International Journal of Population Data Science

Journal Website: www.ijpds.org

A review of synthetic data terminology for privacy preserving use cases

Lora Frayling[1], Shah Suraj Bharat[2], Elizabeth Pattinson[2], Joshua Stock[3], Fiona Lugg-Widger[3], Emma Gordon[2], and Emily Oliver[2,*]

Submission History

Submitted:	20/12/2024
Accepted:	27/06/2025
Published:	15/10/2025

[1] Health Data Insight CIC, CPC4, Capital Park, Fulbourn, Cambridge, CB21 5XE, United Kingdom

[2] Administrative Data Research UK, Economic and Social Research Council, Polaris House, North Star Avenue, Swindon SN2 1UJ, United Kingdom

[3] Centre for Trials Research, College of Biomedical & Life Sciences, Cardiff University, 7th Floor, Neuadd Meirionnydd, Heath Park, Cardiff, CF14 4YS, United Kingdom

Abstract

Synthetic data is emerging as a key area of development for supporting research that involves secure forms of administrative and health data, both in the United Kingdom and globally. In practice, key challenges in the generation and adoption of synthetic data are closely tied to the need for agreed and consistent terminology for describing it. The absence of standardised language hinders the setting of quality standards, establishment of governance and guidelines and effective sharing of knowledge and best practices. This has implications for research that uses synthetic healthcare and administrative data, particularly when such data are generated from protected personal data.

This commentary paper reviews existing literature on synthetic data to explore how key terms are currently defined in practice, with a focus on privacy-preserving use cases. Our analysis reveals that terms describing properties of synthetic data are often lacking and inconsistent, largely due to the breadth of synthetic data types, contexts and use cases. Context-specific terminology with nuanced meanings complicates efforts for the development of universally agreed definitions, particularly for privacy-preserving synthetic data that captures characteristics from protected data sources.

To address this, we propose broad definitions for key terms including *synthetic data*, *utility*, *utility measure* and *fidelity*. We conclude by offering a set of recommendations emphasising the need for consensus on terminology and encouraging clearer descriptions in future literature that specify both the intended use of the data and the measures used to describe it.

Keywords

synthetic data; administrative data; health data; statistical disclosure control; data utility; data confidentiality; privacy preservation

*Corresponding Author:
Email Address: emily.oliver@esrc.ukri.org (Emily Oliver)

https://doi.org/10.23889/ijpds.v10i2.2967

Introduction

Synthetic data is emerging as a key area of development for supporting research that involves highly sensitive administrative and healthcare data [1]. It requires generating new data that do not originate from real-world events but preserve specific properties of the real data, without retaining any identifiable information about a particular individual or allowing for reidentification through data analyses and inference. Such data can be made available for a variety of use cases without compromising privacy. This can be useful even when access to real data is possible in controlled settings. For example, it can enable researchers to efficiently explore datasets, plan their projects, collaborate with others and develop code prior to requesting access to real data. Synthetic data can also be useful in its own right, without the intention to access real data, for example in education and training.

With secure forms of data sources rapidly growing in number and size, the availability of synthetic data has the potential to improve research efficiency and enhance societal benefits, while reducing regulatory burdens on data custodians. However, several challenges still hinder the widespread adoption of synthetic data. Building on barriers identified in the deployment of synthetic data, we categorise the challenges as follows [2]:

- Capacity and infrastructure: lack of knowledge, experience, and resources to create synthetic data; lack of infrastructure for production and hosting of synthetic data resources
- Distribution and use: lack of mechanisms to provide access to synthetic data collections and support for effective use
- Standards: lack of identified synthetic data characteristics and quality standards
- Governance: lack of shared experiences, use cases, knowledge and practices to formulate policies and determine the governance models to ensure maximum benefits of synthetic data production, distribution, and use.

In addition to these, we posit that having unsettled definitions of key synthetic data terms is an underlying challenge that cuts across these categories. We understand that differing interests, knowledge, motivations and priorities across stakeholders naturally leads to divergence in conceptual definitions. On one hand, data scientists and researchers may be more focused on the structure and privacy-preserving properties of synthetic data that are useful for analysis and training machine learning (ML) algorithms. On the other hand, policy makers and the public may be more concerned about balancing benefits and risks, using data for informing policy decisions and ensuring that privacy-protecting features are described using accessible language. Consequently, the lack of a shared language and understanding amongst different stakeholders, such as data owners, processors and providers and users, makes it difficult to coordinate and collaborate.

We aim to understand the landscape of terminology used around synthetic data and seek clarity on definitions of commonly used terms that can help to address existing challenges. We have approached this through our analysis of the literature, homing in on the use of four key terms: *synthetic data*, *utility*, *utility measure*, and *fidelity*. To add value and reach consensus, we offer definitions for these terms that are relevant both in the context of privacy preservation and across other domains.

Methodology note

To identify key terms that are related to synthetic data and understand how they are defined across literature, we surveyed a range of academic and public-facing grey literature, which totalled 44 peer-reviewed journal articles and 112 public-facing items. Academic literature was found using the Cardiff University eLibrary and grey literature was found using Google's Advanced search function, with initial searches for the term 'synthetic data' followed by searches for other key terms. Further details of the search criteria and key findings are set out in detail in our Supplementary file.

The academic sources found typically discuss methods for the generation and evaluation of synthetic data generation. In contrast, public-facing articles are often produced by technology companies offering synthetic data services and promoting awareness or by researchers sharing coding guidance or research findings, however, these are not generally peer-reviewed. While a wider literature review was beyond our scope, we considered additional publications about data synthesis projects, initiatives and solutions to further understand the current state of privacy-preserving synthetic data, as cited throughout this paper. We applied a principled approach to selecting publications that shine light on the need for:

- A shared understanding of synthetic data scope and purpose
- Methods for synthetic data characterisation and evaluation for a given set of objectives, e.g., privacy protection, accessibility and usefulness for its intended purpose
- Clarity on how quality control of synthetic data needs to be managed in relation to the existing data protection regulations, such as GDPR
- Effective guidelines for practitioners involved in generating, distributing or using synthetic data [3].

Our final proposed definitions for terms related to synthetic data were reviewed and evaluated by members of ADR UK's Synthetic Data Working Group, a panel of experts convened to ensure accuracy, clarity and alignment with current standards in the field (see 'Acknowledgements' for further explanation). The Working Group also advised on further literature and had the remit to ensure the definitions were both accessible to non-specialists and technically precise, balancing the need for usability with domain-specific rigour.

What is synthetic data?

To understand the core concepts and principles, it is essential to address the key terms that occur in the literature, starting

from the definition of synthetic data and differentiating between its application for privacy preservation and other purposes.

Currently there is no universally accepted definition of synthetic data. In both the academic and public-facing literature, we find it to be underdefined, with the author often assuming a tacit understanding by the reader or only describing the type of synthetic data being discussed. This leads to highly individualised definitions with specific and nuanced meanings in accordance with context, users and audiences.

Historical overview

In 2022, the Alan Turing Institute and Royal Statistical Society proposed the following definition:

> "Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)". [4]

This definition leaves open the question of where the mathematical model comes from. It can be a model derived from real data or fully constructed by mathematical and computational means.

This approach to defining synthetic data illustrates ambiguity in the broader literature, where synthetic data is often not explicitly defined, but instead, the process for creating it within a specific context is specified. In instances where synthetic data is described, it is defined more narrowly, typically with a reference to properties of a real-world dataset that is retained and the privacy level it protects. For example, Arnold and Neunhoeffer (2020) define synthetic data as "an artificial copy of the original dataset that ideally shares the same statistical properties" [5]. While this makes sense within a specific context such as privacy preservation, this definition does not differentiate other types of synthetic data that are not based on real data.

Furthermore, the earliest references to synthetic data in the literature date back to the 1940s, specifically in the context of Monte Carlo simulation models. These models generate data in the absence of real datasets to simulate real-world phenomena using algorithms to repeatedly sample from probability distributions, often based on mathematical or theoretical frameworks that represent real-world processes [6]. Similarly in recent years, generative artificial intelligence (AI), such as Large Language Models (LLMs), that are trained on vast amounts of data, can be repurposed to generate synthetic data for Machine Learning (ML) that does not necessarily look like a real dataset [7].

Delanius and Liew introduced synthetic data as an alternative to real data in their early development of data anonymisation techniques (Delanius, 1977) and methods for data distortion (Liew, 1985) [8, 9]. Meta-analyses of synthetic data trace the term to D. B. Rubin's 1993 seminal paper "Discussion: Statistical disclosure limitation" [10–14]. In the paper Rubin proposed "the idea to generate synthetic data as a tool for broadening access to sensitive microdata" and the release of synthetic data instead of real-world data. He suggested using methods based in multiple imputation strategies as outlined in his earlier work, which also discussed innovative disclosure protection strategies [15, 16]. In the same year, R. J. A. Little discussed the data quality limitations of Rubin's fully synthetic data approach, and proposed partially synthetic data as an alternative, where only some data items are replaced with synthetic values [17]. Replaced data items can be those used for direct identification or sensitive items that should not be disclosed.

One of the best-known early applications of synthetic data is from 1997, when the US Federal Reserve Board replaced monetary values at high risk of disclosure in the Survey of Consumer Finances with synthetic values. In the UK, a well-known application of synthetic data is its adoption to facilitate access to the Scottish Longitudinal Study. This application links census data with other sensitive information from health records and death registers. Researchers can request bespoke synthetic datasets that were tailor-made to the research questions that they were trying to answer. This enables them to develop analysis that could then be run on the real data. The R package 'Synthpop', a widely used tool for generating synthetic datasets, was developed as part of this project [18–20].

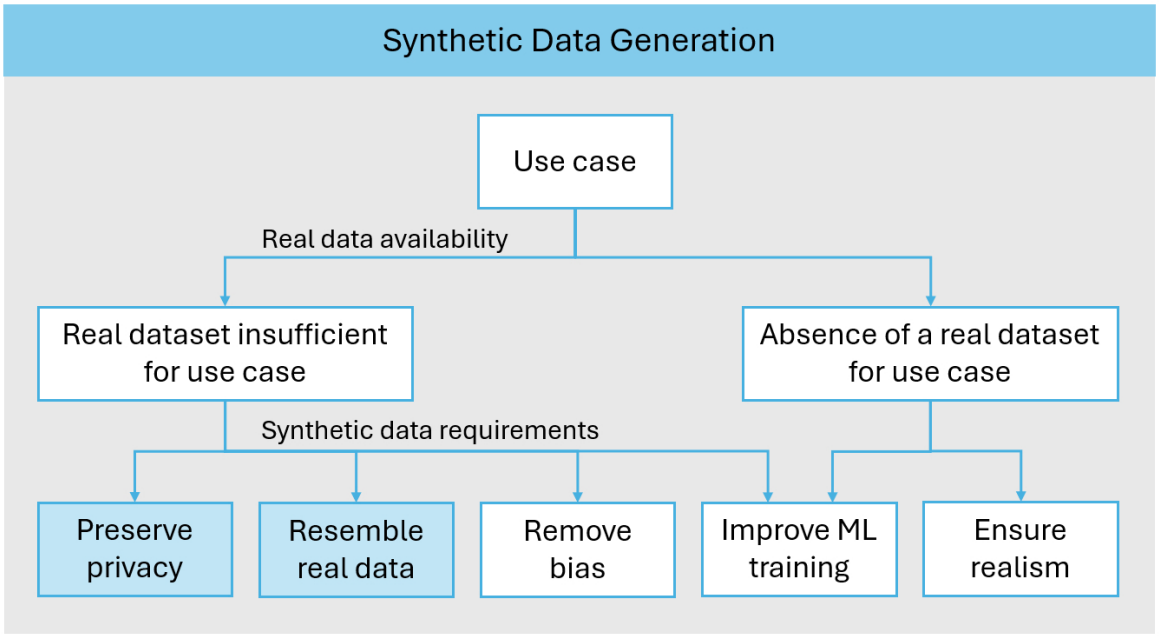
Drechsler and Haensch also note that while these methods were initially developed by the statistical community, in parallel, the computer science community developed synthetic data to provide training datasets for ML models, acknowledging its potential for mitigating disclosure risks [21]. Both approaches often have the aim to achieve the same answers for analyses that are run on both the real and synthetic datasets. However, there is also a broader set of use cases such as de-biasing data and data augmentation for ML [4]. De-biasing aims to remove biases that are found within the data, often by rebalancing minority groups in the data, which can then be used to train ML models. Data augmentation involves creating synthetic data that can be used to train ML models and gain higher accuracy. Across these examples, synthetic data is created to be highly useful for a specific case where the real data is insufficient.

Definition scope

Given the breadth of types of synthetic data, for any meaningful discussion it is helpful to clarify the type being discussed. Figure 1 provides a visualisation of the synthetic data types highlighting the importance of (a) whether the synthetic data has been generated based on an existing real dataset and (b) the desired properties of the synthetic data that ensure its usefulness for its intended purpose. Both the existence of real data and the intended purpose guide the choice of the generative model. While accessibility and privacy might seem implicit to usefulness, the chosen data generation approach ultimately determines the properties of the synthetic data and whether it fulfils the requirements of the use case.

In the case of personal data, the goal is to generate a dataset that preserves properties of the real data that are relevant for the intended use case, while ensuring no identifiable information remains. This is commonly referred to as privacy-preserving synthetic data in the literature. As previously mentioned, other applications of synthetic data focus on objectives such as removing bias from real data or augmenting real datasets for training ML models. There can be multiple aims but often they are competing, e.g.,

Figure 1: Diagram that illustrates the different types of synthetic data and key considerations for choosing a synthetic data generation approach, specifically: (1) the intended use case, (2) the availability and adequacy of a real dataset and (3) the required properties of the synthetic data to meet the use case needs. When a real dataset is available but insufficient due to privacy constraints, privacy-preserving synthetic data can be generated to resemble the real data while ensuring individual privacy is protected (highlighted in blue)



preserving fidelity can often conflict with privacy or removing bias [4]. Furthermore, we acknowledge that partially synthetic data fits as another category here, as it is based on a real dataset.

This contrasts with synthetic data which are generated independently of existing datasets that aim to model real-world phenomena or provide diverse training data for ML models. Its usefulness may depend on how well it represents real-world scenarios, as in the use cases of real datasets, or on its ability to enhance training of ML models by providing numerous, realistic and varied samples.

More generally, synthetic data types can be diverse and include tabular (static or time series), image, audio, and multi-modal formats. Applications for synthetic data are equally wide-ranging across fields, such as data science, healthcare, economics, engineering and autonomous systems, as seen across the reviewed academic literature. While the context of privacy-preserving synthetic data is generally well described in academic publications, the term *synthetic data* is rarely explicitly defined. Literature concerning other types of synthetic data often require deeper reading to determine whether they are based on a real dataset and what the aim of the data synthesis is: two critical considerations when considering a generative model and assessing privacy risks.

Utility, utility measures and fidelity

Given a wide range of potential use cases for synthetic data, it is important to consider use case objectives and define key properties of synthetic data. Utility, fidelity and privacy levels are often discussed in a context of privacy preservation. Ideally, these should be translated into measurable metrics that can guide practice. Through our literature review, we found that, as with the term synthetic data, utility and fidelity are often undefined or defined inconsistently.

Utility and utility measures

Data utility is often used to describe how useful a dataset is within a specific usage context, for instance when it is used to inform real-world processes and decisions, while also being framed more broadly in terms of its 'benefit to society' [22]. Taking this definition, data owners can create synthetic data for a specific use case and evaluate its utility, considering both the inherent qualities of the synthetic data and external factors that affect how it can be used.

However, there are nuances beyond this broad definition. For real data, quality and utility are often described as multi-dimensional and cover aspects such as accuracy, completeness, metadata and documentation and ease of access [23, 24]. As such, we expect that synthetic data utility can also be multi-dimensional.

Differences between definitions of utility arise in literature when there are implicit assumptions about the use case. For example, the utility of anonymised data is typically defined as its ability to be used in the place of original data and achieve the same results for analysis, its implicit use case [25]. Similarly, for synthetic data, utility has been described as "the degree to which the synthetic data effectively mimics the original data for analytical or modelling purposes", which also links directly to the use case [26]. Both scenarios typically call for measuring utility by looking at similarity with the real data.

More generalised synthetic data papers acknowledge the broader range of synthetic data use cases [4, 27]. Consequently, the utility of synthetic data may be measured differently to the utility of real or anonymised data, dependent on the intended use case. For example, privacy-preserving synthetic data can be used for educational training, data exploration to determine suitability of real data for research, understanding the structure of the real data, writing programming code for analyses of real data and, in rarer scenarios, serving as a substitute for real data analysis [28]. For each of these, the utility of the synthetic data depends on multiple data properties and will require a specific type of generative model to achieve these properties.

Our literature review revealed that the utility of synthetic data is typically evaluated using various *utility measures*, which are broadly grouped into two types: *general utility measures* and *specific utility measures*. In the context of privacy-preserving synthetic data, general utility typically focuses on the degree to which the synthetic data resembles the real data in terms of general statistical characteristics, whereas specific utility compares the performance of the synthetic and real data in particular data science tasks [29]. When generalising to broader use cases and considering other dimensions of utility, we propose the following classification: (1) *general utility measures*, which quantify how useful the synthetic data is likely to be for a broad range of expected use cases; and (2) *specific utility measures*, which quantify how well the synthetic data can be used for a specific use case.

Fidelity

The closeness in resemblance of synthetic data to real data is typically referred to as its *fidelity*, however terms such as *resemblance* and *quality* are also used as synonyms. Since general utility measures are typically based on fidelity, the terms *utility* and *fidelity* are often used interchangeably. However, the relationship between them is not always straightforward since fidelity, as well as utility, can be multi-dimensional. For example, synthetic data may adhere to the real data structure, i.e., structural fidelity, or statistical properties, i.e., statistical fidelity. Statistical properties also capture inherent qualities of the real data such as completeness, coverage and timeliness. Different use cases may require fidelity of different dimensions.

People often try to reason about whether data is *low fidelity* or *high fidelity* to help guide them in knowing whether the data is suitable for their use cases. In the literature, we found that high fidelity is often defined, but low fidelity is not, thereby introducing an opportunity for confusion and assumed boundaries for these terms. In addition, the multi-dimensional aspect of fidelity means it is difficult to categorise. For example, data may have high structural fidelity but low statistical fidelity. The Office for National Statistics (ONS) addresses this multi-dimensional problem with a categorisation of synthetic data types based on structural and statistical fidelity and associates them with the data's analytical value [30]. However, this still does not capture the full complexity of the many dimensions that impact analytical value for different applications.

It is important to note that different dimensions of fidelity are highly important for some use cases and less so for others. When fidelity is important for utility, increasing fidelity past a certain point will not always increase the utility further. Thus, we emphasise that utility and fidelity are separate but often related concepts, with the former depending on context and the latter being an inherent quality of the synthetic data. In the literature, this is not always fully clarified and the two are used interchangeably.

When synthetic data is based on real personal data, the goal is typically to ensure that it contains no identifiable information. This can require balancing a privacy guarantee against fidelity and striking the right privacy-fidelity trade-off. Some literature misleadingly refers to a privacy-utility trade-off for synthetic data, assuming implicitly that the utility is directly tied to fidelity. As discussed, utility is not always closely linked to fidelity. There are use cases where utility can remain unaffected while the privacy measures are tightened and fidelity decreases [4].

Other terminology

In the previous section we outline key concepts, their distinction and their multi-dimensional nature. We point out the need to avoid misunderstandings to enable proper evaluation and discussion of synthetic data. However, in addition to the outlined terminology, there are other terms and concepts that remain ambiguous, as found in our review of the academic and public-facing literature.

Across the academic literature, which largely focuses on methods for generation and evaluation of privacy-preserving synthetic data, we found that domain-specific terms are often undefined. For example, papers often discuss *random noise* in the context of *differential privacy*, both of which are statistical concepts relevant for synthetic data generation [31]. While random noise is a strongly understood concept within the domain of statistics, the absence of an explicit definition may hinder the understanding of readers from other disciplines and therefore limit discussions about synthetic data generation methods. Similarly, the term *microdata*, which often describes datasets relating to individuals but may also refer to other entities such as households or organisations, is often used without precise definition, leading to the possibility of confusion over whether the microdata constitutes personal data.

In the public-facing grey literature, there are key terms that are not often explicitly defined, such as *microdata*, *metadata*, *big data*, *artificial intelligence (AI)* and *machine learning (ML)*. There are also a range of terms used interchangeably with *synthetic data*, such as *mock*, *dummy*, *fake*, *artificial* and *simulated data*, often without precise definitions and with conflicting meanings across sources. Lack of clear definitions and inconsistent use of language may mislead members of the public who are learning about synthetic data from the internet. Paradoxically, specialised terms are often more clearly and regularly defined, such as *partial synthetic data*, *differential privacy*, *Bayesian Networks* and *general adversarial networks (GANs)*, perhaps due to less of an expectation that the reader has relevant expertise. As such, a focus on agreed definitions for generally but imprecisely used terminology may yield the most benefit to mutual understanding and clarity among stakeholders.

Table 1: Proposed definitions for terminology related to synthetic data

	Proposed definitions
Synthetic data	
Synthetic data	Data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s). [4]
Privacy-preserving synthetic data	Synthetic data that preserve specific properties of a real dataset without retaining any personally identifiable information.
Fidelity (when synthetic data is based on real data)	
Fidelity	How closely the synthetic data resembles real data overall in terms of its characteristics and properties.
Structural fidelity	How closely the structure of synthetic data matches the structure of the real data.
Statistical fidelity	How closely the statistical properties of the synthetic data resemble those of the real data.
Utility	
Utility	How useful the synthetic data are for a given use case or set of use cases.
Utility measure	A way to measure how useful synthetic data are.
General utility measure	A utility measure that quantifies how useful the synthetic data are for a broad set of common use cases. Note: This is often a measure of statistical fidelity for privacy-preserving use cases.
Specific utility measure	A utility measure that quantifies how useful the synthetic data are for a specific use case. Note: This is often a measure of the similarity of outcomes of a specific data science task when using the synthetic data compared to real data.

Proposed definitions

To help build consensus around the ambiguously used terms discussed in this commentary, we propose definitions that are intended for use in the context of privacy preservation, as set out in Table 1, while remaining generalisable to other contexts. We focus on terms directly related to synthetic data, fidelity and utility, as these are central to discussions on synthetic data methods and evaluation. The definition of other key terms is left out for future consideration. As stated in our Methodology Note, these proposed definitions were reviewed and evaluated by the Synthetic Data Glossary Working Group. We acknowledge that the proposed definitions may not yet be acceptable to the general public since they have not been widely tested, and that further work is needed to ensure public understanding.

Conclusion and recommendations

Inconsistent use of terminology can lead to miscommunication and misunderstandings among key stakeholders such as data providers, data users, policymakers and the public. This may be especially problematic in the context of synthetic data, where lack of clarity may result in misalignment of expectations around utility, privacy and governance. Defining terminology goes beyond providing clarity: it plays an important role in how technologies are developed, how policy and legal agendas evolve and the direction of public discourse, particularly when related to the safeguarding of personal data, a heavily debated topic.

In the absence of agreed definitions surrounding this topic, we have initiated the development of a *non-static glossary* of key synthetic data terminology intended for use in the context of privacy preservation, but generalisable to other contexts. By *non-static*, we propose a resource designed to evolve over time, rather than one that enforces fixed terminology, definitively resolves lexical debates or argues that all successive work should adopt our proposed definitions. Instead, this glossary aims to provide greater coherence in the current discourse, recognise where terms remain contested or vague and promote greater clarity and precision in the language used across the field.

To move this discussion forward, we provide recommendations to key stakeholders in this field, including data owners, synthetic data creators, providers and researchers in academia and industry who work with and write about synthetic data. The recommendations are as follows:

- Agree on our suggested definition of synthetic data, as originally proposed by the Alan Turing Institute, while acknowledging the breadth of synthetic data types based on whether they are derived from real data and their intended utility. Privacy-preserving synthetic data should be acknowledged as a distinct type.
- Clearly state the type of synthetic data being discussed in any literature or discussions about synthetic data
- Clearly define utility measures with respect to the intended use cases of the synthetic data
- Develop frameworks for evaluating fidelity and utility that capture their multiple dimensions and provide a clear way to link the two across use cases
- Avoid the use of terminology that is ambiguously defined, such as *dummy*, *mock*, *fake, artificial* or *simulated* data, unless clearly contextualised
- Agree on the usage and definition of general terms, such as *microdata*, and encourage explicit definition of

domain-specific terms, such as *random noise,* when they are used

- Undertake further research to explore how to ensure that terminology for synthetic data is acceptable to a range of users and stakeholders, including the public.

Acknowledgements

The authors are grateful for the advice and guidance received from members of the Glossary Sub-Group of the Synthetic Data Working Group. The membership of this group grew from a previous working group chaired by HDR UK to include additional collaborators from across the ADR UK partnership including synthetic data award holders. Names are available on request.

Statement of conflicts of interest

None

Ethics statement

Ethics approval was not sought as this work was based entirely on literature searches.

Data availability statement

Data relating to this commentary is to be found in the Supplementary File entitled: A Review of Synthetic Data Terminology for Privacy Preserving Use Cases.

References

1. Kokosi T, De Stavola B, Mitra R, Frayling L, Doherty A, Dove I, Sonnenberg P, Harron K. An overview of synthetic administrative data for research. International Journal of Population Data Science. 2022;7(1). https://doi.org/10.23889/ijpds.v7i1.1727

2. Arthur L, Costello J, Hardy J, O'Brien W, Rea J, Rees G, et al. On the Challenges of Deploying Privacy-Preserving Synthetic Data in the Enterprise [Internet]. In: Proceedings of the 1st Workshop on Challenges in Deployable Generative AI; 2023 Jul [cited 2025 Apr 18]; [about 2 p.]. Available from: https://doi.org/10.48550/arXiv.2307.04208

3. Beduschi A. Synthetic data protection: Towards a paradigm change in data regulation? *Big Data & Society*. 2024;11(1). https://doi.org/10.1177/20539517241231277

4. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, Cohen SN, Weller A. Synthetic Data–what, why and how? ArXiv. 2022 May 6. Available from: https://arxiv.org/abs/2205.03257

5. Arnold C, Neunhoeffer M. Really useful synthetic data: A framework to evaluate the quality of differentially private synthetic data. arXiv. 2020 Apr 16. Available from: https://arxiv.org/abs/2004.07740.

6. Metropolis N, Ulam S. The Monte Carlo method. Journal of the American Statistical Association. 1949; 44(247):335-341. https://doi.org/10.1080/01621459.1949.10483310

7. Lin L, Wang R, Xiao R, Zhao J, Ding X, Chen G, Wang H. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In: Findings of the Association for Computational Linguistics: ACL 2024; 2024 Jul 23-28; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics; 2024. p. 11065-11082. Available from: https://arxiv.org/abs/2406.15126.

8. Dalenius T. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*. 1977;15:429–444.

9. Liew CK, Choi UJ, Liew CJ. A data distortion by probability distribution. *ACM Trans Database Syst*. 1985;10(3):395–411. https://doi.org/10.1145/3979.4017

10. Drechsler J, Haensch AC. 30 years of synthetic data. *Statist Sci*. 2024 May;39(2):221–42. https://doi.org/10.1214/24-STS927

11. Drechsler J, Bender S, Rässler S. Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Trans Data Privacy*. 2008;1(3):105–30.

12. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*. 2003;19:1–16.

13. Reiter JP. Data confidentiality. *WIREs Comput Stat*. 2011 Sep/Oct;3(5):450–6. https://doi.org/10.1002/wics.174

14. Rubin DB. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*. 1993;9:462–8.

15. Rubin DB. Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. In: Proceedings of the Survey Research Methods Section of the American Statistical Association. Vol. 1. Alexandria, VA: American Statistical Association; 1978. p. 20–34.

16. Rubin DB. Multiple imputation for nonresponse in surveys. *Wiley Series in Probability and Statistics*. 1987 Jun 9. https://doi.org/10.1002/9780470316696

17. Little RJ. Statistical analysis of masked data. *Journal of Official Statistics*. 1993;9:407–26.

18. Nowok B, Raab GM, Dibben C. synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control [Internet]. Available from: https://cran.r-project.org/web/packages/synthpop/index.html.

19. Nowok B, Raab GM, Dibben C. Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. Stat J IAOS. 2017;33(3):785–96. https://doi.org/10.3233/SJI-150153

20. Dibben C, Raab GM, Nowok B, Williamson L, Adair L. Synthpop: A tool to enable more flexible use of sensitive data within the Scottish Longitudinal Study. In: Drechsler J, editor. Handbook of Sharing Confidential Data: Differential Privacy, Secure Multiparty Computation, and Synthetic Data. S.l.: Chapman and Hall/CRC; 2024.

21. Drechsler J, Haensch AC. 30 years of synthetic data. *Statist Sci*. 2024 May;39(2):221–42. https://doi.org/10.1214/24-STS927

22. Woo M-J, Reiter JP, Oganian A, Karr AF. Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. JPC [Internet]. 2009 Apr. 1 [cited 2024 Dec. 12];1(1). https://doi.org/10.29012/jpc.v1i1.568

23. Hassenstein MJ, Vanella P. Data quality—Concepts and problems. *Encyclopedia*. 2022;2(1):498-510. https://doi.org/10.3390/encyclopedia2010032. https://doi.org/10.1136/bmjhci-2020-100303

24. Gordon B, Barrett J, Fennessy C, Cake C, Milward A, Irwin C, Jones M, Sebire N. Development of a data utility framework to support effective health data curation. *BMJ Health Care Inform*. 2021;28:e100303. https://doi.org/10.1186/s12911-024-02427-0

25. Woo M-J, Reiter JP, Oganian A, Karr AF. Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. JPC [Internet]. 2009 Apr. 1 [cited 2024 Dec. 12];1(1). https://doi.org/10.29012/jpc.v1i1.568

26. Isasa I, Hernandez M, Epelde G, et al. Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis. *BMC Med Inform Decis Mak*. 2024;24:27. https://doi.org/10.1186/s12911-024-02427-0

27. United Nations Economic Commission for Europe. Synthetic data for official statistics: A starter guide. Geneva: United Nations Economic Commission for Europe; 2022. Available at: Synthetic Data for Official Statistics - A Starter Guide | UNECE.

28. Lloyd M, Lee S, Choi J, et al. Synthetic data in health care: A narrative review. *PLOS Digital Health*. 2023;2(1):e0000082. https://doi.org/10.1371/journal.pdig.0000082

29. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. Journal of the Royal Statistical Society Series A: Statistics in Society. 2018 Jun;181(3):663-88. https://doi.org/10.48550/arXiv.1604.06651

30. Office for National Statistics. Synthetic data pilot. *ONS Methodology Working Paper Series 16*. Available from: ONS methodology working paper series number 16 - Synthetic data pilot - Office for National Statistics.

31. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: Privacy via distributed noise generation. Adv Cryptology - Eurocrypt 2006, Proc. 2006;4004:486-503. https://doi.org/10.1007/11761679_29