Discussion Paper

Synthetic Data Terminology in Official Statistics

Rosanne J. Turner
Reinoud Stoel
Peter-Paul de Wolf

July 2026

Abstract

Synthetic data finds its foundation in statistical disclosure control. It was formalized by Rubin in 1993 as a technique to enable safer distribution of microdata to the public. Today, synthetic data has evolved into an interdisciplinary field spanning statistics, computer science, and artificial intelligence. It is no longer used solely for disclosure control but also for a wide range of applications, including data sharing, machine learning development, testing algorithms, and addressing data scarcity. At the same time, ongoing research continues to address serious challenges related to data utility, disclosure risk, and evaluation methods. In this paper, we aim to create clarity in the terminology surrounding synthetic data, specifically for application in official statistics. We first give a brief history of synthetic data methodology to place the definition of synthetic data in a historical context. Subsequently, we describe the bulk of the terminology related to synthetic data, propose definitions and scopes for each term, and reflect on the different methods from the view of different application scenarios in official statistics.

Keywords: data generation, simulation, AI-generated data, statistical disclosure control

This is an internally reviewed preprint ahead of journal publication

Contents

1 Introduction

In essence, all models for generating data are wrong, but some are useful (a modification of the famous quote by Box and Draper (1987)). Originally coined in Rubin (1993), the term synthetic data only really gained traction in recent years. Synthetic data is generated by a statistical process that extracts information from an actual dataset, and is then expressed as a collection of artificial or synthetic datasets. This allows, in theory, for dissemination of the informational content of the actual microdata and, at the same time, in theory limits the exposure to potential inadvertent or malicious disclosure of sensitive information about the subjects (Raghunathan 2021).

However, the definition of synthetic data above is broad and includes a broad spectrum of data-generation methods, from small "white-box" models to deep neural networks. Further, there is a plethora of related terms that are often used interchangeably in applications. Some examples of these terms are dummy data, random data, mock data, simulated data and AI-generated data. It is important to be specific while communicating about data generation and applications, as potential utility and privacy aspects can vary widely between methods.

Key while generating data is the intended goal. There are many potential applications of generated data, such as testing, education, supporting scientific claims, prediction and knowledge generation. Not all data-generation methods capture the right nature of information to match all these applications (Bauwelinckx et al. 2025). Because of this, it is crucial to match the data-generation method to the intended goal, especially when the goal of releasing a generated dataset is to support scientific claims, prediction or knowledge generation. On the other hand, one should not apply methods overly complex for the intended goal to avoid unnecessary costs and possible high disclosure risks.

Disclosure risk is a second important aspect in differences between data-generation methods. Because the nature of the information captured from the real-world data is different for each data-generation technique, the choice of a specific technique has great impact on the ability to safely disseminate (synthetic) generated data and/ or models for data generation. Some models only capture univariate summary statistics or correlation patterns, whereas others are able to exactly copy and reproduce a real-world dataset and hence generate data that should adhere to the same disclosure control rules as real-world data.

Without a shared language, it is difficult to communicate about and collaborate with others regarding data-generation methods. An example of this is dummy data. Often, the term dummy data is used to describe placeholder data used for testing computer systems that have no connection to any real-world dataset and hence no or a negligible disclosure risk. At the same time, in computer science dummy data can also concern carefully constructed synthetic data suitable for studying complex database attacks, incorporating a lot of sensitive information. A researcher who uses the term dummy data for arbitrary placeholder data might think the project data from the latter researcher can be published without the need for statistical disclosure control checks.

In this paper, we aim to create clarity in the terminology surrounding synthetic data, specifically for application in official statistics. We first give a brief history of synthetic data methodology to place the definition of synthetic data in a historical context. Subsequently, we describe the bulk of the terminology related to synthetic data, propose definitions and scopes for each term, and place all techniques within a spectrum of utility and privacy based on the level of detail at which the model captures information from the original data.

2 A Brief History of Synthetic Data Methodology

In the following section we present a brief history of synthetic data methodology: from the origin in statistical disclosure control and official statistic to the shift towards computer science and machine learning approaches. A more detailed overview can be found in Drechsler and Haensch (2024).

The concept of synthetic data was formalized as a tool for privacy protection by Donald Rubin in 1993. He proposed using multiple imputation to replace microdata releases with imputed microdata, potentially allowing datasets to be shared without revealing confidential information (Rubin 1993), extending previous work by Little and others on protecting microdata by masking parts of records by synthetic values (Little 1993). In this framework, synthetic data is generated from statistical models estimated on the original data, and multiple datasets are created to reflect the inherent uncertainty. These datasets can then be analysed separately, and the results combined using specific rules, enabling valid statistical inference despite the release of original microdata (Reiter and Raghunathan 2007). This idea established the foundation for synthetic data as a principled method for statistical disclosure control (SDC, also called statistical disclosure limitation in the US).

Although related ideas existed earlier (e.g. Liew et al. (1985)), synthetic data remained largely confined to the statistical community for several years. One explanation for this limited adoption lies in the dominant privacy frameworks in computer science at the time. Models such as k-anonymity (Sweeney 2002), as well as extensions like l-diversity and t-closeness, defined privacy in terms of the properties of the released dataset. These approaches required modifying the dataset directly to meet certain criteria, such as ensuring that individuals could not be uniquely identified. Synthetic data, however, does not necessarily satisfy these criteria, even though they may provide strong intuitive privacy protection. As a result, synthetic data did not fit naturally into these frameworks and were not widely adopted in computer science during this period.

A major conceptual shift occurred with the introduction of differential privacy (Dwork et al. 2006). Differential privacy redefined privacy by focusing on the mechanism used to generate or release data rather than on the dataset itself. It requires that the inclusion or exclusion of any single individual has only a limited impact on the output, thereby providing a formal and quantifiable privacy guarantees. Certain key characteristics of differential privacy such as the need to add noise to individual statistical outputs, resulting in inconsistencies between outputs, and determining an appropriate privacy budget upfront make direct application in official statistics difficult (Domingo-Ferrer et al. 2025). Nevertheless, this shift in perspective aligns naturally with synthetic data, since privacy can be ensured through the data-generation process. Following this development, synthetic data became increasingly relevant in the computer science literature, and a variety of differentially private data synthesis methods were proposed. These include approaches based on contingency tables, Bayesian networks, and copula models, all designed to generate data while satisfying formal privacy guarantees.

Another major turning point came with advances in machine learning, particularly the introduction of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). GANs consist of two neural networks—a generator and a discriminator—that are trained simultaneously in a competitive process. The generator produces synthetic data, while the discriminator attempts to

distinguish between real and synthetic observations. Through this iterative process, the generator learns to produce increasingly realistic data. Although GANs were initially developed for image and signal data, they were later adapted to microdata, more common in official statistics and social sciences.

The application of GANs to microdata introduced new challenges, such as handling categorical variables, modelling complex dependencies, and dealing with relatively weak correlations between variables. To address these issues, various extensions and alternative generative models were developed, including variational autoencoders, diffusion models, Bayesian networks, and copula-based methods. These approaches improved the ability to capture high-dimensional relationships and made synthetic data generation more robust and flexible. As a result, the field experienced rapid growth, with increasing integration of statistical and machine learning techniques.

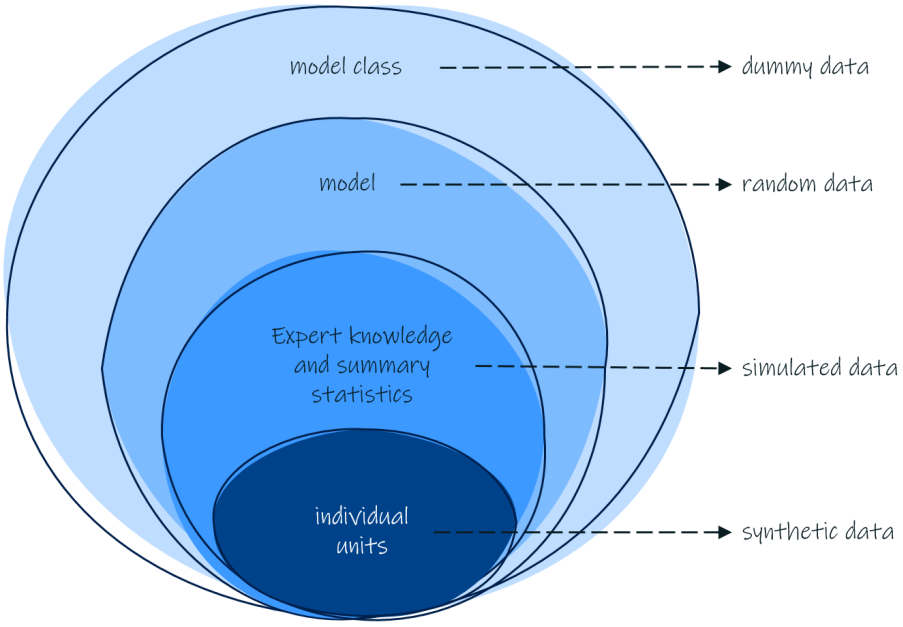
Today, synthetic data has evolved into an interdisciplinary field spanning statistics, computer science, and artificial intelligence. It is no longer used solely for disclosure control but also for a wide range of applications, including data sharing, machine learning development, testing algorithms, and addressing data scarcity. At the same time, ongoing research continues to address serious challenges related to data utility, disclosure risk, and evaluation methods. This evolution reflects a broader shift from traditional parametric modelling toward flexible, data-driven generative approaches, making synthetic data a central tool in modern data science.

3 Terminology Surrounding Synthetic Data Methodology

The evolution of synthetic data and continuous expansion of applications have also led to a shift and expansion in terminology, at some points losing the connection with the original origins in statistical disclosure control and the release of microdata. Since synthetic data is generated data, in this section we will start with defining and describing the most common terms surrounding methodology for generating data. We only consider machine-generated data here. Data generation for example by performing experiments by hand, like rolling dice, are outside the scope. A classification of these terms into synthetic and non-synthetic data is proposed based on properties of real-world data incorporated in the methods generating the data. This classification is illustrated in Figure 3.1: the terms are ranked based on the richness of information the models capture from real-world data. For official statistics, this property is key: on the one hand, the information captured determines which down-stream uses of generated data could potentially satisfy the strict quality, consistency and transparency standards held in official statistics. On the other hand, the detail of information captured by models determine the level of disclosure control needed. The terms are each described in this section, first the related terms, then synthetic data and lastly, imputation is discussed.

In this paper we focus on generating microdata. The microdata that is being replaced or mimicked by the generated microdata is called the real-world data. The terminology we present could be applied to multi-modal data (images, videos, free text) as well, but in our discussion we focus on applications relevant to official statistics and hence mainly on microdata, as multi-modal data is not yet often part of the information stored and disseminated by statistical offices. We omit the direct obfuscation of tabulated data (frequency tables and magnitude tables), covered extensively in statistical disclosure control literature (see for example Hundepool et al. (2012)).

Figure 3.1 Properties of real-world data incorporated in different types of data-generation methods. Arbitrary data are not included in this overview because they do not incorporate any information from real-world data. Dummy data only incorporate information on the model classes that can describe the real-world data. Random data incorporate information on model class and specific models apt for illustrating certain theories or properties. Simulated data models incorporate expert knowledge and information from summary statistics. Finally, synthetic data incorporates information from individual units in the real-world data into a data-generation model.



3.1 Related Terms

3.1.1 Arbitrary data

Arbitrary data is completely random data, generated with completely arbitrary methods, without any model-specification based on real-world information at all. In a sense, white noise without any assumptions on the distribution of the noise is a form of arbitrary data. This means that no data or metadata are used to define the process for generating arbitrary data. An example could be a placeholder matrix containing machine default values. Arbitrary data is sometimes referred to as "fake data" or "mock data". However, we propose to avoid these terms, as they can evoke the association of misleading or manipulative data and are often used in that context (Muammar and Shaalan 2026; Piller 2026; Rane and Subhashini 2026).

3.1.2 Dummy data

Closely related to the terms above, dummy data concerns data that is purely intended to be used as a placeholder, that is, a "dummy". The term is often used in computer science in relation to testing and constructing databases. A dummy dataset contains no information derived from individual entries in the real-world dataset it is a placeholder for. It can contain information on the model class the data is generated from, solely based on metadata. For example, for categorical data, categories can be specified. Another example is integer variables that are specified to fall within a certain range (example in OpenSAFELY (2021)). No information from real-world data is directly incorporated in the dummy data-generation process, which also means that outliers or new categories in real-world data are not necessarily incorporated in dummy data. The choice of the data-generation distribution from the model class is arbitrary. An example could be filling a database with age placeholders drawn from a uniform distribution, the choice for the uniform distribution made purely out of convenience, because it is integrated in standard software testing libraries.

3.1.3 Random data

Random data, also called fictitious data in the literature, is data where the data-generation process is restricted to a specific model based on an experiment setup (e.g. Calder et al. (2022)). This model choice is based on certain theoretical properties, to enable a paper's author to support performance claims or to illustrate new methods. Values of model parameters are determined based on the purpose of the random data, or are set arbitrarily. No information related to a real-world dataset is incorporated. An example could be a researcher performing power analyses, drawing samples from distributions with various sample and effect sizes to illustrate a newly developed test statistic. Another example could be drawing p-values from a uniform distribution ($U(0, 1)$) for educative illustration purposes on the expected behaviour of a p-value under a null hypothesis.

3.1.4 Simulated data

Simulated data should not only serve as a placeholder or to illustrate certain theoretical properties, but mimic real-world data more closely. We define simulated data as data where metadata and high-level aggregate statistics from previously collected real-world data, enriched with expert knowledge and/or literature knowledge, are used to define a data-generation process. This term does not include data-generation models that use unit-level real-world data. Another term sometimes used to describe these methods is generated data. We omit this term as it is too broad, since data in all methods in this section are in essence "generated" by a process or model. An example of simulated data could be to extract a mean and bivariate correlation from a dataset with normally distributed variables to specify a distribution to draw new samples

from to illustrate software used in a scientific publication, where the original study data cannot be published due to privacy concerns. Another example could be generating an exam scores dataset for courses where students can receive a score between 0 and 100 for each exam by drawing from a uniform distribution ($U(0, 100)$), based on previous studies reporting that grades approximately follow this distribution.

3.2 Synthetic Data

We propose to define synthetic data as data on the unit level that are generated by means of a statistical model that is estimated on real-world microdata. In line with the literature (Hundepool et al. 2012) we distinguish between fully synthetic data and partially synthetic data.

3.2.1 Fully synthetic data

Methods that fit algorithms on a unit level to real-world datasets with the purpose of generating entirely new versions of these datasets are methods for fully synthetic data generation (see Figure 3.2). There is a wide range of methods that fall under this definition, for example multiple imputation (as mentioned in Section 2), Latin hypercube sampling, machine learning approaches like classification and regression trees (CART), bagging, random forests, support vector machines, general adversarial networks (GAN) and variational auto-encoders (VAE). We deem AI-generated data a subcategory of fully synthetic data. It is defined as data generated with deep learning or algorithms with a similar level of complexity as deep learning algorithms, trained with the purpose of synthesizing data. AI-generated data also incorporates the output of generative AI models, such as text and images.

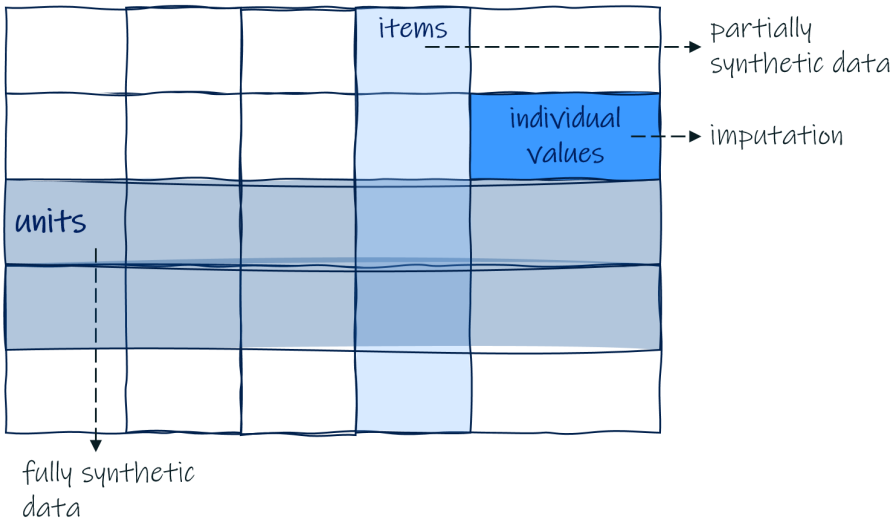
A special category of synthetic data generation is agent-based modelling, recently becoming more researched in the fields of social sciences and official statistics for generating demographic microdata (Chapuis et al. 2022). Agent-based modelling might appear to have its foundation in simulated data as described above, with agents and their interactions being defined with expert knowledge and summary statistics (Bonabeau 2002). However, fine-tuning and validating sound agent-based models require intensive use of real-world data on the unit level. Further, some types of agent-based models utilize synthetic data as input data (Chapuis et al. 2022). Hence, we deem data generated by agent-based models synthetic data.

3.2.2 Partially synthetic data

Partially synthetic data is data where only certain, often sensitive, variables or items are synthesized (Hundepool et al. (2012); also see Figure 3.2). All non-sensitive entries in the real-world data are published in their original form, but the sensitive variables are synthesized. Many methods for generating partially synthetic data exist, ranging from more classical statistical approaches such as regression and Cholesky decomposition (Hundepool et al. 2012), to data science and machine learning-like methods (Slokom et al. 2020). Closely related is hybrid data, where a weighted combination is made of original and synthetic unit-level data. Both partially synthetic and hybrid data retain the same number of units in the synthetic dataset as in the real-world dataset, having implications for disclosure risk.

Of the terms listed above, partially synthetic data, fully synthetic data and AI-generated data fall under the term “synthetic data”. The other forms of data described should not be called synthetic data, and could be placed under the broader term “generated data”. This may avoid confusion about the attractive properties but also the privacy risks associated with synthetic data.

Figure 3.2 Types of synthetic data and the different elements of a dataset they comprise.



3.2.3 Simulated data versus synthetic data

Some synthetic data definitions include purely "knowledge-based" methods, similar to what we call simulated data. There are however important distinctions between what we describe as simulated data and synthetic data that are especially relevant for applications in official statistics.

Would a statistical office want to enable people to simulate data based on their statistics, they would only need to share high-level aggregated statistics with the outside world for them to sample observations from. They could for example share a vector of means and covariate matrix to enable linear regression through structural equation modelling. The release of high-level aggregated statistics instead of the simulated dataset has a few advantages. The statistical office could perform general SDC to ensure an acceptable and explainable level of disclosure risk in the shared statistics. The low level of detail of information captured by these summary statistics and hence limited applicability would be immediately clear, preventing unsupported claims to be drawn from analyses with the release.

Labelling and directly releasing a simulated dataset by a statistical office as "synthetic" can be confusing, as this evokes expectations about the level of detail captured by the data and about possible disclosure risks. With synthetic data moving from the statistical to the data science realm and with many state of the art synthesis techniques being of high complexity, expectations around the level of detail captured in synthetic microdata are generally high. For example, the general sentiment surrounding synthetic data is that it can be utilized for training complex AI models (Capasso 2025). However, when data-generation methods are used that only capture high-level summary statistics, the resulting generated dataset is not apt for reflecting complex dependencies in the real-world data. It is most likely not apt for prediction tasks or studying higher-order interactions between variables. When users of the released simulated microdata would perform such analyses anyway, this would lead to invalid results. We describe this in more detail in Section 4. At the same time, as described in Section 2, synthetic microdata requires an entirely different SDC approach compared to simulated data. Requesting a privacy officer or SDC expert to review the planned release of a synthetic dataset would result in an entirely different (more labour-intensive) course of action than when requesting to review a release of high-level summary statistics.

In summary, we advocate for refraining from referring to data generated with high-level

summary statistics as “synthetic data” and for releasing the summary statistics directly, instead of a simulated dataset based on these statistics.

3.3 Imputed Data

Imputation is a technique closely related to data-generation methods and was originally primarily aimed at completing data with missing values at the item-level (Kalton and Kasprzyk 1982). This is illustrated in Figure 3.2. One or more models are fit on the incomplete data, and used to generate data to fill in missing values. One could wonder why one would aim to complete data directly before executing an analysis, instead of performing an analysis on incomplete data or using other estimates or sufficient statistics for analyses. This might stem from the origin of official statistics in the census, where direct counting of units was a central task, or from the scarcity of user-friendly statistical software for the analysis of incomplete data in previous decades.

Often-used techniques are mean imputation, regression imputation and donor-based imputation (Waal et al. 2011). Multiple imputation (which is mentioned in Section 2, and was the start of the development of synthetic data techniques) was originally proposed by Rubin as an extension that makes it relatively easy to evaluate the accuracy of estimates based on imputed data. Mass imputation is another technique and is often used for data-integration when combining registers and surveys, to fill all missing values for missing observations in one of the two datasets using auxiliary variables (Daalmans 2017). All imputation techniques (although mean imputation less than the others, depending on the precise implementation) use unit-level observations to estimate distributions to draw or predict missing values.

When imputation is used as illustrated in Figure 3.2, to fill in some missing values at the unit-level, we do not deem the resulting dataset a synthetic dataset but an imputed dataset. Imputation to fill in missing values in this way has its origins in dealing with missing values, not in SDC, in contrast to synthetic data. When mass imputation or multiple imputation are applied with the goal of filling missing observations for all values of one dataset or to create entirely new observations, the resulting datasets are partially synthetic and fully synthetic, respectively. These techniques do have a connection to SDC. In practice, at least some background characteristics of units to be imputed are usually known, such as main economic activity and size class of a company or municipality for a person. Hence, in official statistics, when imputation is applied for synthesis this most often results in partially synthetic datasets.

4 Congeniality and Applicability of Synthetic Data in Official Statistics

In this section we will discuss the relation between different forms of generated data, analysis goals and disclosure risk. We will first introduce the concept of congeniality and, closely related, testing utility of synthesized data. Then we will discuss in detail Table 4.1 depicted below, illustrating the aptness of different forms of generated data for several applications in official statistics. We consider using generated data for software testing, educative purposes, predictive modelling and distributing synthetic microdata as public use files for on-the-fly analysis by any member of the public.

Table 4.1 Different methods for data generation, their relation to four types of intended use in official statistics and associated risk of disclosure.

Type of generated data	Intended use				Risk of disclosure
	Software testing	*Education*	*Statistical modelling*	*Public use file*	
Arbitrary data	very limited	not apt	not apt	not apt	no risk
Dummy data	apt in simple scenarios	not apt	not apt	not apt	no risk
Random data	apt in specific scenarios	not apt	not apt	not apt	no risk
Simulated data	apt in most scenarios	apt in specific scenarios	apt in very limited scenarios	not apt	low, able to test with standard SDC techniques
Synthetic data	unjustifiable cost in most scenarios	apt in most scenarios: check congeniality	apt in specific scenarios: check congeniality	most often uncongenial or biased	high, most often impossible to test with standard SDC techniques
AI-generated synthetic data	unjustifiable cost in most scenarios	apt in most scenarios: check congeniality	apt in specific scenarios: check congeniality	most often uncongenial or biased	very high and often unclear how to protect model and synthetic data

4.1 Congeniality

Uncongeniality in synthetic data can occur when the data-generation method does not correspond to the methods that will be used for analyses on the synthetic data (Meng 1994). Congeniality is thus a property of the relationship between the synthesis model and the analysis model. The degree to which this property is attained in practice can subsequently be assessed empirically. A simple example of congeniality are data simulated with the sample mean vector and covariance matrix for an analyst who wants to perform linear regression with structural equation modelling (although one could in this case directly share these sufficient statistics instead of entirely newly generated microdata, see also Section 3.2). A trivial example of uncongeniality: when generating synthetic data by modelling each column of the real-world data independently, one cannot expect to find valid bivariate correlations in the resulting synthetic dataset. A less obvious example is that modelling a joint probability distribution under regularization (as many generative models like GANs and VAEs do) is not the same as modelling a causal process. Synthetic data generated by an arbitrary GAN do not necessarily preserve the original causal structure of the real-world dataset, leading to spurious findings in time series analysis of synthetic data and sometimes even reversed effects (Bauwelinckx et al. 2025).

While synthetic data was previously mostly being developed in computer science with a focus on predictive or discriminative machine learning modelling, congeniality was not much of an issue (Drechsler and Haensch 2024). However, when exploring the aptness of synthetic data for replacing public use files or data under contract in official statistics, where it is impossible to anticipate on all analyses, congeniality becomes very relevant. Rubin, in the original work on multiple imputation for creating synthetic data, already stated that an imputation model that correctly captures relationships among data and corresponding estimation methods is essential. He also noted it is generally impossible for the imputer to anticipate on all possible estimations in advance (Loong and Rubin 2017). Generally, for imputation, if the imputer's model is more general than the analyst's model, inferences are often valid (Murray 2018).

The reader might think these are specific challenges of the classical statistical "white box" models applied by Rubin and colleagues. What about machine learning models, and especially deep learning models, that are able to capture such complex patterns in data? Here, the principle of Occam's razor is key. To avoid overfitting and achieve acceptable external validity, most machine learning models are trained with regularization of parameters to ensure a model that is as simple and generalisable as possible is selected. However, this might not lead to selection of a model that preserves the properties in the real-world data needed for specific analyses, for example preserving causal or timely relations (Bauwelinckx et al. 2025). Ex ante, it is impossible to state that a synthetic dataset synthesized with machine learning techniques will lead to universally valid analysis results.

Interestingly, all of the above highlights that the party deciding on the synthetic data algorithm, also decides which reality is synthesized (Susser and Seeman 2024). They have the power: the information they capture with their synthesis methods will determine the results and validity of the analyses performed on the synthetic data. The choice of synthesis model can even reinforce the choice for a specific analysis model in downstream analysis (Fössing and Drechsler 2025). This is especially important with regard to responsible machine learning, as synthetic data algorithms can introduce bias for small groups or vulnerable groups, especially when those groups are categorized by protected variables (Bhanot et al. 2021).

4.2 Utility of Synthetic Data

Given the above, one can obviously not generate a synthetic dataset using arbitrary machine learning methods for any given analysis goal. The need for testing the utility of the synthesized data for the desired analyses is indispensable to assure valid analysis results. Testing if the generated synthetic data provides an appropriate reflection of the information in the real-world data is generally done with utility measures. These are divided into two categories: general and specific utility measures. General utility measures are aimed at evaluating if the synthesized dataset resembles the real-world data, for example if the joint distributions resemble each other. These are dominant in computer science literature. Specific utility measures are aimed at evaluating if specific analyses give the same result when applied to the synthetic and real-world data.

As follows from the congeniality principle described above, high general utility does not necessarily imply high specific utility for any type of analyses (for example Snoke et al. (2018) and Hittmeir et al. (2019)). An open question is how specific utility could be derived from general utility. This means that even if we use machine learning to generate synthetic datasets, either the technique applied should be adapted such that it is sure to be congenial with the analysis goals, or all analyses we wish to execute on the synthetic dataset need to be validated using specific utility measures with a real-world dataset, nullifying the original purpose of generating a synthetic dataset.

4.3 Intended Use

Using the principle of congeniality, we now describe in more detail the aptness of different types of generated data for different application scenarios in official statistics (Table 4.1). We also discuss how these types differ with respect to privacy risk.

4.3.1 Software testing

Generated data plays a big role in software testing, both while developing new software (systems) and while updating existing software. Depending on the complexity of the software pipeline and the need for consistent relations between input and output of different functional components, different types of generated data might be apt. Because software testing is so frequent and test data might have to be adapted based on changing requirements, choosing the most efficient option for data generation is generally advisable. For example, most unit tests can be performed with dummy data. Some might even be performed with arbitrary data, such as assessing if certain variables exist. For testing statistical software, random data can be used, where data with a specific distribution is generated to test behaviour of computed values under specific theoretical conditions.

For testing more complex software pipelines involving complex data entries and database queries, more realistic generated data could be warranted. Often, the properties of the data tested can be defined in advance based on expert knowledge and estimates of population parameter values, making simulated data an apt option for software testing. When relations between variables are very complex, and part of software testing requires ensuring that (deterministic) relations are retained by a software pipeline, synthetic data might be the only appropriate solution. Such scenarios are sometimes relevant in the statistical process, for example for complex statistical pipelines based on combined data from registers. However, the process of generating an appropriate synthetic dataset should not take more time than obtaining appropriate test data through alternative ways, for example through designing more unit-level tests based on specific test cases.

4.3.2 Education and illustrative purposes

Generated data can also be used for education and illustrative purposes, for example to let students experiment with data science techniques or to illustrate the used scientific software published alongside a paper, where public release of the corresponding microdata is not possible. For a realistic experience, at least expert knowledge or univariate distributions of variables should be used to generate this data. Hence, arbitrary, dummy and random data are not apt here. Depending on the desired analysis to illustrate, simulated data might be sufficient, or the use of synthetic data or even AI-generated synthetic data could be warranted. It should still be checked whether the generated synthetic data retains analytical validity with specific utility measures, to ensure that the goal of education about or illustration of data science techniques is reached.

4.3.3 Statistical modelling

Fitting statistical models, either to study (causal) relations in data or to produce estimates of variables of interest is a key task of statistical offices (although fitting models with the goal of predicting at an individual level most often is not). Generated data could be used for at least two purposes here: to create an extra layer of data protection by training a model on generated data instead of real data, and to enable collaboration in distributed data scenarios. For the latter scenario, a synthetic data approach can sometimes have advantages compared to other privacy preserving techniques (United Nations Committee of Experts on Big Data and Data Science for Official Statistics, New York 2023). For example, in a collaboration between Australian and Canadian hospitals, a synthetic dataset generated using CART and analysed with multivariate regression with one-way interactions yielded results consistent with a federated approach. Setting up the synthetic data experiment took 1 month, whereas setting up the federated analysis took 1.5 years (Azizi et al. 2023).

Again, congeniality is key when using generated data for predictive modelling. Arbitrary, dummy and random data are obviously not apt here. The generated data should be able to capture the information in the data needed for the nature of the prediction task at hand. Building predictive models with the aim of studying (causal) relations, or explanation, requires a different approach than building those models with the aim of estimation in a data pipeline. When releasing or publishing about models based on synthetic data in official statistics, the appropriate tests for specific utility should always be performed.

4.3.4 Public use files

The last category of applications we consider was the one originally posed by Rubin in 1993 with the introduction of the idea of synthetic data: synthetic microdata files released as public use files. From the discussions of model congeniality above it is evident that multi-purpose public use files based on synthetic data are (currently) not feasible. What about public use files released with very specific instructions, limiting the scope of analyses? A second problem, besides model congeniality, is estimation in synthetic data. Although estimations based on synthetic data might seem unbiased, p-values and standard errors estimated with a synthetic dataset can be overly optimistic and misleading. This especially happens with deep learning approaches for generating synthetic data and is difficult to correct for (Decruyenaere et al. 2024). This might decrease the possibilities for statistical offices to release general, multi-purpose synthetic datasets for inferential analysis to the public even further. Even if such datasets would be released, it would be very important to clearly narrow down the scope to that set of analyses that would result certainly in analytically valid conclusions.

4.3.5 Risk of disclosure

We have discussed congeniality and potential utility of generated data, but the other key aspect in working with generated data for statistical offices is the inherent risk of disclosure of sensitive information. For arbitrary, dummy and random data this risk is non-existent, as these methods do not incorporate information from unit-level data. For simulated data, only summary statistics need to be released and hence standard disclosure control rules can be applied. However, for synthetic data and AI-generated synthetic data, both the model used for synthesis and the synthetic dataset pose non-standard disclosure control scenarios. Generally, the more information captured by the model, the higher the disclosure risk, resulting in a privacy-utility trade-off.

When the goal is to mitigate privacy risks in a release, synthetic data does not automatically offer an improvement on the privacy-utility trade-off compared to other disclosure control methods. It is not yet possible to generalize across use cases where their performance is superior (Sarmin et al. 2025). Most risk measures studied in synthetic data again are generic risk measures, applicable to any synthetic dataset. However, for disclosure control in official statistics, specific risk measures that also take into account the context of a publication are warranted. Further, a unique aspect of disclosure control in synthetic data is that a published synthetic dataset might be erroneously perceived and used as a real-world dataset by the public (Raab et al. 2024). Concluding, the process of statistical disclosure control in synthetic data is a complex and tailored process.

5 Conclusion

We aimed to create clarity around the terminology surrounding synthetic data through the lens of official statistics in this paper. We placed all data-generation techniques within a spectrum of utility and privacy based on the level of detail at which the techniques capture information from the original data.

We deem those data generation techniques synthetic data that capture information at the unit-level in original data. The generation techniques cannot be viewed separately from the final purpose of the analyses on the synthetic data. Here, our definition is in line with, but might be slightly stricter, than for example previous definitions by UNECE or the UK office for national statistics (Burnett-Isaacs et al. 2022; Bates et al. 2019). We believe that it is important to be very specific in the definition of synthetic data, as the advantages and challenges associated with unit-level based synthetic data are unique, especially in official statistics. Further, the technique originated from the goal of replacing microdata releases to the public, and the imputation and later machine learning methods used for synthesis all are based on unit-level data. The broadening of the definition to sometimes even include dummy data evolved later, but generated data like dummy data and simulated data have existed since long before synthetic data was proposed as a concept by Rubin in 1993.

We also showed the importance of congeniality and the final goal of analysis in synthetic data, and that a multi-purpose public use file is currently unattainable for statistical offices. A central important question then is what the added value of synthetic data could be for data release and access problems, compared to other disclosure control methods or privacy preserving technologies. When sending a party a synthetic dataset they can only use for linear regression estimates with up to second order interactions, that you have specifically validated for this

purpose by fitting the regression on the original data as well, what is then the remaining added value of the synthetic data compared to sending the regression models to that party directly? Or would there perhaps be more added value in giving them indirect access to the original data through other privacy preserving techniques, to perhaps check for third order interactions or try out other models?

Acknowledgments

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands.

We thank Nynke Krol for her early contributions to this paper. We thank Ton de Waal and Sander Scholtus for their observant reviews and valuable additions on imputation and uncongeniality to this paper.

References

Azizi, Z., S. Lindner, Y. Shiba, V. Raparelli, C. M. Norris, K. Kublickiene, M. T. Herrero, A. Kautzky-Willer, P. Klimek, T. Gisinger, et al. (2023). "A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health." In: *Scientific reports* 13.1, p. 11540.

Bates, A. G., I. Špakulová, I. Dove, and A. Mealor (2019). "ONS methodology working paper series number 16 — Synthetic data pilot." In: *UK Office of National Statistics*.

Bauwelinckx, Y.-C., J. Dhaene, M. van den Heuvel, and T. Verdonck (2025). "On the causality-preservation capabilities of generative modelling." In: *Journal of Computational and Applied Mathematics* 457, p. 116312.

Bhanot, K., M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett (2021). "The problem of fairness in synthetic healthcare data." In: *Entropy* 23.9, p. 1165.

Bonabeau, E. (2002). "Agent-based modeling: Methods and techniques for simulating human systems." In: *Proceedings of the national academy of sciences* 99.suppl_3, pp. 7280–7287.

Box, G. E. and N. R. Draper (1987). *Empirical model-building and response surfaces*. Wiley.

Burnett-Isaacs, K, C Girard, K Sallier, G Raab, A Ramsden, M Slokom, C Task, M Salamh, A Wardlaw, E Ghashim, et al. (2022). "Synthetic data for official statistics: A starter guide." In: *United Nations Economic Commission for Europe, Geneva*.

Calder, J., N. García Trillos, and M. Lewicka (2022). "Lipschitz regularity of graph Laplacians on random data clouds." In: *SIAM Journal on Mathematical Analysis* 54.1, pp. 1169–1222.

Capasso, M. (2025). "Synthetic data as meaningful data. On Responsibility in data ecosystems." In: *Big Data & Society* 12.4, p. 20539517251386053. DOI: 10.1177/20539517251386053.

Chapuis, K., P. Taillandier, and A. Drogoul (2022). "Generation of synthetic populations in social simulations: a review of methods and practices." In: *Journal of Artificial Societies and Social Simulation* 25.2.

Daalmans, J. (2017). *Mass imputation for census estimation*. Statistics Netherlands.

Decruyenaere, A., H. Dehaene, P. Rabaey, C. Polet, J. Decruyenaere, S. Vansteelandt, and T. Demeester (2024). "The real deal behind the artificial appeal: inferential utility of tabular synthetic data." In: *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pp. 966–996.

Domingo-Ferrer, J., D. Sánchez, and K. Muralidhar (2025). “Statistical Disclosure Control: Moving Forward.” In: *Journal of Official Statistics* 41.3, pp. 820–826.
Drechsler, J. and A.-C. Haensch (2024). “30 years of synthetic data.” In: *Statistical Science* 39.2, pp. 221–242.
Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor (2006). “Our data, ourselves: Privacy via distributed noise generation.” In: *Annual international conference on the theory and applications of cryptographic techniques*. Springer, pp. 486–503.
Fössing, E. and J. Drechsler (2025). “Does the Synthesis Model Influence a Subsequent Prediction of the Same Type.” In: *UNECE Expert Meeting on Statistical Data Confidentiality 2025*.
Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). “Generative adversarial nets.” In: *Advances in neural information processing systems* 27.
Hittmeir, M., A. Ekelhart, and R. Mayer (2019). “Utility and privacy assessments of synthetic data for regression tasks.” In: *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5763–5772.
Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf (2012). *Statistical disclosure control*. Wiley.
Kalton, G. and D. Kasprzyk (1982). “Imputing for missing survey responses.” In: *Proceedings of the section on survey research methods, American Statistical Association*. Vol. 22. American Statistical Association Cincinnati, p. 31.
Liew, C. K., U. J. Choi, and C. J. Liew (1985). “A data distortion by probability distribution.” In: *ACM Transactions on Database Systems (TODS)* 10.3, pp. 395–411.
Little, R. J. (1993). “Statistical Analysis of Masked Data.” In: *Journal of Official Statistics* 9.2, p. 407.
Loong, B. and D. B. Rubin (2017). “Multiply-imputed synthetic data: advice to the imputer.” In: *Journal of Official Statistics* 33.4, pp. 1005–1019.
Meng, X.-L. (1994). “Multiple-imputation inferences with uncongenial sources of input.” In: *Statistical science*, pp. 538–558.
Muammar, S. and K. Shaalan (2026). “A PySpark-based KNN classification framework for detecting fake product reviews in e-commerce.” In: *Telematics and Informatics Reports* 21, p. 100275.
Murray, J. S. (2018). “Multiple Imputation: A Review of Practical and Theoretical Findings.” In: *Statistical Science* 33.2.
OpenSAFELY (2021). *Dummy data and expectations*. URL: `https://docs.opensafely.org/legacy/study-def-expectations/` (visited on 07/27/2026).
Piller, C. (2026). “Fake data from trial sites ruin studies, drug firms say.” In: *Science* 391.6781.
Raab, G. M., B. Nowok, and C. Dibben (2024). “Practical privacy metrics for synthetic data.” In: *arXiv preprint arXiv:2406.16826*.
Raghunathan, T. E. (2021). “Synthetic data.” In: *Annual review of statistics and its application* 8.1, pp. 129–140.
Rane, R. and R Subhashini (2026). “An Automated Fake News Detection Framework Using Residual Convolutional Bi-LSTM with Improved Optimisation-Based Weighted Feature Representation.” In: *Journal of Information & Knowledge Management*, p. 2550135.
Reiter, J. P. and T. E. Raghunathan (2007). “The multiple adaptations of multiple imputation.” In: *Journal of the American Statistical Association* 102.480, pp. 1462–1471.
Rubin, D. B. (1993). “Statistical disclosure limitation.” In: *Journal of official Statistics* 9.2, pp. 461–468.
Sarmin, F. J., A. R. Sarkar, Y. Wang, and N. Mohammed (2025). “Synthetic data: revisiting the privacy-utility trade-off: F. Jahan Sarmin et al.” In: *International Journal of Information Security* 24.4, p. 156.
Slokom, M., M. Larson, and A. Hanjalic (2020). “Partially synthetic data for recommender systems: Prediction performance and preference hiding.” In: *arXiv preprint arXiv:2008.03797*.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic (2018). “General and Specific Utility Measures for Synthetic Data.” In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 181.3, pp. 663–688. DOI: 10.1111/rssa.12358.

Susser, D. and J. Seeman (2024). “Critical provocations for synthetic data.” In: *Surveillance & Society* 22.4, pp. 453–459.

Sweeney, L. (2002). “k-anonymity: A model for protecting privacy.” In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05, pp. 557–570.

United Nations Committee of Experts on Big Data and Data Science for Official Statistics, New York (2023). *United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*. URL: https://unstats.un.org/bigdata.

Waal, T. de, J. Pannekoek, and S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Vol. 563. Wiley Online Library.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague, Netherlands
www.cbs.nl

Prepress
Statistics Netherlands Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information