International Journal of Population Data Science

Journal Website: www.ijpds.org

An overview of synthetic administrative data for research

Theodora Kokosi[1,*], Bianca De Stavola[1], Robin Mitra[2], Lora Frayling[3], Aiden Doherty[4,5], Iain Dove[6], Pam Sonnenberg[7], and Katie Harron[1]

Submission History

Submitted:	21/12/2021
Accepted:	25/03/2022
Published:	23/05/2022

[1] Department of Population, Policy and Practice, UCL Great Ormond Street Institute of Child Health, University College London, London, UK
[2] School of Mathematics, Cardiff University, Cardiff UK
[3] Health Data Insight, UK
[4] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[5] Nuffield Department of Population Health, University of Oxford, Oxford, UK
[6] Office for National Statistics, Titchfield, UK
[7] Department of Infection & Population Health, Institute for Global Health, University College London, London, UK

Abstract

Use of administrative data for research and for planning services has increased over recent decades due to the value of the large, rich information available. However, concerns about the release of sensitive or personal data and the associated disclosure risk can lead to lengthy approval processes and restricted data access. This can delay or prevent the production of timely evidence. A promising solution to facilitate more efficient data access is to create synthetic versions of the original datasets which are less likely to hold confidential information and can minimise disclosure risk. Such data may be used as an interim solution, allowing researchers to develop their analysis plans on non-disclosive data, whilst waiting for access to the real data. We aim to provide an overview of the background and uses of synthetic data and describe common methods used to generate synthetic data in the context of UK administrative research. We propose a simplified terminology for categories of synthetic data (univariate, multivariate, and complex modality synthetic data) as well as a more comprehensive description of the terminology used in the existing literature and illustrate challenges and future directions for research.

Keywords

synthetic data; administrative datasets; data linkage; statistical disclosure control; data utility; data confidentiality

*Corresponding Author:
Email Address: dora.kokosi@ucl.ac.uk (Theodora Kokosi)

https://doi.org/10.23889/ijpds.v7i1.1727

Introduction and background

Data collected during the administration of public services e.g., health (such as by the National Health Service in the UK), education, or employment (by the courts or the benefits system) are increasingly being used, especially in the UK, by researchers and policymakers to conduct meaningful research, make informed decisions and deliver impact. In addition, administrative data offer a long-term perspective particularly relevant and useful for policymaking that might otherwise be difficult to examine using small-scale (and shorter-term) data [1]. These administrations keep records of the interactions with the public in order to provide services in an effective way, but they could be more extensively exploited, for example, to provide a representative picture of public service uses and needs within the UK. For example, de-identified data shared with researchers via Administrative Data Research UK (ADR UK) are used for this purpose, with several safeguards put in place to prevent re-identification [2]. These safeguards can prevent disclosure risk to some extent by removing crucial information that could be traced back to individuals (i.e., names, contact details or any unique identifiers), thus allowing researchers to conduct safer research.

One of the greatest benefits of using administrative data for research is when datasets created by one government department or public service are linked with other datasets with the potential to provide a rich dataset to researchers [3]. Linkage between administrative datasets can add further value by bringing together data across multiple sources and thus providing a comprehensive database relevant for researchers from a range of different disciplines including, but not limited to, social science, public policy, finance, and medicine as well as facilitating cross-disciplinary research questions. Access to linked administrative data, which often contain sensitive and/or personal data, is carefully scrutinised. Despite several safeguards that are put into place to ensure data privacy, linked datasets hold the additional risk of a potential "linkage attack" which refers to the attempt of an adversary to re-identify individuals in the linked data by using side information owned by the adversary. This information could be obtained by directly observing target individuals, mining suitable open data or gaining access to the original data through data breaches [4]. Hence, in the UK and elsewhere, measures to protect the confidentiality of data fall under the "Five safes" framework: safe data, safe people, safe projects, safe outputs, and safe settings [5]. Whilst important, these measures can pose a number of challenges for researchers in practice.

The application and approval processes put in place to ensure safe access to administrative data are often lengthy and complicated, taking months and years of research time. As very well documented in a recent paper describing the experiences of applying for linked national data for research, there are several challenges related to process delays, bureaucracy and a lack of clarity between data holders, data controllers/processors, and researchers [6]. These can lead to serious delays in accessing data, jeopardising the timelines of funded research studies and limiting timelines of analyses based on administrative data [7, 8]. Safe settings are highly restrictive, e.g., preventing access to the internet (which can provide useful reference material). Not all secure settings can be accessed remotely, which has been particularly problematic during the COVID-19 pandemic and for researchers needing to travel to access safe settings outside of their usual areas of work. Additionally, there is a further burden in terms of resources required to ensure any outputs leaving the safe setting are not disclosive, which can involve a lengthy checking process before and after using secure lab settings. Finally, data collected in clinical settings can pose additional challenges due to their high dimensionality. Common de-identification methods used to preserve the privacy of individuals while allowing something of value to be extracted from personal information have been useful. However, there is usually at least some risk that those individuals might still be re-identified from de-identified datasets [9]. This could be achieved through auxiliary data an adversary (or intruder) might have available and which they could use to match and re-identify individuals within a dataset.

A promising solution to the challenges faced by users of administrative data is to create 'synthetic' versions of the data. Synthetic data are artificial data that look like the original data sources without containing information on any 'real' individuals, but that attempt to preserve some of the statistical properties of the original data sources. Similar methods of statistical disclosure control (SDC), such as differential privacy, have traditionally been applied to rule out the risk of meaningful information being released [10]. Consequently, privacy-preserving techniques and synthetic data can share similar features; however, the aim of this article is to improve researchers' understanding of some common approaches to generating synthetic data, which might support access to administrative data in the UK. We provide an overview of important considerations and options for generating synthetic data with a focus on UK administrative datasets since this is a field in which synthetic data has a clear potential to support more timely and efficient research in the social and health sciences. In addition, the rationale to focus on UK administrative datasets derives from our experience, and the experience of many researchers in the UK, of the challenges of working with these data [6]. Similar issues might also arise in other countries collecting routine data from government departments (e.g., Australia, Canada, Brazil, and the U.S) [3]. There has been substantial investment in the UK to bridge the gaps between government departments and policy bodies and to make these data more easily available to enable vital and cost-effective research to inform policy.

Given the wide-ranging terminology currently used by researchers for describing the different categories of synthetic data, we propose a simple and consistent terminology that could be used across research teams working in this area. In order to assemble all the information presented in this overview, we first identified the available synthetic datasets and relevant projects on synthetic data in the UK context. Then we expanded our search into the associated concepts of evaluating the utility of the synthetic data as well as disclosure risk. Following that process, we identified available software that has been used for these projects.

We begin by describing different specifications of synthetic data, the different settings in which they might be used, and terminology for different categories of synthetic data. We then explore different synthetic data generation methods, drawing examples from the literature and current research practice. Finally, we consider ongoing challenges and future directions for synthetic data. We focus on the use of administrative data in the UK, but the implications will be generalisable to different contexts including large population-based surveys or longitudinal cohort studies.

A spectrum for synthetic data

The idea of creating synthetic data was introduced in 1993 by Rubin [11] who proposed the development of a technology for releasing, for users of public-use data, synthetic datasets of the U.S Census using generation methods based on multiple imputation. This means that all data points in the original data observations would be used to create fully synthetic datasets of artificial units. Although generating synthetic values for all data points holds many benefits, this is not always necessary and so Little [12] proposed the idea of partially synthetic data, in which only the sensitive to public disclosure variables of the original data are synthesised. However, producing accurate imputation models based on parametric approaches proved rather difficult [13] and thus, non-parametric machine-learning approaches started becoming more popular, such as Classification and Regression Trees [14], support vector machines [15], bagging [16] and random forests [17] for data synthesis [18]. More recently and with the developments in deep learning, there has been an increasing interest in the use of generative adversarial networks (GANs) to generate synthetic data [19, 20].

Synthetic data are artificially generated data designed to mimic real data as closely as possible, but without containing data directly collected from real individuals and hence without personally identifiable information. There is a balance between the extent to which the statistical properties of the original data are retained and the risk of disclosure. We can therefore think of synthetic data as a spectrum, whereby the lower end of the spectrum includes very basic representations of the original data that have no disclosure risk (e.g., by preserving only the data type, format and structure and univariate characteristics). We refer to this end of the spectrum as "low fidelity" synthetic data. The higher end of the spectrum includes very detailed and accurate representations of the original data for which disclosure control will be critical (e.g., by also preserving complex interrelationships between the variables, joint distributions, biological relationships etc.) [21]. We refer to this end of the spectrum as "high fidelity" synthetic data. In the following section, we first propose a simplified terminology that could be consistently used by researchers in this area (see Table 1) and then we also bring together existing terms for categories of synthetic data (Supplementary Table 1) for a more comprehensive presentation of the existing methods commonly used for the generation of synthetic data.

Uses of synthetic data

Synthetic data can be used both alone and in advance of accessing the real data. For example, a sole purpose might be to create a synthetic dataset that can be used for training purposes: a training course on Hospital Episode Statistics (HES; the administrative hospital dataset in England) might use a dataset that has some of the same variables and structure as HES, without including any real data. This would allow a variety of course participants to attend and assess a non-disclosive dataset for practical training sessions, without needing any approvals. Similarly, synthetic data could be used for the evaluation and development of different methodological approaches (including "benchmarking") [22], or as training data for machine learning methods [20, 23, 24]. If the synthetic data were similar enough to the real data, access to the real data may not be required for these purposes.

Alternatively, researchers may want to use synthetic data to develop code before deployment in real datasets, or to conduct preliminary hypothesis generation and testing. These activities could be conducted while approvals to access the real data are being considered, thereby making more efficient use of time. Whilst some suggest that synthetic datasets could also be used in more complex settings such as clinical trials, either as a proxy for real clinical trial data, to make data more broadly available for secondary analysis [25], most agree that making decisions that affect the service or care that an individual receives should not be based solely on synthetic data, given that there is a lack of consensus on how to measure the quality of the synthetic data. For example, a recent summary of synthetic data in the pharmaceutical industry lists a number of uses but excludes decision making or statistical inference and concludes that medical and scientific acceptance is required before results based on synthetic data can be published [26].

Consequently, final analyses are rarely (if ever) conducted on synthetic data and access to the real data (by the researcher themselves or by a data holder) would usually be required. For example, OpenSAFELY[1] (software to support analysis of electronic health records) produces simulated, randomly generated "dummy data" that have the same structure as the real data without disclosing any personal information and allows external researchers to submit code, which is then deployed on NHS health records that are only accessed by a small number of approved analysts within a Trusted Research Environment [27]. This code can be tested against dummy patient data minimising the interactions with real patient data and allowing data users with technical skills to check and apply their methods. In the same vein, low fidelity synthetic data could facilitate data sharing across different government departments, avoiding any potential legal risks and uncertainty about governance requirements [28]. To increase trust in the results obtained from synthetic data, the idea of validation and verification servers has also been proposed. This allows for validating how well the synthetic data generators reproduce the estimates and other characteristics of the original data. The use of validation servers also allows the execution of statistical programmes developed and debugged on synthetic data to run on confidential data with noise added to the estimates to preserve privacy [29]. One example is Synthea, a leading synthetic data generator, method, and software mechanism that examines and validates the quality of healthcare synthetic data [30]; however, it is noted that further research is needed as its capability to model heterogeneous health outcomes is still limited. Ultimately, the appropriateness of using synthetic data in place of original data will depend on the importance of the decisions to be informed.

The purposes for which synthetic data are intended will determine the type of synthetic data that are required (i.e.,

[1] https://www.opensafely.org/about/#:~:text=In%20OpenSAFELY%2C%20the%20data%20management,none%20of%20the%20disclosive%20risks

at which point along the spectrum data will be generated). Synthetic data at the lower end of the spectrum could be used for training purposes as mentioned above as no identical structure and statistical resemblance of the original data is needed. On the other hand, synthetic data at the higher end of the spectrum that resemble the original data could be used for training in data analysis and for exploring and comparing populations or subgroups of populations for clinical trials/interventions. In the context of administrative data, synthetic data can be useful as an interim solution to data access whilst approvals for access to the real data are sought.

There are many different terms in the literature used by researchers, government analysts, and industry stakeholders to describe different categories of synthetic data (Supplementary Table 1). Some are used interchangeably (e.g., dummy data [31] with Synthetic valid or structural [32] as well as the different categories for synthetically-augmented data under the ONS spectrum), which creates confusion around which terms to use (and whether there are any subtle differences among them). Therefore, a need to agree on a consistent simplified terminology has emerged. In Supplementary Table 1, we provide a comprehensive overview and description of the terminology for synthetic data currently used in the literature. In Table 1, we then propose a simplified categorisation of the terminology used in Supplementary Table 1, creating three broad categories for synthetic data based on the characteristics of the data generated: univariate, multivariate, and complex modality synthetic data. In both tables (Table 1 and Supplementary Table 1) we describe how data at the lower end of the spectrum, for example, will retain the structure of the original data (e.g., the number and type of variables) but do not attempt to mirror their statistical properties. Data at this end of the spectrum will have a very low risk of disclosure. In contrast, synthetic data at the higher end of the spectrum will retain statistical relationships between the variables (including any multilevel structure) and allow researchers to test out various analytical strategies and generate preliminary synthetic results. However, there may be a higher risk of disclosure. Both types of data (depending on the level of analysis conducted) could be useful for researchers who want to appreciate the overall layout of the data with the latter helpful for identifying any potential analytical complications before applying for access.

Synthetic data generation methods

Synthetic data generation methods have been previously categorised into two distinct classes: data-driven and process-driven methods [44]. Data-driven methods derive synthetic data from generative models that use original (or observed) data. Some of the most frequently used data-driven methods are i) imputation-based methods [11, 12, 45], ii) full joint probability distribution methods [46], and iii) function approximation method [47]. However, it is noted that some of these techniques might overlap, e.g., multiple imputation methods using Bayesian networks that estimate joint probability distributions [48]. Process-driven methods derive synthetic data from computation or mathematical models of an underlying real-world process and include techniques such as i) numerical simulation [49], ii) Monte Carlo simulations [50], iii) agent-based modelling [51] or iv) discrete-event simulations [52]. Although the above classification of data- and process-driven methods is meaningful, we now present a simplified classification of synthetic data generation methods which could be grouped into three general classes: 1) Generative models, 2) Sampling, and 3) Prediction. Table 2 gives an overview of existing synthetic data generation methods based on those three general classes with examples of current use in the context of UK administrative data (see also the Glossary in Supplementary Table 2 in the Appendix for further explanation of some terms). Many of these approaches have the flexibility to generate data at both ends of the synthetic data spectrum, depending on how the models are specified.

The choice of synthetic data generation methods depends on the field of research and the type of data researchers aim to synthesise. For example, the most frequently used techniques to generate synthetic administrative datasets draw from the statistical methods of sampling and prediction. Specifically, prediction builds on multiple imputation methods, in which values for missing data are imputed based on relationships with original variables and is also the de facto method for generating synthetic data in the context of Statistical Disclosure Control [44]. Prediction approaches effectively treat all values as missing; synthesised variables are generated through imputation based on relationships observed in the real data. Synthetic data are generated by sequentially predicting the value of each variable depending on the value of other variables. On the other hand, with sampling, synthetic data are generated by sampling from distributions estimated from the original data (e.g., as described in the Bayesian Network technique in Table 2).

Evaluating synthetic data

Although synthetic datasets may never produce exactly the same results as the original datasets, it is still important to evaluate how well the synthetic data resemble the real data [68], i.e., whether the structure/type/format of the data and their statistical properties have been preserved. It is generally recommended that the validity of the data should be measured including how well the datatypes and the format of the data have been preserved, followed by a range of more comprehensive evaluation methods to assess whether the synthetic dataset has preserved any of the statistical properties of the original data by comparing important statistical estimates [21].

To evaluate and measure the quality of the synthetic data, different measures of utility are used. Utility of the data refers to the extent to which results from the synthetic data agree with those from the original data and is grouped into two categories, ***general utility*** and ***specific utility*** [69]. To determine the type of evaluation, it is first worth mentioning that this is linked to the purpose of the data that were synthesised, i.e., how the data will be used. For example, simple validity checks might be sufficient for low fidelity univariate synthetic data generated for training and education purposes only. On the other hand, specific utility might need to be assessed for high fidelity multivariate synthetic data which could be used for extended code testing, data analysis,

Table 1: Proposed simplified categorisation of synthetic data (a more comprehensive summary of existing terminology for synthetic data can be found in Supplementary Table 1)

Data utility	Category of synthetic data	Description	Expectations	Uses
Minimally disclosive, minimal analytic value,low fidelity	**Univariate synthetic data**	-Preserves the type, structure and format -Does not contain any original data -Impossible to identify any single entity	-Variables in the synthetic dataset should have similar fundamental aggregated statistics to the ground truth (original data) for both continuous and categorical variables. These include: • Population/cohort-based distribution data (e.g., age, income distributions etc.) • Categorical proportion data (e.g., % of ethnic groups) **-Requires none or light disclosure processes**	i) Basic code or advanced code testing including data management or cleaning ii) Education and training for data analysis iii) Sharing (allows easier sharing of data within or between government departments)
↓	**Multivariate synthetic data**	-Preserves complex inter-relationships between variables -Close representation of values in real individuals in a specific population is expected - Can preserve multivariate distribution for higher-level or low-level geographies and household structure -Can preserve real/logical relationships (joint distributions, e.g., marital status and age) -Can preserve biological relationships (e.g., excessive thirst due to diabetes)	-Preserves statistical distributions of variables and at least some relationships (e.g., correlations) between them -Preserves the original confounding structure of the data • These datasets should be informed by subject-matter expertise or familiarity with real-world distributions. • The more relationships are preserved, the more realistic they can be **-Requires stringent disclosure processes**	i) Extended code testing ii) Extended education and training for data analysis iii) Testing experimental methods iv) Exploring and comparing populations v) Understanding and examining specific subgroups of populations for study or trial/intervention planning
More disclosure risk; more analytic value, high fidelity	**Complex modality synthetic data**	-Data created from perturbations using accurate forward models (i.e., models that simulate outcomes given specific inputs), physical simulation or AI-driven generative models [33]. -This includes specific modalities such as cardiac and radiology images, physiological longitudinal data (e.g., from wearables), genomes, longitudinal data on interactions with public services, prescriptions data etc.	-Quality, effectiveness, and robustness of the synthetic data rely on the quality of the ground truth data supplied to train algorithms -Careful consideration is required around the appropriate generation of high-volume data modalities (e.g., issues such as geometric distortions of body parts inherent to MRIs). **-Requires stringent disclosure processes**	i) Producing high-dimensional data distributions in image synthesis [34, 35] (e.g., image-to-image translation) [36, 37] and speech synthesis i) Increasing robustness and adaptability of AI models and testing stability of machine learning for medicine and healthcare (e.g., predicting early Alzheimer's disease from brain imaging data or medical imaging such as skin lesions [38], pathology slides [39] and other imaging modalities) [40–43]. ii) Hypothesis generation by facilitating the use of data in understanding social and human behaviour (e.g., diagnosing mental health conditions from a smartphone mood diary app or audio recordings without using personally identifiable data)

Table 2: Methods and examples of synthetic data generation (further explanations are provided in the Glossary in Supplementary Table 2)

Techniques	Description	Examples	
Prediction			
Multiple Imputation	Data are generated by simulating multiple copies of the population from which respondents have been selected. All data with missing values are filled in by multiple imputations. A random sample from each of these synthetic populations is then released [53].	**US Census Bureau's Small Area Income and Poverty Estimates** **https://www.census.gov/library/fact-sheets/2021/what-are-synthetic-data.html**	
Classification and Regression Tree (CART)	CART works by using a series of conditional models to sequentially *predict* and impute the value of each variable, depending on the value of other variables, some of which have already been synthesised. Synthetic data are generated by sequentially predicting the value of each variable, depending on the value of other variables. This approach builds on multiple imputations, in which values for missing data are imputed based on the relationships with original variables. Prediction approaches effectively treat all values as missing and synthesised variables are generated through imputation based on relationships observed in the real data [18]	**Synthetic versions of the Scottish Longitudinal Study (SLS)** **https://sls.lscs.ac.uk/guides-resources/synthetic-data/**	R package *synthpop*[2] [54]
Autoregressive models	An autoregressive (AR) model predicts future behaviour based on past behaviour. It is used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them. The process is a linear regression of the data in the current series against one or more past values in the same series. The AR process is an example of a stochastic process, which has degrees of uncertainty or randomness built-in. Randomness means that while future trends can be predicted based on past data, predictions will never be 100% accurate. Usually, the process gets "close enough" for it to be useful for generating synthetic data. AR models are also called conditional models, Markov models, or transition models [55].		
Sampling			
Bayesian Network	In this approach, the relationships between the variables are specified within a graphical structure (e.g., directed acyclic graph) and joint probability distributions (or contingency tables if categorical) for all the variables are derived from the original data [17]. Synthetic data are generated by sampling from these distributions.	**The National Cancer Registry ('Simulacrum'), generated by Health Data Insight and Public Health England** **https://healthdatainsight.org.uk/project_category/synthetic-data/**	**"Simulacrum codebase"** in **MATLAB**
Bayesian models using MCMC	MCMC chains are used to sample from the joint posterior distribution of the variables to be synthesised. A mixture of variable types can be handled by transforming the observed joint distribution into a multivariate normal distribution. Multi-level data can be synthesised (i.e., to preserve hierarchical structures such as pupils within schools or patients within hospitals) [56, 57].	**UK Primary care data for public health research** **i) CPRD cardiovascular disease dataset[3] [46, 58]** **ii) CPRD COVID-19 symptoms and risk factors synthetic dataset[4] [46, 58]**	R package **bnlearn**[5] R package ***jomo*** [59]
Synthetic minority over-sampling (SMOTE)	SMOTE is essentially performing resampling (creating new data points for the minority class) and instead of duplicating observations, it creates new observations along the lines of a randomly chosen point and its nearest neighbours, i.e., it synthesises new data examples between existing (real) examples. SMOTE was initially developed to help address the problems with applying classification algorithms to unbalanced datasets and thus, it does not replicate data in the general region of the minority samples, but on exact locations. This method was proposed in 2002 [60] and could be used to train neural network classifiers for medical decision making [61].		

Continued

Table 2: Continued

Techniques	Description	Examples
Static spatial microsimulation	This technique generates synthetic populations by combining individual-level samples (i.e., microdata or seed) and aggregated census data (i.e., macro data or target). Its intended use is for cases where spatial microsimulation is desirable, i.e., where the individuals belong to household and regional hierarchical structures. The main method used for synthetic reconstruction is called Iterative Proportional Fit [62, 63] which builds up a synthetic dataset for a small area using Census tables leading to an entirely synthetic dataset that is created by a joint probability distribution specified using attributes conditional on existing (known) attributes.	
	Generative models	
General Adversarial Networks (GANs)	GANs generate two competing neural network models. One takes noise as input and generates *samples* (the generator or forger). The other model (the discriminator) receives samples from both the generator and the training data and attempts to distinguish between the two sources. There is a continuous game between the two networks where the generator is learning to produce more and more realistic samples while the discriminator is learning to get better and better at distinguishing generated data from real data. GAN is successful if these two networks co-operate well, and both learn at the expense of one another and attain equilibrium over time [64].	**Quantifying utility and preserving privacy in synthetic datasets (QUIPP) – The Alan Turing Institute** [Conditional GAN for Tabular data (CTGAN)] **https://www.turing.ac.uk/research/research-projects/quipp-quantifying-utility-and-preserving-privacy-synthetic-data-sets** **Synthetic data in Machine learning for healthcare** **https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/** (PATE-GAN[6], ADS-GAN[7], and TimeGAN techniques[8])
Autoencoders and Variational autoencoders (VAEs)	The autoencoder's deep network consists of two individual deep neural networks. The first of these networks is called the encoder and compresses the input (original) data into a shortcode. The second deep network is called the decoder and it is a mirror image of the encoder and its purpose is to decompress the shortcode generated by the encoder into a representation that closely resembles the original data [65]. The variational autoencoders (VAEs) are a more modern version of autoencoders. VAEs use the same architecture as autoencoders but impose added constraints on the learned encoded representation. Those two techniques use the *sampling* method to produce new samples which are similar to those in the original dataset but not exactly the same [66].	
Recurrent Neural Networks (RNNs)	RNNs are networks with loops in them, allowing information to persist, i.e., they remember things from prior inputs while generating outputs. This architecture equips RNNs with a form of internal state (memory) enabling them to exhibit temporal dynamic behaviours and to process sequences of inputs. In RNNs the aim is to build a generative model which captures the joint probability, p(x, y), of the inputs x and the output y. This probability can be used to sample data or to make predictions by using Bayes rules to calculate the posterior probability p(y|x) and then estimating the most likely output [67].	

[2] For access to code for the synthpop package: https://github.com/cran/synthpop.

[3] https://www.cprd.com/content/synthetic-data#CPRD%20cardiovascular%20disease%20synthetic%20dataset.

[4] https://www.cprd.com/content/synthetic-data#CPRD%20COVID-19%20symptoms%20and%20risk%20factors%20synthetic%20dataset.

[5] For access to code and details about the R package bnlearn: https://cran.r-project.org/web/packages/bnlearn/index.html.

[6] http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html.

[7] https://ieeexplore.ieee.org/document/9034117.

[8] https://papers.nips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html.

exploring specific populations or as proxies for real patient data in clinical trials to inform intervention planning and accelerate research.

The general utility of the data should summarise the utility of the dataset overall and can be assessed by inspecting the marginal distributions of key variables [68] against the original data and then estimating the standardised propensity mean square error (pMSE) [69, 70] or distances between distributions (e.g., Kullback-Leibler or Hellinger distance). Other simple methods for evaluating synthetic data against the original data may include 1) visual comparisons of univariate distributions (e.g., inspecting cell counts or the frequency of the distributions (for categorical data) or comparing histograms and box plots (for continuous variables). Specific utility can be evaluated according to how the data will be used, for example by comparing estimated coefficients from selected models between the real and the artificial datasets [71]. Other ways of evaluating specific utility are calculating the relative or absolute differences in mean scores of standardized values or examining whether point estimates from synthetic data fall within the 95% confidence interval of the point estimates obtained from the real data (larger overlap = higher similarity) [72]. Otherwise, z-scores can be used to measure the difference in statistics (e.g. the Chi-squared coefficient) that are used to analyse associations within the original and synthetic data [21]. More complex comparisons can be done among multivariate distributions that would involve comparing the model parameters estimated from each dataset.

To measure the extent to which the statistical properties of the real data have been retained in synthetic multivariate structures, correlations between variables in the synthetic datasets can be inspected. For example, pairwise Pearson correlations between the synthetic and the original data (if continuous and symmetrically distributed) could be used to visualise the differences in the correlation matrices and calculate an overall mean of the correlation differences [73]. Also, other model-based approaches such as logistic regressions for binary variables and multiple linear regressions for continuous variables can be applied to examine multivariate structures and compare confidence intervals for the estimated parameters.

In addition, dimension reduction techniques can also explore whether statistical properties of the real data are appropriately reflected in the synthetic data (e.g., whether certain factors that can be derived from the original data can also be derived in the synthetic datasets). These methods include Principal Component Analysis (PCA) [74], t-Distribution Stochastic Neighbor Embedding (t-SNE) [75] and Multiple Correspondence Analysis (MCA) [76]. Finally, with the rise of synthetic data in the field of data science and machine learning another measure of utility has been developed – that is measuring the difference in the performance of supervised machine learning models trained on synthetic and on original data. Although a recent study suggested that the majority of the models trained on synthetic data had lower accuracy compared to the models trained on original data, these deviations are expected and manageable and highlight the potential of synthetic data and how we could further evaluate their robustness [77].

Evaluating disclosure risk

In addition to data utility, information disclosure risk should also be taken into consideration as these two metrics relate to equally important, and typically complementary but also competing, objectives, and thus should be considered jointly. Traditionally, one solution to deal with disclosure risk and the associated challenges is to allow the release of data with reduced confidentiality risk to an acceptable level under the differential privacy concept [10, 78], using statistical disclosure control (SDC) methods. Early SDC approaches involved recoding data, data swapping, or adding noise, and k-anonymity (performed via suppression and generalization of the data). However, methods used to prevent re-identification of individuals or attribute disclosure risk could degrade the data to such an extent that they are no longer fit for purpose. Concerns around disclosure risk may also remain, particularly when increasing numbers of variables are brought together from different sources through linkage.

Although the purpose of the synthetic data is to control confidentiality risk in place of the traditional SDC, this does not mean that the synthetic datasets do not pose any disclosure risks. Specifically, information disclosure is about the extent to which the real data may be revealed, directly or indirectly, by the synthetic data [44]. There are two classes of privacy disclosure risk, ***identity disclosure*** and ***attribute disclosure***, where the former refers to the risk of an attacker identifying an individual within a confidential and sensitive dataset and the latter is about the risk of an attacker identifying sensitive information about an individual in a dataset from a set of attributes known to the attacker, e.g., by matching relevant variables. For example, there is the possibility of identification disclosure in partially synthetic data, as discussed by Reiter and Mitra [79], as the probabilities of identification in the released data could be computed. On the other hand, it is less likely that personal records will be identified in fully synthetic data, as these do not contain any personal information on any individual. Thus, one way to minimize disclosure risk would be to ensure that the synthetic data depend as little as possible on the original data, so that an intruder who has partial knowledge of a particular population would not be able to make inferences based on values of sensitive variables for that population.

Disclosure risk in synthetic data depends on the methods used to generate the data, and the original data themselves; there is a lack of evidence on which generation methods are "safest" for which purposes [80]. How to measure disclosure risk in synthetic data is an ongoing and active field of research. Recent work in this area includes a method called Targeted Correct Attribution Probability by Elliot and Taub [80], Correct Relative Attribution Probability [68], and a Generalized Method [81] of Taub's initial approach of Correct Attribution Probability which is about assigning a risk probability for the exposure of the real value of the corresponding individual in the original dataset instead of ignoring those non-matches or assigning probability 0. Privacy mechanisms have been added to some software that implements the synthetic data generation methods, in order to handle disclosure risk post synthetic data generation. For instance, synthpop in R includes tools for statistical disclosure control that label and remove unique replicates of unique actual individuals in the original data (e.g., bottom and/or top coding, smoothing, and excluding variables from being identified as unique).

Challenges and future directions for synthetic data

Challenges of using synthetic data

Generation of synthetic data is an important development for research using administrative data and other longitudinal population-based studies as it makes data accessible for exploration that can benefit the research community before applying for and gaining access to the datasets. However, there are a number of challenges that have to date prevented the wider use of synthetic data for administrative data research. For one, disclosure risk can remain a concern, as there is still the possibility of unintentionally identifying individuals and disclosing private information, especially when synthetic data are too accurate or resemble the original data too closely. This means that even when the data are fully synthetic, specific outliers that occur in patient populations with rare diseases, for example, could be identified when the data generated have preserved the same statistical properties. Therefore, if synthetic data are to be made widely available (outside of safe settings, for example), or involve sensitive topics such as sexual behaviour, then some level of SDC (e.g., low number suppression) may still need to be considered to minimize the residual risk of re-identification. This may involve removing or editing outliers or unique records, or in some cases, applying SDC methods to an entire dataset. Alternatively, we could make more use of hide-and-seek challenges, where trusted users are invited to try to re-identify individuals (to provide additional confidence that the data are not disclosive) [82].

Second, comparisons of synthetic data generation methods can be difficult. While a variety of synthetic data generation methods exist, evaluation metrics are not consistently applied across different approaches. According to Walonoski and colleagues [83] who developed Synthea, an open-source software package that simulates the lifespans of synthetic patients, validation of claims of success and methodologies in synthetic data generation methods are often superficial and focus only on the overall structural appearance or general statistical comparison under the concept of evaluating the general utility of the synthetic data. Apart from that, there are several challenges present when measuring the utility of both low- and high-fidelity synthetic data. For example, although low-fidelity synthetic data do not hold any disclosure risk, there is the likely accusation of poor-quality data because of undesirable combinations of variables from different data fields (e.g., age and marital status with children being shown as married) if data have not been through checks. On the other hand, these types of utility challenges would not be present with high-fidelity data, but it is important to highlight that, as we attempt to produce complex, high-quality synthetic data and as the accuracy increases, the disclosure risk also increases and differential privacy methods, such as adding noise, might be applied to ensure that disclosure risk remains at an acceptable level [31]. Consequently, the modelling performance of those data may be compromised making them invalid[9]. Thus, a consistent and appropriate evaluation of the utility of synthetic data (alongside an assessment of disclosure risk) is needed [44].

A third challenge is how to enable repeatability and reuse of different methodologies, which requires comprehensive and detailed documentation of the synthetic data generation methods [24]. The quality of synthetic data depends on the quality of the input data and the data generation model, as far as data-driven methods are concerned. Creating synthetic datasets that are fit for purpose requires time, effort, and output controls to ensure accuracy when comparing the synthetic with the original datasets. Having a well-defined purpose for the data can make the generation and evaluation easier. Therefore, being clear about the purpose will provide a more complete assessment of the quality of the synthetic data [44].

Fourth, administrative data often contain very large numbers of variables and can include complex data types when linked to other sources such as clinical data that include imaging, free-text, and genomics. Thus far, researchers have attempted to generate synthetic data from a small number of variables with certain data types (e.g., numerical) (see [84]) due to the complexity of the synthetic data generation. However, in order to generate synthetic versions of administrative datasets, there is a need to synthesise datasets with hundreds of thousands or millions of records and several hundreds of variables with different data types (e.g., categorical or string). The most effective ways of synthesising these different variable types are not yet known. Furthermore, additional considerations required for synthesising linked data may need to be taken into account (e.g., whether to synthesise data before or after they are linked, and whether relationships between variables in different datasets can be retained) and specific algorithms to deal with large-scale data should be applied. As discussed by Raghunathan and colleagues [85], further research needs to be conducted on the synthetic data generation methods for longitudinal studies as those pose additional challenges. Those challenges can include difficulties in adding new information to the already synthesised data that has changed across waves and increased disclosure risk when linking longitudinal data across waves as more information is being released about the individuals or households than if the data were not linked [86].

Future directions for synthetic data

Finally, the lack of consensus on the appropriate use of synthetic data has created scepticism about the benefits of synthetic data. While some argue that it is better to create multivariate synthetic datasets that preserve not only the data format, and types but also the complex statistical relationships between variables [85], others recognise the potential of univariate synthetic data for planning and training [31]. Additionally, acceptability to the public has not yet been assessed and it is important for data holders to effectively communicate the benefits of synthetic data for research and to develop trust with data users and the public. This may

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

be done most easily by starting with low fidelity data, which may help to build acceptance of synthetic data. Recently, our research team engaged with a panel of patients, data users and advisers who are part of the "use MY data" team[10] that supports and promotes the protection of patient data, in order to get feedback on public perceptions of the acceptability of releasing synthetic versions of data or making these publicly available. The general feedback from the discussion was that the group found synthetic data an interesting idea, however, they were unfamiliar with the term and its potential uses. One of the main issues that came up was the extent to which an individual's private information would be protected in a synthetic dataset, and whether the use of synthetic data could be a cost-effective way of accessing data in the future. Further public engagement is needed to communicate and promote the potential benefits arising from the uses of synthetic data as well as to understand public misconceptions and perceptions around creating synthetic versions of administrative datasets.

Conclusions

We provide an overview of synthetic data generation methods in the context of UK administrative data research and propose a simplified categorisation of synthetic data: univariate, multivariate, and complex modality synthetic data. We discuss benefits, potential challenges and future directions in the field. While access to administrative datasets can facilitate meaningful and impactful research, it is also important to retain the privacy and confidentiality of personal and sensitive data collected using public funds and services. Generating synthetic datasets can minimise some of the main challenges related to information disclosure risk as they can be used in place of the real data either for training reasons or to accelerate research [44]. Further understanding is needed from the research community on the uses of synthetic data and how those are generated as well as public engagement and collaboration between data producers and researchers/data users.

Acknowledgements

We would like to thank Prof Neil Sebire at Health Data Research UK (HDRUK) for his input on this overview.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Ethics statement

This research article did not require ethical approval because it involves information freely available in the public domain.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or 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]. This research was supported in part by the NIHR Great Ormond Street Hospital Biomedical Research Centre. AD is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), Novo Nordisk, the British Heart Foundation Centre of Research Excellence (grant number RE/18/3/34214), the Alan Turing Institute and the British Heart Foundation (grant number SP/18/4/33803), and Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities.

References

1. Penner AM, Dodge KA. Using administrative data for social science and policy. RSF Russell Sage Found J Soc Sci. 2019;5(3):1–18. https://doi.org/10.7758/RSF.2019.5.3.01

2. Mc Grath-Lone L, Libuy N, Harron K, Jay MA, Wijlaars L, Etoori D, et al. Data Resource Profile: The Education and Child Health Insights from Linked Data (ECHILD) Database. Int J Epidemiol. 2022;51(1):17–17f. https://doi.org/10.1093/ije/dyab149

3. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data Soc. 2017;4(2):2053951717745678. https://doi.org/10.1177/2053951717745678

4. Fiore M, Katsikouli P, Zavou E, Cunche M, Fessant F, Le Hello D, et al. Privacy of trajectory micro-data: a survey. 2019; https://doi.org/10.48550/arXiv.1903.12211

5. Ritchie F. The 'Five Safes': A framework for planning, designing and evaluating data access solutions [Internet]. Available from: https://uwe-repository.worktribe.com/output/880713/the-five-safes-a-framework-for-planning-designing-and-evaluating-data-access-solutions

6. Taylor JA, Crowe S, Pujol FE, Franklin RC, Feltbower RG, Norman LJ, et al. The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement. BMJ Open. 2021;11(8):e047575. https://doi.org/10.1136/bmjopen-2020-047575

7. Dattani N, Hardelid P, Davey J, Gilbert R. Accessing electronic administrative health data for research takes time. Arch Dis Child. 2013;98(5):391–2. https://doi.org/10.1136/archdischild-2013-303730

8. Morris H, Lanati S, Gilbert R. Challenges of administrative data linkages: experiences of Administrative DataResearch Centre for England

[10] https://www.usemydata.org/about.php.

(ADRC-E) researchers. Int J Popul Data Sci IJPDS. 2018;3(2). https://doi.org/10.23889/ijpds.v3i2.566

9. Elliot M, O'hara K, Raab C, O'Keefe CM, Mackey E, Dibben C, et al. Functional anonymisation: Personal data and the data environment. Comput Law Secur Rev. 2018;34(2):204–21. https://doi.org/10.1016/j.clsr.2018.02.001

10. Dwork C. Differential privacy: A survey of results. In Springer; 2008. p. 1–19. https://doi.org/10.1007/978-3-540-79228-4_1

11. Rubin DB. Statistical disclosure limitation. J Off Stat. 1993;9(2):461–8.

12. Little RJ. Statistical analysis of masked data. J Off Stat-Stockh-. 1993;9:407–407.

13. Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Comput Stat Data Anal. 2011;55(12):3232–43. https://doi.org/10.1016/j.csda.2011.06.006

14. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 2017; https://doi.org/10.1201/9781315139470

15. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In 1992. p. 144–52. https://doi.org/10.1145/130385.130401

16. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40. https://doi.org/10.1007/BF00058655

17. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324

18. Nowok B. Utility of synthetic microdata generated using tree-based methods. UNECE Stat Data Confidentiality Work Sess. 2015;

19. Trampert P, Rubinstein D, Boughorbel F, Schlinkmann C, Luschkova M, Slusallek P, et al. Deep neural networks for analysis of microscopy images—synthetic data generation and adaptive sampling. Crystals. 2021;11(3):258. https://doi.org/10.3390/cryst11030258

20. Tremblay J, Prakash A, Acuna D, Brophy M, Jampani V, Anil C, et al. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In 2018. p. 969–77. https://doi.org/10.48550/arXiv.1804.06516

21. Bates A, Špakulová I, Dove I, Mealor A. Synthetic data pilot. Working paper series, Office for National Statistics; 2018. Available from: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot#authors

22. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data. 2019;6(1):1–18. https://doi.org/10.1038/s41597-019-0103-9

23. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. ArXiv Prepr ArXiv171204621. 2017; https://doi.org/10.48550/arXiv.1712.04621

24. Sankaranarayanan S, Balaji Y, Jain A, Lim SN, Chellappa R. Learning from synthetic data: Addressing domain shift for semantic segmentation. In 2018. p. 3752–61. https://doi.org/10.48550/arXiv.1711.06969

25. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. 2021;11(4):e043497. https://doi.org/10.1136/bmjopen-2020-043497

26. James S, Harbron C, Branson J, Sundler M. Synthetic data use: exploring use cases to optimise data utility. Discov Artif Intell. 2021;1(1):1–13. https://doi.org/10.1007/s44163-021-00016-y

27. Curtis HJ, Inglesby P, Morton CE, MacKenna B, Walker AJ, Morley J, et al. Trends and clinical characteristics of COVID-19 vaccine recipients: a federated analysis of 57.9 million patients primary care records in situ using OpenSAFELY. MedRxiv. 2021; https://doi.org/10.1101/2021.01.25.21250356

28. Bogiatzis Gibbons D, Friemann P, Kolker E, Collerton E, Sutherland A. Applying Behavioural Insights to Cross-government Data Sharing [Internet]. Available from: https://www.adruk.org/fileadmin/uploads/adruk/Documents/BIT-ADRUK_Applying_BI_to_HMG_Datasharing_Dec20.pdf

29. Burman LE, Engler A, Khitatrakun S, Nunns JR, Armstrong S, Iselin J, et al. Safely expanding research access to administrative tax data: creating a synthetic public use file and a validation server. Technical report US, Internal Revenue Service; 2019. Available from: https://www.urban.org/research/publication/safely-expanding-research-access-administrative-tax-data-creating-synthetic-public-use-file-and-validation-server

30. Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Med Inform Decis Mak. 2019;19(1):1–9. https://doi.org/10.1186/s12911-019-0793-0

31. Calcraft P, Thomas I, Maglicic M, Sutherland A. Accelerating public policy research with synthetic data [Internet]. Available from: https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf

32. Office for National Statistics. ONS methodology working paper series number 16 - Synthetic data

pilot [Internet]. Available from: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot

33. Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5(6):493–7. https://doi.org/10.1038/s41551-021-00751-8

34. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. Wavenet: A generative model for raw audio. ArXiv Prepr ArXiv160903499. 2016; https://doi.org/10.48550/arXiv.1609.03499

35. van den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. In PMLR; 2016. p. 1747–56. https://doi.org/10.48550/arXiv.1601.06759

36. Wang T-C, Liu M-Y, Zhu J-Y, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In 2018. p. 8798–807. https://doi.org/10.48550/arXiv.1711.11585

37. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In 2017. p. 2223–32. https://doi.org/10.48550/arXiv.1703.10593

38. Ghorbani A, Natarajan V, Coz D, Liu Y. Dermgan: Synthetic generation of clinical skin images with pathology. In PMLR; 2020. p. 155–70.

39. Mahmood F, Borders D, Chen RJ, McKay GN, Salimian KJ, Baras A, et al. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. IEEE Trans Med Imaging. 2019;39(11):3257–67. https://doi.org/10.1109/TMI.2019.2927182

40. Costa P, Galdran A, Meyer MI, Niemeijer M, Abràmoff M, Mendonça AM, et al. End-to-end adversarial retinal image synthesis. IEEE Trans Med Imaging. 2017;37(3):781–91. https://doi.org/10.1109/TMI.2017.2759102

41. Frangi AF, Tsaftaris SA, Prince JL. Simulation and synthesis in medical imaging. IEEE Trans Med Imaging. 2018;37(3):673–9. https://doi.org/10.1109/TMI.2018.2800298

42. Nie D, Trullo R, Lian J, Wang L, Petitjean C, Ruan S, et al. Medical image synthesis with deep convolutional adversarial networks. IEEE Trans Biomed Eng. 2018;65(12):2720–30. https://doi.org/10.1109/TBME.2018.2814538

43. Zhou T, Fu H, Chen G, Shen J, Shao L. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. IEEE Trans Med Imaging. 2020;39(9):2772–81. https://doi.org/10.1109/TMI.2020.2975344

44. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol. 2020;20(1):1–40. https://doi.org/10.1186/s12874-020-00977-1

45. Yucel RM, Zhao E, Schenker N, Raghunathan TE. Sequential hierarchical regression imputation. J Surv Stat Methodol. 2018;6(1):1–22. https://doi.org/10.1093/jssam/smx004

46. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med. 2020;3(1):1–13. https://doi.org/10.1038/s41746-020-00353-9

47. Abay NC, Zhou Y, Kantarcioglu M, Thuraisingham B, Sweeney L. Privacy preserving synthetic data release using deep learning. In Springer; 2018. p. 510–26. https://doi.org/10.1007/978-3-030-10925-7_31

48. Di Zio M, Scanu M, Coppola L, Luzi O, Ponti A. Bayesian networks for imputation. J R Stat Soc Ser A Stat Soc. 2004;167(2):309–22. https://doi.org/10.1046/j.1467-985X.2003.00736.x

49. Kim J, Fox MF, Field A, Nam YU, Ghim Y-C. Conditions for generating synthetic data to investigate characteristics of fluctuating quantities. Comput Phys Commun. 2016;204:152–8. https://doi.org/10.1016/j.cpc.2016.04.004

50. Kiviet JF. Monte Carlo simulation for econometricians. Found Trends®Econom. 2012;5(1–2):1–181. https://doi.org/10.1561/0800000011

51. Amadi M, Shcherbacheva A, Haario H. Agent-based modelling of complex factos impacting malaria prevalence. Malar J. 2021;20(1):1–15. https://doi.org/10.1186/s12936-021-03721-2

52. Bartz-Beielstein T, Bartz E, Rehbach F, Mersmann O. Optimization of High-dimensional Simulation Models Using Synthetic Data. ArXiv Prepr ArXiv200902781. 2020; https://doi.org/10.48550/arXiv.2009.02781

53. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. J Off Stat. 2003;19(1):1. Retrieved from https://www.proquest.com/scholarly-journals/multiple-imputation-statistical-disclosure/docview/1266794989/se-2

54. Nowok B, Raab GM, Dibben C. synthpop: Bespoke creation of synthetic data in R. J Stat Softw. 2016;74:1–26. https://doi.org/10.18637/jss.v074.i11

55. Penny W, Roberts S. Bayesian methods for autoregressive models. In IEEE; 2000. p. 125–34. https://doi.org/10.1109/NNSP.2000.889369

56. Goldstein H, Carpenter JR, Browne WJ. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. J R Stat Soc Ser A Stat Soc. 2014;177(2):553–64. https://doi.org/10.1111/rssa.12022

57. Lee MC, Mitra R, Lazaridis E, Lai A-C, Goh YK, Yap W-S. Data privacy preserving scheme using generalised linear models. Comput Secur. 2017;69:142–54. https://doi.org/10.1016/j.cose.2016.12.009

58. Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Comput Intell. 2021;37(2):819–51. https://doi.org/10.1111/coin.12427

59. Quartagno M, Grund S, Carpenter J. Jomo: a flexible package for two-level joint modelling multiple imputation. R J. 2019;9(1). https://doi.org/10.32614/rj-2019-028

60. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953

61. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Netw. 2008;21(2–3):427–36. https://doi.org/10.1016/j.neunet.2007.12.031

62. Birkin M, Clarke M. Synthesis—a synthetic spatial information system for urban and regional analysis: methods and examples. Environ Plan A. 1988;20(12):1645–71. https://doi.org/10.1068/a201645

63. Birkin M, Clarke M. The generation of individual and household incomes at the small area level using synthesis. Reg Stud. 1989;23(6):535–48. https://doi.org/10.1080/00343408912331345702

64. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Process Syst. 2014;27.

65. Zhang X, Fu Y, Zang A, Sigal L, Agam G. Learning classifiers from synthetic data using a multichannel autoencoder. ArXiv Prepr ArXiv150303163. 2015; https://doi.org/10.48550/arXiv.1503.03163

66. Wan Z, Zhang Y, He H. Variational autoencoder based synthetic data generation for imbalanced learning. In IEEE; 2017. p. 1–7. https://doi.org/10.1109/SSCI.2017.8285168

67. Behjati R, Arisholm E, Bedregal M, Tan C. Synthetic test data generation using recurrent neural networks: a position paper. In IEEE; 2019. p. 22–7. https://doi.org/10.1109/RAISE.2019.00012

68. Raab GM, Nowok B, Dibben C. Practical data synthesis for large samples. J Priv Confidentiality. 2016;7(3):67–97. https://doi.org/10.29012/jpc.v7i3.407

69. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. J R Stat Soc Ser A Stat Soc. 2018;181(3):663–88. https://doi.org/10.1111/rssa.12358

70. Woo M-J, Reiter JP, Oganian A, Karr AF. Global measures of data utility for microdata masked for disclosure limitation. J Priv Confidentiality. 2009;1(1). https://doi.org/10.29012/jpc.v1i1.568

71. Steinbakk GH, Langsrud Ø, Løland A. A brief overview of methods for synthetic data for official statistics. 2020; Available from: https://nr.brage.unit.no/nr-xmlui/bitstream/handle/11250/2682092/methods-for-synthetic-data-SAMBA-23-20.pdf?sequence=1

72. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. Am Stat. 2006;60(3):224–32. https://doi.org/10.1198/000313006X124640

73. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes. 2019;12(7):e005122. https://doi.org/10.1161/CIRCOUTCOMES.118.005122

74. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010;2(4):433–59. https://doi.org/10.1002/wics.101

75. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11).

76. Greenacre M, Blasius J. Multiple correspondence analysis and related methods. Chapman and Hall/CRC; 2006. https://doi.org/10.1201/9781420011319

77. Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. JMIR Med Inform. 2020;8(7):e18910. https://doi.org/10.2196/18910

78. Willenborg L, De Waal T. Elements of statistical disclosure control. Vol. 155. Springer Science & Business Media; 2012.

79. Reiter JP, Mitra R. Estimating risks of identification disclosure in partially synthetic data. J Priv Confidentiality. 2009;1(1). https://doi.org/10.29012/jpc.v1i1.567

80. Taub J, Elliot M. The Synthetic Data Challenge. 2019; Available from: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf

81. Hittmeir M, Mayer R, Ekelhart A. A baseline for attribute disclosure risk in synthetic data. In 2020. p. 133–43. https://doi.org/10.1145/3374664.3375722

82. Jordon J, Jarrett D, Saveliev E, Yoon J, Elbers P, Thoral P, et al. Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-identification. In PMLR; 2021. p. 206–15. https://doi.org/10.48550/arXiv.2007.12087

83. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230–8. https://doi.org/10.1093/jamia/ocx147

84. Kaloskampis I. Data science for public good Synthetic data for public good. Available from: https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

85. Raghunathan TE. Synthetic data. Annu Rev Stat Its Appl. 2021;8:129–40. https://doi.org/10.1146/annurev-statistics-040720-031848

86. Mitra R, Blanchard S, Dove I, Tudor C, Spicer K. Confidentiality challenges in releasing longitudinally linked data. Trans Data Priv. 2020;13(2):151–70. Available from: https://orca.cardiff.ac.uk/146471/

87. Wu G, Heppenstall A, Meier P, Purshouse R, Lomax N. A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. Sci Data. 2022;9(1):1–11. https://doi.org/10.1038/s41597-022-01124-9

88. Caterini AL, Chang DE. Deep neural networks in a mathematical framework. 2018; https://doi.org/10.1007/978-3-319-75304-1

89. Ballard DH. Modular learning in neural networks. In 1987. p. 279–84.

90. Napierala K, Stefanowski J. Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst. 2016;46(3):563–97. https://doi.org/10.1007/s10844-015-0368-1

91. Chen SH, Pollino CA. Good practice in Bayesian network modelling. Environ Model Softw. 2012;37:134–45. https://doi.org/10.1016/j.envsoft.2012.03.012

92. Soni D. Introduction to Bayesian Networks. Available from: https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e

93. Brownlee J. Neural Networks are Fucntion Approximation Algorithms. Available from: https://machinelearningmastery.com/neural-networks-are-function-approximators/

Supplementary Table 1: Existing categorisations of synthetic data currently used in the literature. A simplified categorisation is proposed in Table 1 of the main text

Data utility	Fully/partially synthetic	Common terms for synthetic data type	Description	Exemplar data case
Minimally disclosive, minimal analytic value, low fidelity	**Fully synthetic data**[11]	**- Dummy data [31]**	-Preserves the type, structure and format of real data but not the statistical properties of the variables -Do not contain any original data - Almost impossible to identify any single entity	i) Basic code or advanced code testing including data management or cleaning ii) Technical development (API, tool or pipeline testing) iii) Education and training (for data analysis)
		- Synthetic datasets[12] **[32]** • Valid	- Preserves only the structure, format, and data types of the variables - Constructed based only on available metadata; values are generated from ad-hoc distributions and open sources - Contains only values present in the original (univariate) data. -No disclosure risk	
		• Structural	-Preserves the format and record-level plausibility (i.e., values use plausible distributions) and replicates marginal (univariate) distributions where possible -Produced dataset passes the sanity check (validation condition or edit rules) the real dataset would need to go through. - Missing value codes, errors and inconsistencies of the original data are present -Minimal disclosure risk	
	Fully & Partially synthetic data	**- Population level synthetic data [87]** **- Patient level synthetic data [46]**	- Key characteristics of variables in the original data (e.g., distributions) are preserved -Complex inter-relationships between variables are not considered -Preserved complex inter-relationships between variables (for each individual) -Close representation of values in real people in a specific population	i) Explore and compare populations ii) Education and training (for data analysis) iii) Testing experimental methods iv) Extended code testing v) Understanding and examining specific groups or populations for study or trial planning vi) Develop analysis plans vii) Produce preliminary results prior to accessing the real data
		- Synthetically-augmented datasets [32] • Synthetically-augmented plausible	-Preserves the format and record-level plausibility and replicate marginal (univariate) distributions where possible -Constructed based on real dataset, values are generated based on original distributions (with added fuzziness and smoothing) -Does not preserve relationships	
		• Synthetically-augmented multivariate plausible	-Preserves the format and record-level plausibility and replicate multivariate distribution loosely for higher level geographies -Constructed based on real dataset, values are generated based on original distributions (with added fuzziness and smoothing) -Some key relationships are retained	

Continued

Supplementary Table 1: Continued

Data utility	Fully/partially synthetic	Common terms for synthetic data type	Description	Exemplar data case
↓		- Synthetically-augmented multivariate detailed	-Similar to previous but more effort to match the real relationships (joint distributions), e.g., in smaller geographies and household structure	
		- Synthetically-augmented replica	-Preserves format, structure, joint distributions, missingness patterns, low level geographies. -Constructed based on the real dataset, values are generated based on observed joint or conditional distributions, while de-identification methods are applied. **-In all types of synthetically-augmented datasets, missingness is to be preserved and disclosure control is necessary case by case**	
More disclosure risk; more analytic value, high fidelity		**-Complex modality synthetic data**	-This includes specific modalities such as radiology images, ECG time series data	i) Machine learning for e.g., medicine and healthcare ii) Facilitating the use of data in understanding social and human behaviour

[11] The term "Fully synthetic data" refers to datasets in which all variables are generated, and original values are included. The term "Partially synthetic data" refer to datasets in which only some variables, typically those with sensitive information, are generated while some of the original variables are still present. Also, for the terms assigned in the "Fully and Partially synthetic data" category, datasets can either contain fully synthesized data or partially synthesized data based on the purpose of the research and the statistical modelling applied.

[12] The terms "Synthetic datasets" and "Synthetically-augmented datasets" refer to a high-level scale to evaluate the synthetic data based on how closely they resemble the original data, their purpose and disclosure risk and is proposed by the Office of National Statistics (ONS). More details can be found here:
https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot

Supplementary Table 2: Glossary of synthetic data generation techniques

Terms	Definition
Deep Neural Networks	Deep learning is part of the machine learning methods based on artificial neural networks (a technology built to simulate the activity of the human brain) with representation learning. As such, deep neural networks are networks with a certain level of complexity and consist of an input layer, an output layer and at least one hidden layer in between. Each layer performs specific types of sorting and ordering ('feature hierarchy process'). These networks use sophisticated mathematical modelling to process data in complex ways [88].
General Adversarial Networks (GANs)	GANs belong to a class of generative models within the artificial intelligence (AI) and machine learning field and represent a powerful way of learning any kind of data distribution based on unsupervised learning. GANs aim to learn the true data distribution of the training (input) dataset and attempt to generate new data points from this distribution with some variations and not just reproducing the old data the model has been trained on. GANs try to use the power of neural networks (as described above) to learn a function to approximate the approach to model a distribution as close as possible to the real data [80].
Autoencoders	Autoencoders were originally introduced as a method for learning meaningful representations from data in an unsupervised manner and the concept of autoencoders in the context of artificial neural networks was first presented by Ballard.....[89]. An autoencoder is a feed-forward deep neural network that first compresses the input data into a more compact representation and then attempt to reconstruct the original input by using an in-between layer which restricts the amount of information that travels within the network. Autoencoders have been frequently used for data compression and dimensionality reduction and can learn nonlinear relationships [84].
Minority class	This term refers to classification predictive modelling in machine learning (ML) and involves predicting a class label for a given observation. Most of the ML algorithms used for classification in predictive modelling were designed with the assumption of an equal number of examples for each class. However, imbalanced classification problems might occur (i.e., where the distribution of examples across the known classes is bias or skewed) and one of the target classes can contain a much smaller number of instances than the other classes (minority class) [90].
Bayesian Network	Bayesian networks represent systems as a network of interactions between variables from primary cause to final outcome, with all cause-effect assumptions made explicit....[91]. They are a type of probabilistic graphical model that uses Bayesian inference for probability computations. Bayesian networks aim to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph. Through these relationships, one can efficiently conduct inference on the random variables in the graph through the use of factors [92].
Function approximation	Function approximation is a technique for estimating an unknown underlying function using historical or available observations from the domain. Artificial neural networks learn to approximate a function [93].