bmjmedicine

Synthetic data in medical research

Theodora Kokosi, Katie Harron

Population, Policy, and Practice Department, UCL Great Ormond Street Institute of Child Health, London, UK

Correspondence to: Dr Theodora Kokosi, Population, Policy, and Practice Department, UCL Great Ormond Street Institute of Child Health, London WC1N 1EH, London, UK; dora.kokosi@ucl.ac.uk

Cite this as: *BMJMED* 2022;1:e000167. doi:10.1136/bmjmed-2022-000167

Received: 14 March 2022
Accepted: 26 August 2022

Synthetic data have the potential to improve medical research while minimising the need to access personal data; Theodora Kokosi and Katie Harron explain what they are and how they are used.

Introduction

Demand to access high quality data at the individual level for medical and healthcare research is growing. Electronic health record data collected on whole populations can help to generate real world evidence and can be used for a range of secondary purposes, including testing new hypotheses and developing and evaluating different methodological and statistical approaches. Secondary analysis of primary research data, such as from clinical trials,[1] is also valuable—for example, to conduct meta-analyses of individual participant data. However, several complex privacy requirements make accessing these data challenging.[2]

Information contained in electronic health records or in clinical trial data are highly sensitive and access to these datasets can be an expensive and lengthy process.[3] Data privacy and protection regulations are the main barriers to accessing these data for healthcare and medical research.[4] Anonymisation (where potentially identifiable variables are removed) is one way to make data available; however, intensive anonymisation can degrade the data to the extent that it is no longer fit for purpose.[5] For example, adding random noise to the data reduces precision and leads to larger confidence intervals. Several reidentification attempts on anonymised data have been successful and have harmed public and regulators' trust in such methods.[6 7] For instance, one study showed that patients could be identified by matching information from patient level data that was publicly available, attributing information obtained from newspapers, and contacting those patients directly.[6]

Use of information from clinical trials and electronic health records of large populations has the potential to benefit medical and healthcare research and makes seeking new approaches to data access imperative. One solution is to use so-called synthetic data, or artificial data, which provide a realistic representation of the original data source. Synthetic data look like the original data source, without containing any information on any real individuals. Synthetic data can attempt to preserve some of the statistical properties of the original data source (eg, distributions of continuous data, proportions of categorical data, correlations between variables, and other model parameters).

Approaches to generating synthetic data

The aim of data synthesis is to create a dataset that resembles the original individual level data, and retains the same sample size, with rows for each participant and columns for each variable. Characteristics of the original data, including missing values and patterns, are replicated depending on the method chosen to generate the synthetic data. Several methods can be applied for generating such data. In medical research, machine learning methods have predominantly been used, given the complex and high dimensional nature of patients' data. Machine learning methods for constructing synthetic data from the original data sources are typically based on generative models. These models are built to capture and accurately estimate the correct correlations and distributions between different variables in the original data source. Additionally, the models draw on inferences from the original data using bayesian networks via sampling techniques or deep learning via neural networks,[5] such as generative adversarial networks.[8 9] Generative adversarial networks have become particularly popular for use in synthetic data and are used to generate not only synthetic samples but also synthetic images (versions of medical images produced by a wide range of imaging methods) and image translations (conversion of one image representation to another image representation (eg, a grayscale photo to a coloured photo).[10] These techniques attempt to generate synthetic data while dealing with privacy issues as well as patient data that are imbalanced, biased, or from a small sample.[11] However, correction of imbalances can also worsen model performance by leading to poor calibration of risk predictions or wrong absolute risks.[12] Alternative approaches to generative adversarial networks have also been developed more recently, such as ADS-GAN (anonymisation through data synthesis using generative adversarial networks), PATE-GAN (private aggregation of teacher ensembles), and Time-GAN (time-series generative adversarial networks).[13 14]

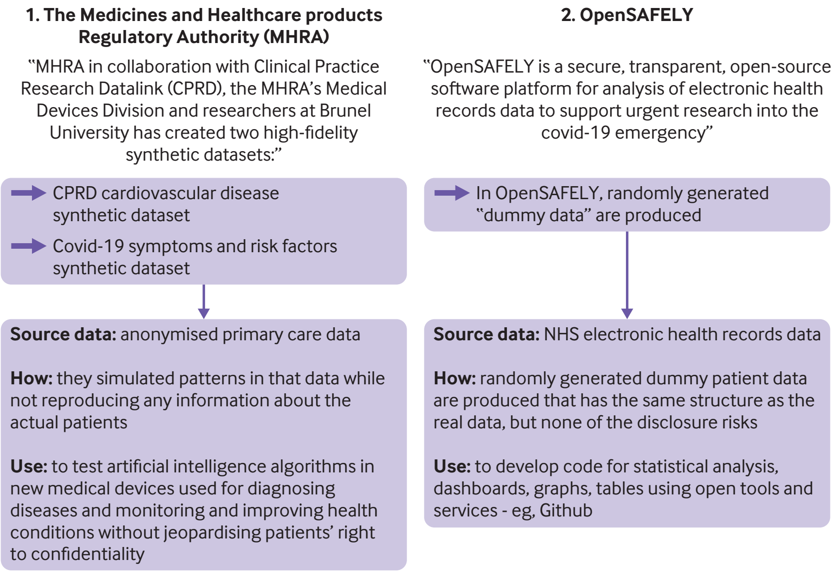
KEY MESSAGES

- ⇒ Synthetic data are artificial data that can be used to support efficient medical and healthcare research, while minimising the need to access personal data
- ⇒ More research is needed to determine the extent to which synthetic data can be relied on for formal analysis, the cost effectiveness of generating synthetic data, and how to accurately assess disclosure risk

Uses and benefits of synthetic data

Some of the most valuable uses for synthetic data are developing code or conducting preliminary hypothesis generation and testing before deployment in real datasets. Researchers can then develop and validate methods for a particular task before accessing real data. This process saves time because data access applications can be conducted in parallel or while waiting for data access to be granted. Synthetic data also help to preserve privacy because the amount of time that researchers need to access sensitive patient information is reduced. This type of data can also be used to improve the reproducibility of research because synthetic datasets can readily be shared with other researchers or third parties to verify models and analysis strategies.[4] Synthetic data can also be used to accelerate methodological developments in medical research and facilitate training and capacity building in methods for handling medical data that are high dimensional and challenging to model. Additionally, synthetic data could be a solution to researchers who are already synthesising clinical study evidence. For example, researchers of a meta-analysis of individual participant data using sufficient statistics from aggregate data and who want to combine data from trials that provide individual participant data in addition to from those that do not.[15] Similarly, synthetic data could be used in simulation studies for sample size calculations for a meta-analysis of individual participant data to account for previous knowledge (eg, number of studies promising individual participant data) in the information available.[16]

Figure 1 presents two examples of how synthetic data are being used in medical research.



1. The Medicines and Healthcare products Regulatory Authority (MHRA)

"MHRA in collaboration with Clinical Practice Research Datalink (CPRD), the MHRA's Medical Devices Division and researchers at Brunel University has created two high-fidelity synthetic datasets:"

- CPRD cardiovascular disease synthetic dataset
- Covid-19 symptoms and risk factors synthetic dataset

Source data: anonymised primary care data

How: they simulated patterns in that data while not reproducing any information about the actual patients

Use: to test artificial intelligence algorithms in new medical devices used for diagnosing diseases and monitoring and improving health conditions without jeopardising patients' right to confidentiality

2. OpenSAFELY

"OpenSAFELY is a secure, transparent, open-source software platform for analysis of electronic health records data to support urgent research into the covid-19 emergency"

- In OpenSAFELY, randomly generated "dummy data" are produced

Source data: NHS electronic health records data

How: randomly generated dummy patient data are produced that has the same structure as the real data, but none of the disclosure risks

Use: to develop code for statistical analysis, dashboards, graphs, tables using open tools and services - eg, Github

Figure 1 | Examples of synthetic data in medical research.[20 21]

Evaluation of synthetic data

Understanding how closely synthetic data replicate original data sources is vital for understanding what the data can be used for; a factor that can be thought of in terms of fidelity.[17] Figure 2 shows the difference between low fidelity data (which do not preserve associations between different variables) and high fidelity data (which do preserve these associations). Low fidelity data can be useful for educational purposes (eg, methodological and software education) and initial data exploration, whereas high fidelity data are more useful for developing models.

The extent to which the synthetic data resemble the original data can be measured in several ways. Metrics include data usefulness, which evaluates the extent to which synthetic data resemble the statistical properties of the original data, and information disclosure, which measures how much of the real data can be shown by the synthetic data.

Approaches for measuring data usefulness include comparing univariate or multivariable distributions of variables between the original and synthetic data, or comparison of model parameters and estimates for multivariate or multivariable models, and interval overlap of confidence intervals.[18] Figure 3 gives an example of a bivariate comparison between original (observed) and synthetic data. The similarity can also be measured between the relative performance of two algorithms (trained and tested) on the synthetic data and their relative performance (when trained and tested) on the original data.

To evaluate disclosure risk, two concepts are considered: identity disclosure, which refers to the risk of an intruder identifying an individual within a sensitive dataset, and attribute disclosure, which refers to the risk of an intruder identifying an individual based on other sensitive attributes of a patient record (eg, medical tests and diagnoses).[10] Several methods can quantitatively assess disclosure risk and attribute disclosure, such as hamming distance, targeted correct attribution probability, and correct relative attribution probability.[10 19]

Challenges and future directions

Although synthetic data methods were introduced more than 30 years ago, these data are not yet widely used in medical and health research, and are associated with several challenges . One area of concern is whether synthetic data would ever be used for decision making or whether final analyses will always need to be conducted on the original data.[4] Furthermore, disclosure risk is minimised in synthetic datasets. However, the risk of including even a small number of unique observations, owing to the nature of the health data (ie, rare diseases or outliers), can pose an

Original data
Birth weight and length of gestation in HES for births in 2019-20

- Birth weight increases with length of gestation
- Some combinations of birth weight and gestation age are invalid (likely due to miscoding of geastational age in days rather than weeks)

High fidelity synthetic data
Each point in the plot was generated by the Classification and Regression Tree technique that subsequently predicts and imputes the value of each variable based on the value of previous variables

- Birth weight increases with length of gestation
- Invalid combinations of birth weight and gestational age are represented in the synthetic data
- High fidelity data preserve the relation between gestational age and birth weight present in the original data, and also preserve some of the "messiness" of the original data (ie, gestational age recorded with errors)

Low fidelity synthetic data
Each point in the plot is randomly generated from unvariable data on birth weight and length of gestation in HES

- Birth weight does not visibly increase with length of gestational
- Many more combinations of birth weight and gestational age are invalid
- Low fidelity data preserve only the univariable distributions of gestational age and birth weight

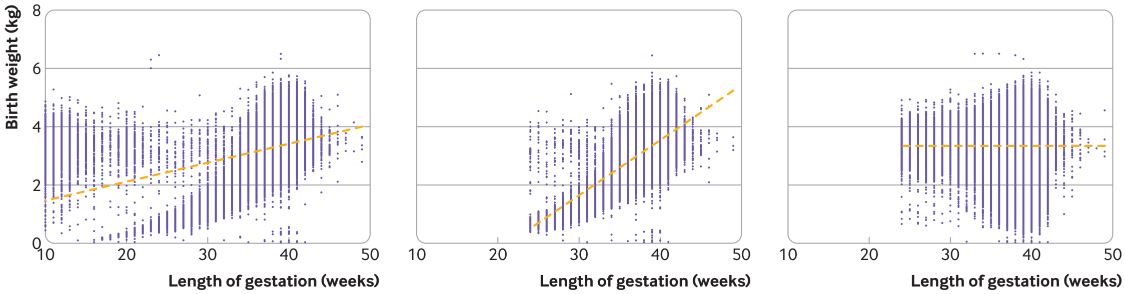


Figure 2 | Examples of high and low fidelity synthetic data. In this example, values of birth weight and length of gestation recorded on birth records in Hospital Episode Statistics (HES) data were used to illustrate high fidelity and low fidelity synthetic data. The lines on the scatterplots represent the regression lines for birth weight on length of gestation.

additional challenge to attribute disclosure. This challenge involves accurately estimating the high dimensional distribution of these data without replicating the information of the individual. Furthermore, additional research is also needed to understand the cost effectiveness of generating synthetic data—that is, whether potential benefits outweigh the time and effort required to generate synthetic data that are fit for purpose.

Twitter Theodora Kokosi @dora_kokosi

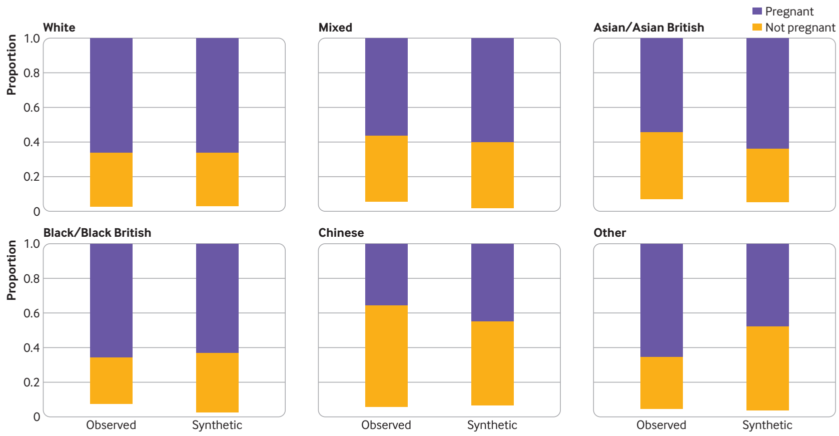


Figure 3 | Example of visual comparison of bivariate distributions between original(observed) and synthetic data. This example is obtained from an analysis of the third National Survey of Sexual Attitudes and Lifestyles (Natsal-3)[19] and shows the proportion of the female survey respondents who answered that they had ever been pregnant, by ethnic group. The method used to generate the synthetic data was Classification and Regression Tree (known as CART).

Contributors KH had the idea for the article, and reviewed and edited the paper. TK did the literature search, write up, and analysis. Both TK and KH are guarantors.

Funding The authors disclosed receipt of the following financial support for the research, authorship, and publication of this article: this work was supported by the Economic and Social Research Council via the Administrative Data Research UK (grant number ES/V005448/1). The funders had no role in considering the study design or in the analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests Competing interests: We have read and understood the BMJ policy on declaration of interests and declare the following interests: none.

Patient and public involvement Patients and members of the public were involved at the initial stage of this research. We consulted and received input from a panel of patients, data users, and advisers who are part of the 'useMYdata' team that supports and promotes the protection of patient data to get feedback on public perceptions of the acceptability of releasing synthetic versions of data or making these publicly available.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement No data are available. Authors do not have permission to share patient level Hospital Episode Statistics data. Hospital Episode Statistics data are available from the NHS Digital Data Access Advisory Group (enquiries@nhsdigital.nhs.uk) for researchers who meet the criteria for access to confidential data.

ORCID iDs

Theodora Kokosi http://orcid.org/0000-0003-1590-6764
Katie Harron http://orcid.org/0000-0002-3418-2856

REFERENCES

1 Azizi Z, Zheng C, Mosquera L, *et al*. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 2021;11:e043497. doi:10.1136/bmjopen-2020-043497

2 Harutyunyan H, Khachatrian H, Kale DC, *et al*. Multitask learning and benchmarking with clinical time series data. *Sci Data* 2019;6:1–18. doi:10.1038/s41597-019-0103-9

3 Dattani N, Hardelid P, Davey J, *et al*. Accessing electronic administrative health data for research takes time. *Arch Dis Child* 2013;98:391–2. doi:10.1136/archdischild-2013-303730

4 Azizi Z, Zheng C, Mosquera L, *et al*. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 2021;11:e043497. doi:10.1136/bmjopen-2020-043497

5 Abay NC, Zhou Y, Kantarcioglu M. *Privacy preserving synthetic data release using deep learning*. Springer, 2018: 510–26.

6 Sweeney L. Matching Known Patients to Health Records in Washington State Data. *SSRN Electron J* [Internet]. 2013 [cited 2022 Jun 6]. Available: http://www.ssrn.com/abstract=2289850

7 Ghafur S, Van Dael J, Leis M, *et al*. Public perceptions on data sharing: key insights from the UK and the USA. *Lancet Digit Health* 2020;2:e444–6 https://linkinghub.elsevier.com/retrieve/pii/S2589750020301618 doi:10.1016/S2589-7500(20)30161-8

8 Park N, Mohammadi M, Gorde K, *et al*. Data synthesis based on generative adversarial networks. *Proceedings VLDB Endowment* 2018;11:1071–83 http://arxiv.org/abs/1806.03384 doi:10.14778/3231751.3231757

9 Kieran C-C, Thomas S, Julia EV. Generation of Heterogeneous Synthetic Electronic Health Records using GANs. 2019 Dec 13 [cited 2022 Jun 6]. Available: http://hdl.handle.net/20.500.11850/392473

10 Goncalves A, Ray P, Soper B, *et al*. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 2020;20:1–40. doi:10.1186/s12874-020-00977-1

11 Tucker A, Wang Z, Rotalinti Y, *et al*. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digit Med* 2020;3:1–13. doi:10.1038/s41746-020-00353-9

12 den GRvan, van Smeden M, Timmerman D, *et al*. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. 2022 [cited 2022 Jun 7]. Available: https://arxiv.org/abs/2202.09101

13 Jordon J, Jarrett D, Saveliev E. Hide-and-Seek privacy challenge: synthetic data generation vs patient re-identification. *In PMLR* 2021:206–15.

14 Yoon J, Drumright LN, van der Schaar M. Anonymization through data synthesis using generative Adversarial networks (ADS-GAN). *IEEE J Biomed Health Inform* 2020;24:2378–88 https://ieeexplore.ieee.org/document/9034117/ doi:10.1109/JBHI.2020.2980262

15 Papadimitropoulou K, Stijnen T, Riley RD, *et al*. Meta-Analysis of continuous outcomes: using pseudo IPD created from aggregate data to adjust for baseline imbalance and assess treatment-by-baseline modification. *Res Synth Methods* 2020;11:780–94 https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1434 doi:10.1002/jrsm.1434

16 Ensor J, Burke DL, Snell KIE, *et al*. Simulation-Based power calculations for planning a two-stage individual participant data meta-analysis. *BMC Med Res Methodol* 2018;18:41 https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0492-z doi:10.1186/s12874-018-0492-z

17 Calcraft P, Thomas I, Maglicic M, *et al*. Accelerating public policy research with synthetic data [Internet]. Available: https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf

18 Snoke J, Raab GM, Nowok B, *et al*. General and specific utility measures for synthetic data. *J R Stat Soc Ser A Stat Soc* 2018;181:663–88. doi:10.1111/rssa.12358

19 Kokosi T, De Stavola B, Mitra R, *et al*. An overview on synthetic administrative data for research. *Int J Popul Data Sci* 2022;7 https://ijpds.org/article/view/1727 doi:10.23889/ijpds.v7i1.1727

20 University of Oxford. OpenSAFELY, 2022. Available: https://www.opensafely.org/about/#:~:text=In%20OpenSAFELY%2C%20the%20data%20management,none%20of%20the%20disclosive%20risks

21 Medicines & Healthcare products Regulatory Agency. UK data driving real-world evidence. Synthetic data, 2022Regulatory Agency. Available: https://cprd.com/synthetic-data