical analyses, and training in how to use particular data sets. In the case of the SLS, where provision is available for the generation of high-fidelity 'bespoke' synthetic data for testing and exploratory research, according to one of those we spoke to there is less demand for it than for other uses (e.g. training). The reasons for this are unclear.

There is scope for further engagement and research into the extent of and reasons for why uptake of synthetic data appears to be less than expected in some cases. Factors that could underlie it include:

- Geographic: it was suggested that many researchers live near to the secure data storage facility (in this case, in Edinburgh). This means that the number of researchers for whom the effort of requesting access to and travelling to secure data storage facilities outweighs the effort of applying for a bespoke synthetic data set that can be used outside of these facilities would be fewer than expected.
- Metadata quality: it was suggested in another case that poor metadata quality could have been a factor. Poor metadata can make it unclear what fields the researchers need or should request.
- Force of habit: applying for bespoke synthetic data might be seen as an additional or unfamiliar step in an already onerous process. This would make researchers less likely to engage with it.

We recommend that **the factors that could affect uptake of this kind of service should be investigated further**, in order to determine in which contexts a 'bespoke' synthetic data service is worth pursuing. **It would be interesting to examine whether the Covid-19 pandemic has made any difference to the uptake of synthetic data services.**

Training programs for researchers should be supported

A researcher from the UK Data Service (UKDS) is planning a training program in collaboration with ADR UK with the intention of training researchers on the use of administrative synthetic data, which if successful will likely increase demand for synthetic data of various degrees of accuracy amongst academic researchers. **Programmes of this nature should be supported and encouraged**, and we recommend that if successful, **experience from them could be drawn upon to build a standard program introducing these concepts to government analysts and data holders**.

The general public

Synthetic data must be explained

It is not clear that the general public is aware of synthetic data or its potential uses. This increases the risk that a synthetic data set somehow released to the public might be mistaken or misrepresented as being a genuine data set. It also can limit (where relevant) the ability of members of the public to grant or withhold informed consent that their data be used as the basis for a synthetic data set.

Potential solutions to this include:

- **The production of clear, ordinary language explanations of synthetic data** in order that the concepts involved can be understood by as wide a range of people as possible. This should include concrete examples of what each class of synthetic data can disclose as outlined above. We recommend that this draws upon the high- and low-fidelity terminology established in this paper, leaving the finer distinctions to researchers.

- Clear, standardised labelling of synthetic data sources to minimise the potential for confusion. For example, every synthetic data set released on data.gov.uk might have a header along the lines of "**This is a synthetic data set. The records within do not correspond to real individuals**", perhaps with more specialised subheaders according to the fidelity of the synthetic data. As previously mentioned, the potential of digital watermarking as a means of incorporating identification of the data as synthetic into the data itself should be investigated.

There is a need to find out what the public prefers

In addition, since some practitioners observed potential mismatches between public and institutional tolerance regarding the risks associated with data used for different purposes, **it would be worthwhile researching public attitudes towards the risks of synthetic data**. This could involve experimental trials gauging how the public attitude towards risk as the conditions under which data are used are varied.

While public opinion by itself is unlikely to be a reliable guide to good policy in this area (it is always reasonable to ask how well in general the technical concepts are understood, and legal and ethical constraints should always apply regardless), this might reassure policy makers and implementers. As a result it could mitigate any tendency towards excessive caution when assessing the risks of releasing synthetic data of a particular level of quality for a particular purpose.

5. Limitations and further work

In this report we have engaged with practitioners and non-practitioners within government to establish the potential use cases and appetite for the generation of synthetic administrative data, as well as an expert view on potential problems that might arise.

This engagement was necessary to determine the requirements that tools for the creation of synthetic data must fulfil, and so formed the foundation for the creation of our prototype of a synthetic data generation tool. This tool is intended to generate the kind of safe, low-fidelity synthetic data which we believe is likely to be of immediate use to researchers and analysts if rolled out broadly. We have also presented some recommendations based on our engagement as to how to further the use of synthetic administrative data.

Because of the relatively limited level of engagement (six interviews with interested parties and follow up conversations with senior government analysts and data policy advisors) over a short period of time, the findings of this report should be considered to be exploratory rather than conclusive. Further work and discussions will be needed to not only develop a semi-automated synthetic data tool, but to gather a deeper understanding of attitudes towards the potential benefits and risks of synthetic data within government, amongst academic researchers and the general public, along with how best to implement the recommendations of this report.

Appendix: Implementations of synthetic data

Here we provide a short, non-exhaustive list of interesting implementations of synthetic data, some discussed in the text, others not.

Implementations discussed in the text

United States Census Bureau Synthetic SIPP Data

United States Census Bureau 2020 Census

UK Longitudinal Survey SYLLS data

Northern Ireland Longitudinal Survey Univariate Synthetic Data

Scotland Longitudinal Survey Synthetic Data

Simulacrum

Implementations not discussed in the text

CPRD cardiovascular disease and COVID-19 risk factors synthetic data sets

COVID-19 synthetic data set generated with Synthea

National Covid Cohort Collaborative (N3C) database (including synthetic components)

PaySim synthetic data generator for mobile phone banking fraud detection

Bootstrapping Amazon Alexa’s language capabilities with synthetic data