

Supplementary File:

A Review of Synthetic Data Terminology for Privacy Preserving Use Cases

Authors: Shah Suraj Bharat, Lora Frayling, Elizabeth Pattinson, Joshua Stock, Fiona Lugg-Widger, Emma Gordon, Emily Oliver

Summary

This is the Supplementary File that supports the commentary paper, “A Review of Synthetic Data Terminology for Privacy Preserving Use Cases”. It describes the methodology and some key findings from the initial work that surveyed a range of academic and public-facing grey literature to identify consistent and contested terminology. The methodology for this work and reference list is detailed in this file.

Contents

Summary	1
Academic Literature Search	2
Search Methodology	2
Identification of Key Terminology	3
Definition Analysis for Synthetic Data and Utility	4
Grey Literature Search.....	6
Search Methodology	6
Identification of Key Terminology	7
Definition Analysis for Secondary Search Terms	7
Academic Literature Reference List.....	9
Grey Literature Reference List.....	12

Academic Literature Search

Search Methodology

The academic literature search was conducted using the Cardiff University eLibrary system. Within this system, the 'articles' filter (i.e., journal papers) was applied to all searches to ensure that results focused on peer-reviewed academic sources. In cases where the number of results under the 'articles' filter was limited, the 'books' filter was applied to expand the search scope.

Initial Search Term: Synthetic Data (20 reviewed sources)

The first two pages of search results were reviewed, and key terminology relevant to synthetic data was extracted at the discretion of the reviewer. Terminology was recorded in an Excel spreadsheet for further analysis. Articles that were adjacent to synthetic data but had an alternative core focus to privacy-preserving synthetic data use cases (e.g., coding in Python, seismographic models) were excluded from further review. After the initial review, the reference lists of selected key articles were also examined to identify additional relevant sources. These articles were considered to provide a more comprehensive overview of the topic, such as:

- Statistical Disclosure Limitation by Rubin (1993)
- Synthetic Data by Rangunathan (2020)
- Comparative Study of Differentially Private Synthetic Data Algorithms by Bowen & Snoke (2021)

Secondary Search Terms (24 reviewed sources)

Following the initial search, additional search terms were included in a second search based on their frequency of occurrence in the initially reviewed literature. To broaden the scope, each of these secondary terms was searched in conjunction with "synthetic data" to ensure relevance to the topic. The first two items of results for each search were reviewed, excluding articles with a focus outside synthetic data.

The secondary search terms included:

- Full Synthesis
- Partial Synthesis
- High Fidelity
- Low Fidelity
- Multiple Imputation
- Differential Privacy
- Utility
- Microdata
- Monte Carlo Simulation
- Bayesian Networks
- Metadata
- Big Data
- Generative Adversarial Networks (GANs)
- Artificial Intelligence

A full list of reviewed sources is provided in the 'Academic Literature Reference List' section.

Identification of Key Terminology

Key terms of interest across all 44 reviewed articles were identified and were defined as any technical language, words, or phrases that seemed vague or not immediately apparent to a lay audience. These terms were compiled into a single column list in Excel and sorted by their frequency of mention. The top 10 terms were then elaborated in a separate document for further analysis and explanation. These are displayed in Table 1.

Table 1: Top 10 terms from academic literature search

Keyword	Occurrence % in literature (of 44 reviewed)
<i>(Random) Noise</i>	22.9%
<i>Generative adversarial network (GANS)</i>	20.8%
<i>Utility</i>	20.8%
<i>Multiple Imputation</i>	16.7%
<i>Disclosure (risk)</i>	16.7%
<i>Full/Partial Synthesis</i>	14.6%
<i>Machine Learning</i>	12.5%
<i>Micro data</i>	12.5%
<i>Statistical disclosure</i>	12.5%
<i>Re-identification</i>	10.4%
<i>Bayesian Networks</i>	8.3%
<i>Monte Carlo (Simulation)</i>	8.3%
<i>Differential Privacy</i>	6.25%

Some key findings included for the following terms:

- Noise
 - This is often discussed in the context of differentially private generation methods, however, often lacks concrete definitions across articles, assuming prior statistical knowledge of the meaning
 - Furthermore, despite the type of “noise”, e.g., Gaussian or exponentially distributed, being important in the context of specific differential private methods, the different types are not often discussed in these articles.
- Microdata
 - This is often discussed in the context of statistical disclosure control and privacy-preserving synthetic data without exact definition. It is not always clear if this refers to the data of individuals and whether this would constitute personal data.

Definition Analysis for Synthetic Data and Utility

An in-depth analysis was conducted to examine how the terms *synthetic data* and *utility* are defined and used across journal articles. Articles were systematically reviewed to determine:

- Whether the term is mentioned
- Whether it is explicitly defined
- Whether sufficient context is provided to imply a definition
- Whether no explicit definition or sufficient context is provided

A summary of this is in Table 2.

Table 2: Analysis of how 'synthetic data' and 'utility' are defined in academic literature

Term	Mentioned in source	Explicit definition	Sufficient context to imply definition	No definition or context
Synthetic data	38/41	6/41	30/41	2/41
Utility	21/41	3/41	8/41	10/41

An example of definitions:

- Explicit definition of synthetic data:
"synthetic data: an artificial copy of the original data-set that ideally shares the same statistical properties"
(Really Useful Synthetic Data: A Framework to Evaluate the Quality of Differentially Private Synthetic Data, Arnold & Neunhoffer 2021)
- Contextually implied definition:
"Synthetic data, which retains the essential features of the actual data with some or all of the values replaced by simulations from probability distributions, can be a way of overcoming these difficulties"

(Practical data synthesis for large samples, Rabb et al 2016)

For articles where these terms were mentioned with explicit definition or context, the context in which these were discussed were categorised and summarised below. Three articles couldn't be accessed in this follow-up review of papers, so only 41 were reviewed for these definitions

Synthetic data:

- No mention of 'synthetic data': 3/41
- Relationship to real data:
 - *Partially synthetic data*: 2/41
 - *Based on real data*: 30/41
 - *Not based on real data*: 6/41

- Aims of synthetic data for 30 sources based on real data:
 - 16/30 describe it as being artificially generated to preserve the statistical properties of real data.
 - 12/30 discuss synthetic data as preserving statistical properties while ensuring privacy.
 - 2/30 discuss its use to improve the quality of real data or augment datasets for machine learning tasks.

Utility:

- Of the 41 articles, 21 mentioning *utility*:
 - Explicit definitions (3 articles):
 - Utility as beneficial to society.
 - Utility as the similarity between synthetic and real data.
 - Utility as usefulness and similarity to real data.
 - Sufficient Context (8 articles):
 - Often focused on specific use cases, utility measures, or metrics for evaluation.
 - There remaining 10 articles had no definitions or relevant context.

Overall summary:

- Explicit definitions for synthetic data are rare, however most articles provide sufficient context to infer a definition.
- Articles highlight synthetic data types that are based on real data or no data and discuss the statistical and privacy-preserving aims of synthetic data, though exact definitions vary and may be contradictory.
- Utility is less frequently mentioned in synthetic data-related articles, with even fewer explicit definitions.
- Contextual discussions about utility typically focus on utility measures, specific use cases, and types of utility.

Grey Literature Search

Search Methodology

The grey literature search was conducted using Google's Advanced search function, to explore key terminology used in the context of synthetic data across other public-facing material. For all searches, the language was restricted to English with the instruction to look for the defined search terms "allintext". There were no restrictions set on the type of sources to be returned at the outset to maintain a broad scope.

Initial Search Term: Synthetic data (5 reviewed sources)

A preliminary web search using Google's Advanced search function for 'synthetic data', 'simulated data', 'masked data', or 'mock data' with no further restrictions was run. Following this, five sources from this search were selected and were reviewed in depth to collect and identify the terms, definitions, phrases not immediately apparent to a lay audience. These were selected based on their relevance to the overall topic of synthetic data and the depth of content they provided on the subject.

Secondary Search Terms: (112 reviewed sources)

Following the initial search, additional search terms were identified based on their frequency of occurrence in the reviewed literature. Terms were compared between the academic and grey literature, and minor adjustments were made for the appropriateness of the search. For example, the reviewer considered 'Monte Carlo Simulation' too narrow a term to include in a general literature search but felt it was important to search for 'usefulness,' as this term was more likely to appear in non-academic literature.

To broaden the scope, each of these secondary terms was searched in conjunction with "synthetic data" to ensure relevance to the topic. Approximately, the top 20 search results for each search were reviewed. Web pages requiring sign-up/paywalls or containing incoherent sources were then excluded, along with academic articles, online discussion forums and results that repeated sources from previous searches.

The secondary search terms included:

- Synthesis (Synthesis or Full Synthesis or Partial Synthesis)
- Mock data
- Dummy data
- Fidelity (High Fidelity or Low Fidelity)
- Differential Privacy
- Utility
- Usefulness
- Micro data
- Bayesian Networks
- Metadata
- Big Data
- Generative Adversarial Networks (GANs)
- Artificial Intelligence Machine Learning (AIML)

A full list of reviewed sources are provided in the 'Grey Literature Reference List' section.

Identification of Key Terminology

The generative AI tool [PopAi](#) was specifically asked "what are the key words and definitions" from the main body of texts in all literature results. This approach was used to maximize the scope of the reviewed literature, in the time available. Where the tool picked out irrelevant text (e.g., definitions from advertisements) these were removed. The identified keywords were then ranked based on their frequency of appearance across literature and summarised in Table 3.

Table 3: Top search terms from grey literature search

Keyword (other than synthetic data)	Occurrence % in literature (of 112 reviewed)
<i>Generative Adversarial Networks (GANs)</i>	25.89%
<i>Differential Privacy</i>	19.64%
<i>Synthetic Data Generation</i>	16.96%
<i>Utility</i>	16.07%
<i>Machine Learning</i>	11.61%
<i>Privacy</i>	10.71%
<i>Deep Learning</i>	8.04%
<i>Fidelity</i>	8.05%
<i>Synthetic Data vault</i>	8.05%
<i>Artificial Intelligence</i>	7.14%
<i>Anonymisation</i>	7.14%
<i>Bayesian Networks</i>	6.25%

Definition Analysis for Secondary Search Terms

For each webpage identified using secondary search terms, the term (along with specific related terms, where relevant) was examined using the Ctrl+F function. All mentions of the term were reviewed to determine whether it was:

- Mentioned in the source
- Explicitly defined
- Not explicitly defined but accompanied by sufficient context to imply a definition to the reader
- Neither explicitly defined nor accompanied by sufficient context

For example, sources discussing "high fidelity" do not always directly define terms like "fidelity" or "high fidelity." However, they use these terms in the context of how closely synthetic data resembles real data, allowing readers to infer a definition from the surrounding context. A summary of this can be found in Table 4.

Table 4: Summary of how top terminology are mentioned in grey literature

Search term group	No. sources	Term	Mentioned in source	Explicit definition	Sufficient context to imply definition	No definition or sufficient context
<i>Synthetic data / Synthesis</i>	11	<i>Synthetic data</i>	11	9	2	0
		<i>Synthesis</i>	11	2	1	8
		<i>Full synthesis / Fully synthetic data</i>	6	4	2	0
		<i>Partial synthesis / Partially synthetic data</i>	6	6	0	0
<i>Mock data</i>	8	<i>Mock data</i>	8	1	0	7
<i>Dummy data</i>	7	<i>Dummy data</i>	7	2	2	3
<i>Fidelity</i>	11	<i>Fidelity</i>	11	6	1	3
		<i>Low fidelity</i>	4	0	2	2
		<i>High fidelity</i>	7	2	2	3
<i>Differential privacy</i>	8	<i>Differential privacy</i>	8	8	0	0
<i>Utility</i>	5	<i>Utility</i>	5	1	4	0
		<i>Utility measure/metric</i>	4	4	0	0
<i>Usefulness</i>	7	<i>Usefulness</i>	7	0	7	0
<i>Micro data</i>	9	<i>Micro data</i>	9	0	2	7
<i>Bayesian Networks</i>	7	<i>Bayesian Networks</i>	7	5	2	0
<i>Metadata</i>	11	<i>Metadata</i>	11	2	1	8
<i>Big Data</i>	11	<i>Big Data</i>	11	0	0	11
<i>GANS</i>	9	<i>GANS</i>	9	8	0	1
<i>AIML</i>	9	<i>AIML</i>	9	0	0	9

Summary of findings:

- The term *synthetic data* is explicitly defined in 9 out of 11 reviewed sources, whereas *synthesis* is explicitly defined in only 2 out of 11.
- Fully and partially synthetic data are discussed in 6 out of 11 sources. Partially synthetic data is explicitly defined in 100% of sources that mention it, while fully synthetic data is explicitly defined in only 4 out of 6 instances, often relying on contextual implications.
- *Mock data* is rarely explicitly defined, with only 1 out of 8 sources providing a definition. *Dummy data*, on the other hand, shows a mix of definitions.
- *Fidelity* is explicitly defined in 6 out of 11 sources.
- *High-fidelity* and *low-fidelity* data are mentioned inconsistently. Low-fidelity data appears in fewer sources and lacks explicit definitions, while high-fidelity data is explicitly defined in only 2 out of 6 sources.

- *Utility* is generally not explicitly defined, although utility measures and metrics are commonly well-articulated.
- *Usefulness* is not defined explicitly, but sufficient context is provided to infer its meaning.
- Techniques like *differential privacy*, *Bayesian networks*, and *GANs* are explicitly defined in most reviewed sources.
- Concepts like *microdata*, *metadata*, *big data*, and *AI/ML* are typically not explicitly defined.

Academic Literature Reference List

Search term: Synthetic data

1. Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. *J Stat Softw.* 2016;74(11):1–26.
2. Shanley D, Hogenboom J, Lysen F, Wee L. Getting real about synthetic data ethics: Are AI ethics principles a good starting point for synthetic data ethics? *EMBO Rep.* 2024;25(5):2152–2155.
3. Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak.* 2019;19(1):44.
4. Raghunathan TE. Synthetic Data. *Annu Rev Stat Appl.* 2020;7:129–141.
5. Patel P. Synthetic data. *Bus Inf Rev.* 2024;0(0):1–5.
6. Bowen CM, Snok J. Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. *J Priv Confidential.* 2021;11(1):1–26.
7. Rubin DB. Statistical Disclosure Limitation. *J Off Stat.* 1993;9(2):461–468.
8. Little RJA. Statistical Analysis of Masked Data. *J Off Stat.* 1993;9(2):407–426.
9. Raghunathan TE, Reiter JP, Rubin DB. Multiple Imputation for Statistical Disclosure Limitation. *J Off Stat.* 2003;19(1):1–16.
10. Reiter JP. Using CART to generate partially synthetic public use microdata. *J Off Stat.* 2005;21(3):441–462.
11. Raab GM, Nowok B, Dibben C. Practical Data Synthesis for Large Samples. *J Priv Confidential.* 2016;7(3):67–97.
12. Arnold C, Neunhoeffler 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>
13. Bowen CM, Liu F, Su B. Differentially Private Data Release via Statistical Election to Partition Sequentially. *METRON.* 2021;79(1):1–31.
14. Charest AS. How Can We Analyze Differentially-Private Synthetic Datasets? *J Priv Confidential.* 2011;2(2):83–102.
15. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our Data, Ourselves: Privacy via Distributed Noise Generation. In: *Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 2006)*; 2006 May 28–June 1; St. Petersburg, Russia. Berlin: Springer; 2006. p. 486–503.
16. Ping H, Stoyanovich J, Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In: *Proceedings of the 29th International Conference on Scientific and*

Statistical Database Management (SSDBM 2017); 2017 Jun 27–29; Chicago, IL, USA. New York: ACM; 2017. p. 1–5.

17. Quick H. Generating Poisson-Distributed Differentially Private Synthetic Data. *J R Stat Soc Ser A Stat Soc.* 2021;184(3):1093–1108.
18. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and Specific Utility Measures for Synthetic Data. *J R Stat Soc Ser A Stat Soc.* 2018;181(3):663–688.
19. Lucini F. The Real Deal About Synthetic Data. *MIT Sloan Manag Rev.* 2021 Oct 20. Available from: <https://sloanreview.mit.edu/article/the-real-deal-about-synthetic-data/>
20. Emam KE, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. Sebastopol (CA): O'Reilly Media; 2020.

Search term: Synthesis

21. Sallier K. Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis. *Stat J IAOS.* 2020;36(4):1059–1066.
22. Du Y, Cai Y, Jin X, Wang H, Li Y, Lu M. ASIDS: A Robust Data Synthesis Method for Generating Optimal Synthetic Samples. *Mathematics.* 2023;11(18):3891.

Search term: Fidelity

23. You H, Yu Y, Trask N, Gulian M, D'Elia M. Data-Driven Learning of Nonlocal Physics from High-Fidelity Synthetic Data. *Comput Methods Appl Mech Eng.* 2020;365:113029.
24. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digit Med.* 2020;3(1):147.

Search term: Multiple imputation

25. Hu J, Akande O, Wang Q. Multiple Imputation and Synthetic Data Generation with the R Package NPBayesImputeCat. *R J.* 2021;13(2):90–105.
26. Loong B, Rubin DB. Multiply-imputed synthetic data: advice to the imputer. *J Off Stat.* 2017;33(4):1005–1019.

Search term: Differential privacy

27. Xin B, Geng Y, Hu T, Chen S, Yang W, Wang S. Federated synthetic data generation with differential privacy. *Neurocomputing.* 2022;468:1–10.
28. Bao E, Xiao X, Zhao J, Zhang D, Ding B. Synthetic Data Generation with Differential Privacy via Bayesian Networks. *J Priv Confidentiality.* 2021;11(3):1–20.

Search term: Utility

29. Drechsler J, Reiter JP. Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *J Priv Confidential.* 2009;1(2):1–29.

30. Pereira M, Kshirsagar M, Mukherjee S, Dodhia R, Lavista Ferres J, de Sousa R. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *PLOS ONE*. 2023;18(10):e0297271.

Search term: Microdata

31. Little C, Elliot M, Allmendinger R. Comparing the utility and disclosure risk of synthetic data with samples of microdata. *arXiv*. 2022 Jul 2. Available from: <https://arxiv.org/abs/2207.03339>
32. Lin Y, Xiao N. Generating small areal synthetic microdata from public aggregated data using an optimization method. *Prof Geogr*. 2023;75(1):1–15.

Search term: Monte Carlo Simulation

33. Courtenay LA, Aramendi J, González-Aguilera D. Recruiting a skeleton crew—Methods for simulating and augmenting paleoanthropological data using Monte Carlo based algorithms. *Am J Biol Anthropol*. 2023;181(3):454–473.
34. Gan G, Valdez EA. Nested stochastic valuation of large variable annuity portfolios: Monte Carlo simulation and synthetic datasets. *Data*. 2018;3(3):31.

Search term: Bayesian Networks

35. Kaur D, Sobiesk M, Patil S, Liu J, Bhagat P, Gupta A, Markuzon N. Application of Bayesian networks to generate synthetic health data. *J Am Med Inform Assoc*. 2021;28(4):801–811.
36. Gogoshin G, Branciamore S, Rodin AS. Synthetic data generation with probabilistic Bayesian Networks. *Math Biosci Eng*. 2021;18(6):8603–8621.

Search term: Metadata

37. Isasa I, Hernández M, Epelde G, Londoño F, Beristain A, Larrea X, Alberdi A, Bamidis P, Konstantinidis E. Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis. *BMC Med Inform Decis Mak*. 2024;24(1):1–15.
38. Hong T, Macumber D, Li H, Fleming K, Wang Z. Generation and representation of synthetic smart meter data. *Build Simul*. 2020;13(6):1225–1237.

Search term: Big data

39. Nicolaie MA, Füssenich K, Ameling C, Boshuizen HC. Constructing synthetic populations in the age of big data. *Popul Health Metrics*. 2023;21(1):19.
40. Lun R, Siegal D, Ramsay T, et al. Synthetic data in cancer and cerebrovascular disease research: A novel approach. *PLoS One*. 2024;19(2):e0295921.

Search term: Generative adversarial network

41. Narika R, Fukatsu Y, Wang Z, Ogawa T, Adachi Y, Tanaka Y, Ishikawa S. Generative adversarial networks-based synthetic microstructures for data-driven materials design. *Adv Theory Simul*. 2022;5(5):2100470.

42. Yilmaz B, Korn R. Synthetic demand data generation for individual electricity consumers: Generative adversarial networks (GANs). *Energy AI*. 2022;9:100161.

Search term: Artificial intelligence

43. Dindorf C, Dully J, Konradi J, Wolf C, Becker S, Simon S, Huthwelker J, Werthmann F, Kniepert J, Drees P, Betz U, Fröhlich M. Enhancing biomechanical machine learning with limited data: generating realistic synthetic posture data using generative artificial intelligence. *Front Bioeng Biotechnol*. 2024;12:1350135.
44. Mavrogenis AF, Scarlat MM. Artificial intelligence publications: synthetic data, patients, and papers. *Int Orthop*. 2023;47(6):1395–1396.

Grey Literature Reference List

Search term: Synthesis

1. Mostly AI. What is synthetic data? [Internet]. Available from: <https://mostly.ai/what-is-synthetic-data>
2. Turing. Synthetic Data Generation: Definition, Types, Techniques, and Tools [Internet]. Available from: <https://www.turing.com/kb/synthetic-data-generation-techniques>
3. AltexSoft. Synthetic Data for Machine Learning: Its Nature, Types, and Means of Generation. [Internet]. 2022 March 22. Available from: <https://www.altexsoft.com/blog/synthetic-data-generation/>
4. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, Cohen SN, Weller A. Synthetic Data--what, why and how?. *arXiv preprint arXiv:2205.03257*. 2022 May 6.
5. Riemann R. European Data Protection Supervisor. Synthetic data [Internet]. Available from: https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en
6. Dilmegani C. AIMultiple. Synthetic data generation [Internet]. 2024 Nov 11. Available from: <https://research.aimultiple.com/synthetic-data-generation/>
7. IBM. Synthetic data generation: Building trust by ensuring privacy and quality [Internet]. 2023 Nov 29. Available from: <https://www.ibm.com/blog/synthetic-data-generation-building-trust-by-ensuring-privacy-and-quality/>
8. Data Science Campus, ONS. Synthetic data for public good [Internet]. 2019 Feb 21. Available from: <https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/>

9. Mostly AI. Comparison of synthetic data generation methods [Internet]. 2022 Sept 1. Available from: <https://mostly.ai/blog/comparison-of-synthetic-data-generation-methods>
10. Woodcock C. Statistical disclosure control using synthetic data: Presentation [Internet]. 2008 Nov. Available from: https://doku.iab.de/fdz/events/2008/SDC-Workshop_Presentation_Woodcock.pdf
11. Raghunathan TE. Synthetic data. Annu Rev Stat Appl. 2021;8:129-40. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-040720-031848;jsessionid=Dpciwl5gFVRC9oblFCJN_bQdA2kf4i1XmYazVWa4.annurev-liv-e-10-241-10-91

Search term: Mock data

12. NHS AI Lab. Exploring how to create mock patient data: Synthetic data from real patient data [Internet]. 2022 March 14. Available from: <https://transform.england.nhs.uk/ai-lab/explore-all-resources/develop-ai/exploring-how-to-create-mock-patient-data-synthetic-data-from-real-patient-data/>
13. Datprof. Synthetic test data generation [Internet]. Available from: <https://www.datprof.com/solutions/synthetic-test-data-generation/>
14. QuestionPro. Synthetic test data [Internet]. Available from: <https://www.questionpro.com/blog/synthetic-test-data/>
15. Richman A. K2View. Synthetic data creation: 8 Must-Have Features [Internet]. 2023 Nov 30. Available from: <https://www.k2view.com/blog/synthetic-data-creation/>
16. Syntheticus. Everything you need to know about synthetic data [Internet]. Available from: <https://syntheticus.ai/guide-everything-you-need-to-know-about-synthetic-data>
17. Sawant A. Beyond mock data: Synthetic revolution in AI development [Internet]. 2024 March 26. LinkedIn; Available from: <https://www.linkedin.com/pulse/beyond-mock-data-synthetic-revolution-ai-development-anagh-sawant-cm9kf/>
18. Marshall K. Deepchecks. Is synthetic data fake data? [Internet]. Available from: <https://deepchecks.com/question/is-synthetic-data-fake-data/>
19. Syntho. What is synthetic data? [Internet]. Available from: <https://www.syntho.ai/what-is-synthetic-data/>

Search term: Dummy data

20. NHS Digital. Synthetic data in live environments [Internet]. Available from: <https://digital.nhs.uk/services/e-referral-service/document-library/synthetic-data-in-live-environments>
21. Chirazi M. Datamaker. Synthetic data generation [Internet]. 2023 May 15. Available from: <https://datamaker.app/blog/synthetic-data-generation>
22. ADR UK. An interim position statement on synthetic data [Internet]. 2023 July. Available from: <https://www.adruk.org/news-publications/publications-reports/an-interim-position-statement-from-adr-uk-on-synthetic-data-new645d0cdb19d98062623548/>
23. Sogeti. Synthetic data management [Internet]. Available from: <https://www.sogeti.com/ai-for-qe/section-5-manage-data/chapter-2/>
24. Fekete A. Mostly AI. Synthetic data generator for healthy test data [Internet]. 2022 July 4. Available from: <https://mostly.ai/blog/synthetic-data-generator-for-healthy-test-data>
25. Statrice. Open-source tools for synthetic data generation [Internet]. Available from: <https://www.statice.ai/post/synthetic-data-open-source-tools-guide-building-buying>
26. Rosales A. Nasstar. What is synthetic data? [Internet]. Available from: <https://www.nasstar.com/hub/blog/what-is-synthetic-data>

Search term: Fidelity

27. Marshall V, Markham C, Avramovic P, Comerford P, Maple C, Szpruch L. Financial Conduct Authority. Exploring synthetic data validation: Privacy, utility, fidelity. 2023 June 12. [Internet]. Available from: <https://www.fca.org.uk/publications/research-articles/exploring-synthetic-data-validation-privacy-utility-fidelity>
28. Mostly AI. Synthetic data dictionary: Fidelity [Internet]. Available from: <https://mostly.ai/synthetic-data-dictionary/fidelity>
29. Haddad F. AWS. Evaluating synthetic data: Fidelity, utility, and privacy [Internet]. 2022 Dec 16. Available from: <https://aws.amazon.com/blogs/machine-learning/how-to-evaluate-the-quality-of-the-synthetic-data-measuring-from-the-perspective-of-fidelity-utility-and-privacy/>
30. Wang Z, Draghi B, Rotalinti Y, Lunn D, Myles P. High-Fidelity Synthetic Data Applications for Data Augmentation. 2024 Jan 12. DOI:10.5772/intechopen.113884.

31. ADR UK. User guide for synthetic data tool [Internet]. Available from: <https://www.adruk.org/fileadmin/uploads/adruk/Documents/User-Guide-Synthetic-Data-Tool.pdf>
32. CPRD. Synthetic data [Internet]. Available from: <https://www.cprd.com/synthetic-data>
33. YData. Visual evaluation of synthetic data quality [Internet]. 2023 Nov 10. Available from: <https://ydata.ai/resources/how-to-visually-evaluate-your-synthetic-data-quality>
34. Financial Conduct Authority. Using synthetic data in financial services [Internet]. 2024 March. Available from: <https://www.fca.org.uk/publication/corporate/report-using-synthetic-data-in-financial-services.pdf>
35. Yoon J, Arik S. Google Research. EHR-Safe: Generating high-fidelity and privacy-preserving synthetic electronic health records [Internet]. 2022 Dec 21. Available from: <https://research.google/blog/ehr-safe-generating-high-fidelity-and-privacy-preserving-synthetic-electronic-health-records/>
36. Mendelevitch O, Lesh MD. Fidelity and Privacy of Synthetic Medical Data: Review of Methods and Experimental Results. 2021 June. Available from: <https://arxiv.org/pdf/2101.08658>

Search term: Differential privacy

37. Differential Privacy. Synthetic data: An overview [Internet]. Available from: <https://differentialprivacy.org/synth-data-0/>
38. Near J, Darais D. NIST. Differentially private synthetic data [Internet]. 2021 May 3. Available from: <https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data>
39. De Jesus, R. Stalice. Differential privacy: Mechanisms and examples [Internet]. 2023 Dec 12. Available from: <https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples>
40. Stalice. What is Differential Privacy? Definition, Mechanisms, Examples [Internet]. 2022. Available from: <https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples>
41. Watson, A. How To Create Differentially Private Synthetic Data. [Internet]. 2020 Nov 23. Available from: <https://towardsdatascience.com/how-to-create-differentially-private-synthetic-data-ae1a04cff1b9>
42. Programming Differential Privacy. Chapter 3: Differential Privacy in Practice [Internet]. Available from: <https://programming-dp.com/ch3.html>

43. Kopp, A. Microsoft. Privacy-preserving synthetic data with SmartNoise [Internet]. 2021 February 18. Available from: <https://cloudblogs.microsoft.com/opensource/2021/02/18/create-privacy-preserving-synthetic-data-for-machine-learning-with-smartnoise/>
44. Xiao X. Synthesizing Relational Data with Differential Privacy [presentation]. School of Computing, National University of Singapore; [date unknown].

Search term: Utility

45. Taub J. Elliot M, Sakshaug. UNECE. Synthetic data: Methods and use cases [Internet]. Available from: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/4_Taub.pdf
46. K2View. What is synthetic data generation? [Internet]. Available from: <https://www.k2view.com/what-is-synthetic-data-generation/>
47. SBA Research. Synthetic data utility evaluation for machine learning [Internet]. Available from: https://www.sba-research.org/wp-content/uploads/2020/09/MLDM_Synthetic-Data-Utility-Evaluation-for-Machine-Learning_WEB.pdf
48. O'Reilly Media. Practical synthetic data [Internet]. Available from: <https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/ch04.html>
49. Pickens M, Andre J, Morrison, G. Urban Institute. Understanding synthetic data [Internet]. Available from: <https://www.urban.org/sites/default/files/2023-01/Understanding%20Synthetic%20Data.pdf>

Search term: Usefulness

50. Dilmegani, C. AI Multiple. Synthetic data vs. real data [Internet]. 2024 Oct 31. Available from: <https://research.aimultiple.com/synthetic-data-vs-real-data/>
51. Raden, N. Diginomica. Assessing the value of synthetic data [Internet]. 2023 June 21. Available from: <https://diginomica.com/assessing-value-synthetic-data-ai-public-policy-issues-take-center-stage>
52. DWP Digital. Why synthetic data could be useful for a government department [Internet]. 2021 Jun 18. Available from: <https://dwpdigital.blog.gov.uk/2021/06/18/why-synthetic-data-could-be-useful-for-a-government-department/>

53. QuestionPro. Synthetic data vs. real data [Internet]. Available from: <https://www.questionpro.com/blog/synthetic-data-vs-real-data/>
54. Pure Storage. Why synthetic data is on the rise [Internet]. 2024 Feb 23. Available from: <https://blog.purestorage.com/purely-educational/fake-it-til-you-make-it-why-synthetic-data-is-on-the-rise/>
55. Poynter, R. NewMR. Synthetic data: An overview and taxonomy [Internet]. 2024 April 20. Available from: <https://newmr.org/blog/synthetic-data-an-overview-a-taxonomy-some-faqs/>
56. Johnny S. The Power of Synthetic Data Generation in Today's Data-Driven World [Internet]. 2024 Apr 24. Available from: <https://scottjohnny288.medium.com/the-power-of-synthetic-data-generation-in-todays-data-driven-world-60257ac03d19>

Search term: Microdata

57. Louis T, Miranda J. Census Bureau. Synthetic data: Public-use microdata [Internet]. 2014 Oct 14. Available from: <https://www.census.gov/newsroom/blogs/research-matters/2014/10/synthetic-data-public-use-micro-data-for-a-big-data-world.html>
58. The Alan Turing Institute. Quantifying utility and preserving privacy in synthetic data [Internet]. Available from: <https://www.turing.ac.uk/research/research-projects/quipp-quantifying-utility-and-preserving-privacy-synthetic-data-sets>
59. Abowd JM, Lane JJ. Synthetic Data and Confidentiality Protection. Longitudinal Employer-Household Dynamics Technical Paper No. TP-2003-10. U.S. Census Bureau, LEHD Program; 2003 Sep. Available from: [Synthetic Data and Confidentiality Protection](#)
60. Steinbakk GH, Langsrud Ø, Løland A. A brief overview of methods for synthetic data for official statistics. Note no. SAMBA/23/20 [Internet]. 2020 Jul 1. Available from: <https://nr.brage.unit.no/nr-xmloi/bitstream/handle/11250/2682092/methods-for-synthetic-data-SAMBA-23-20.pdf?sequence=1>
61. Drechsler J, Haensch AC. 30 years of synthetic data. *Statist Sci*. 2024 May;39(2):221–42. doi: 10.1214/24-STS927.
62. Stinson M, Abowd J, Benedett G. Synthetic Micro-data Based on Integrating the Survey of Income and Program Participation with Lifetime Earnings and Retirement Benefits Data [presentation]. [date unknown]. Available from: <https://doku.iab.de/fdz/events/2007/Abowd.pdf>

63. Gussenbauer J. UNECE. AI-based privacy preserving census(like) data publication. [presentation]. Poland: 1-3 Dec 2021. Available from: https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Gussenbauer_AD.pdf
64. Rodrigues B. Synthpop: A tool for generating synthetic data [Internet]. 2020 Feb 23. Available from: <https://www.brodrigues.co/blog/2020-02-23-synthpop/>
65. Hittmeir M, Ekelhart A, Mayer R. Synthetic Data Utility Evaluation for Machine Learning. Available from: https://www.sba-research.org/wp-content/uploads/2020/09/MLDM_Synthetic-Data-Utility-Evaluation-for-Machine-Learning_WEB.pdf

Search term: Bayesian networks

66. Philip AM. Synthetic Data Generation Using Bayesian Networks: A Step-by-Step Guide [Internet]. Medium. Available from: <https://medium.com/@annmariaphilip8/synthetic-data-generation-using-bayesian-networks-a-step-by-step-guide-3ff3071e7c59>
67. Adele G. Using Bayesian Networks to Generate Synthetic Health Data [Internet]. Medium. Available from: <https://medium.com/@gideonadele/using-bayesian-networks-to-generate-synthetic-health-data-ed2b24effafe>
68. The Turing Institute. Synthetic Data [Internet]. Available from: <https://www.turing.ac.uk/research/interest-groups/synthetic-data>
69. Maddern K. Clear guidelines needed for synthetic data to ensure transparency, accountability and fairness, study says [Internet]. University of Exeter; 2024 Apr 15. Available from: <https://news.exeter.ac.uk/faculty-of-humanities-arts-and-social-sciences/clear-guidelines-needed-for-synthetic-data-to-ensure-transparency-accountability-and-fairness-study-says/>
70. Röhm R. Prospects and Limitations of Synthetic Data [Internet]. LinkedIn. Available from: <https://www.linkedin.com/pulse/prospects-limitations-synthetic-data-robin-r%C3%B6hm/>
71. Oehm D. Simulating Data with Bayesian Networks [Internet]. 2019 Oct 15. Available from: <https://gradientdescending.com/simulating-data-with-bayesian-networks/>
72. Little C. Generative adversarial networks for synthetic data generation: A comparative study. Conference of European Statisticians, Expert Meeting on Statistical Data Confidentiality; 2021 Dec 1-3; Poland. Available from: https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Little_AD.pdf

Search term: Metadata

73. GenRocket. How to leverage metadata for synthetic data generation [Internet]. Available from: <https://www.genrocket.com/blog/how-to-leverage-metadata-for-synthetic-data-generation/>
74. Pradhan SC. Synthetic data generation using SDV – Part 1 [Internet]. Medium; 2023 Nov 18. Available from: <https://medium.com/@soubhagya.cse.pradhan/synthetic-data-generation-using-sdv-part-1-08e63a8df8a2>
75. Bijit. Why you need synthetic data for machine learning [Internet]. Medium; 2024 Mar 3/ Available from: <https://medium.com/@bijit211987/why-you-need-synthetic-data-for-machine-learning-71e10405ee7e>
76. Kumar S. Synthetic Data Vault: A framework to generate synthetic data [Internet]. 2023 Feb 25. Available from: <https://www.aitude.com/synthetic-data-vault-a-framework-to-generate-synthetic-data/>
77. Owen S. Synthetic data for better machine learning [Internet]. 2023 Apr 12. Available from: <https://www.databricks.com/blog/2023/04/12/synthetic-data-better-machine-learning.html>
78. Office for National Statistics. Synthetic data pilot [Internet]. ONS Methodology Working Paper Series No. 16. Available from: <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpapersseries/onsmethodologyworkingpapersseriesnumber16syntheticdatapilot>
79. Ebert A. Time to Define Synthetic Data [Internet]. 2022 Dec 5. Available from: <https://mostly.ai/blog/define-synthetic-data>
80. Lets Data Science. Synthetic data generation [Internet]. Available from: <https://letsdatascience.com/synthetic-data-generation/>
81. Alizadeh E. SDV library for modeling datasets [Internet]. Available from: <https://ealizadeh.com/blog/sdv-library-for-modeling-datasets/>
82. Cohen R. Synthetic data generation with SDV: A comprehensive overview [Internet]. LinkedIn. Available from: <https://www.linkedin.com/pulse/synthetic-data-generation-sdv-comprehensive-overview-reuven-cohen-2f9rc/>
83. Temenos. Data sheet: Module synthetic [Internet]. 2022 Jun. Available from: https://www.temenos.com/wp-content/uploads/2022/06/Data-Sheet-Module-Synthetic_v02.pdf
84. Ebert A. Time to Define Synthetic Data [Internet]. 2022 Dec 5. Available from: <https://mostly.ai/blog/define-synthetic-data>

Search term: Big data

85. PwC. Synthetic data: The future of data privacy and business innovation [Internet]. 2023 Dec 07. Available from: <https://www.pwc.com/gx/en/issues/technology/synthetic-data.html#:~:text=Synthetic%20data%20is%20exactly%20what,directly%20from%20real%2Dworld%20activities.>
86. Buyuk E. The Importance of Synthetic Data in the Big Data World [Internet]. Towards Data Science; 2021 Nov 12. Available from: <https://towardsdatascience.com/the-importance-of-synthetic-data-in-the-big-data-world-7e50fc64854f>
87. Yasar K, Laskowski N. What is Synthetic data [Internet]. TechTarget. Available from: <https://www.techtarget.com/searchcio/definition/synthetic-data>
88. Gretel.ai. What is synthetic data? [Internet]. Available from: <https://gretel.ai/blog/what-is-synthetic-data>
89. Donovan R. Synthetic data, big data, fewer privacy risks [Internet]. Stack Overflow Blog; 2022 Oct 17. Available from: <https://stackoverflow.blog/2022/10/17/synthetic-data-big-data-fewer-privacy-risks/>
90. CBS. What is synthetic data? [Internet]. Available from: <https://www.cbs.nl/en-gb/about-us/innovation/project/what-is-synthetic-data->
91. IBM. Synthetic data [Internet]. Available from: <https://www.ibm.com/topics/synthetic-data>
92. Technavio. Synthetic data generation market analysis [Internet]. 2024 Jul. Available from: <https://www.technavio.com/report/synthetic-data-generation-market-analysis>
93. BetterDataAI. Can synthetic data be as accurate as real data? [Internet]. LinkedIn. Available from: <https://www.linkedin.com/pulse/can-synthetic-data-accurate-real-betterdataai-saicf/>
94. Toews R. Synthetic data is about to transform artificial intelligence [Internet]. Forbes; 2022 Jun 12. Available from: <https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/>
95. NVIDIA Blog. What is synthetic data? [Internet]. 2021 June 8. Available from: <https://blogs.nvidia.com/blog/what-is-synthetic-data/>

Search term: Generative adversarial network

96. Impetus Technologies. Synthetic data generation using GANs. Impetus. 2023 Dec 7. Available from: <https://www.impetus.com/resources/blog/synthetic-data-generation-using-gans/>
97. XQ. Exploring GANs to generate synthetic data. Medium. 2023 Nov 18. Available from: <https://medium.com/the-research-nest/exploring-gans-to-generate-synthetic-data-ca48f8a4b518>
98. Sharifi N. Generative AI: Synthetic data generation with GANs using PyTorch. Towards Data Science. 2024 Mar 3. Available from: <https://towardsdatascience.com/generative-ai-synthetic-data-generation-with-gans-using-pytorch-2e4dde8a17dd>
99. ONS. Generative adversarial networks (GANs) for synthetic dataset generation with binary classes. Data Science Campus, ONS. Available from: <https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/>
100. Pandya Y. Are Generative Adversarial Networks (GANs) the future of synthetic data generation? Medium. 2023 Dec 19. Available from: <https://medium.com/@yagnesh.pandya/are-generative-adversarial-networks-gans-the-future-of-synthetic-data-generation-e1a04f7b428d>
101. Saxena P. Synthetic data generation using conditional GAN. Towards Data Science. 2021 Aug 12. Available from: <https://towardsdatascience.com/synthetic-data-generation-using-conditional-gan-45f91542ec6b>
102. Sonwane D. Structured synthetic data generation using GANs. AIMind. 2023 Jul 17. Available from: <https://pub.aimind.so/structured-synthetic-data-generation-using-gans-af88932cadcd>
103. Vieira A. Generating synthetic data with referential integrity using GANs. Hazy. 2020 Apr 27. Available from: <https://hazy.com/resources/2020/04/27/generating-synthetic-data-with-referential-integrity-using-gans>

Search Term: Artificial intelligence machine learning

104. El Emam K, editor. Accelerating AI with Synthetic Data. O'Reilly Media; 2023. Available from: [https://www.oreilly.com/library/view/accelerating-ai-with/9781492045991/ch01.html​:contentReference\[oaicite:0\]{index=0}](https://www.oreilly.com/library/view/accelerating-ai-with/9781492045991/ch01.html​:contentReference[oaicite:0]{index=0})
105. Chakraborty J. AI makes data: How synthetic data is revolutionizing data science. Medium. 2024 Apr 1. Available from: <https://medium.com/@joyhit19/ai->

[makes-data-how-synthetic-data-is-revolutionizing-data-science-4dfa242cda56​;contentReference\[oaicite:0\]{index=0}](#)

106. Thomas A. Synthetic data: Fake is the new real. EY. 2022 Dec 1. Available from: [https://www.ey.com/en_in/technology/synthetic-data-fake-is-the-new-real​;contentReference\[oaicite:0\]{index=0}](https://www.ey.com/en_in/technology/synthetic-data-fake-is-the-new-real​;contentReference[oaicite:0]{index=0})
107. Skoon. The importance of synthetic data in AI. LinkedIn. 2023 Jun 6. Available from: [https://www.linkedin.com/pulse/importance-synthetic-data-artificial-intelligence-ai-skoon-data/​;contentReference\[oaicite:3\]{index=3}](https://www.linkedin.com/pulse/importance-synthetic-data-artificial-intelligence-ai-skoon-data/​;contentReference[oaicite:3]{index=3})
108. El Emam K, editor. *Practical Synthetic Data Generation*. O'Reilly Media; 2024. Available from: [https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/ch01.html​;contentReference\[oaicite:4\]{index=4}](https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/ch01.html​;contentReference[oaicite:4]{index=4})
109. Diehl SA, Cabral CG. Synthetic data gets real. Proskauer in Life Sciences. 2021 Jun 29. Available from: <https://www.proskauerinlifesciences.com/2021/06/synthetic-data-gets-real/>
110. Skoon. White paper: The importance of synthetic data in artificial intelligence. LinkedIn. 2023 Jun 16. Available from: [https://www.linkedin.com/pulse/white-paper-importance-synthetic-data-artificial-intelligence/​;contentReference\[oaicite:6\]{index=6}](https://www.linkedin.com/pulse/white-paper-importance-synthetic-data-artificial-intelligence/​;contentReference[oaicite:6]{index=6})
111. Office of the Privacy Commissioner of Canada. Research funding for privacy and knowledge translation: Synthetic data privacy project. 2021-2022. Available from: [https://www.priv.gc.ca/en/opc-actions-and-decisions/research/funding-for-privacy-research-and-knowledge-translation/completed-contributions-program-projects/2021-2022/p_2021-22_10/​;contentReference\[oaicite:7\]{index=7}](https://www.priv.gc.ca/en/opc-actions-and-decisions/research/funding-for-privacy-research-and-knowledge-translation/completed-contributions-program-projects/2021-2022/p_2021-22_10/​;contentReference[oaicite:7]{index=7})
112. Kunc M. Harnessing the power of synthetic data for innovation and insights. E-think. 2023 Apr 15. Available from: [https://www.e-think.ac.at/moderate-harnessing-the-power-of-synthetic-data-for-innovation-and-insights/​;contentReference\[oaicite:8\]{index=8}](https://www.e-think.ac.at/moderate-harnessing-the-power-of-synthetic-data-for-innovation-and-insights/​;contentReference[oaicite:8]{index=8})