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]**	- Preserves only the structure, format, and data types of the variables	
		• Valid	- 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]	-Preserves the format and record-level plausibility and replicate marginal (univariate) distributions where possible	
		• Synthetically-augmented plausible	-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].