Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics

Bayrem Kaabachi[1,*], Jérémie Despraz[1], Thierry Meurers[2], Karen Otte[2], Mehmed Halilovic[2], Fabian Prasser[2], and Jean Louis Raisaro[1]

[1]Biomedical Data Science Center, Centre Hospitalier Universitaire Vaudois, Rue du Bugnon 21, Lausanne , 1003, Switzerland
[1]Medical Informatics Group, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Charitéplatz 1 Berlin 10117 Germany
[*]Corresponding Author, E-mail: mohamed-beyrem.kaabachi@chuv.ch

ABSTRACT

Introduction: Sharing and re-using health-related data beyond the scope of its initial collection is essential for accelerating research, developing robust and trustworthy machine learning algorithms methods that can be translated into clinical settings. The sharing of synthetic data, artificially generated to resemble real patient data, is increasingly recognized as a promising means to enable such a re-use while addressing the privacy concerns related to personal medical data. Nonetheless, no consensus exists yet on a standard approach for systematically and quantitatively evaluating the actual privacy gain and residual utility of synthetic data, de-facto hindering its adoption.
Objective: In this work, we present and systematize current knowledge on the field of synthetic health-related data evaluation both in terms of privacy and utility. We provide insights and critical analysis into the current state of the art and propose concrete directions and steps forward for the research community.
Methods: We assess and contextualize existing knowledge in the field through a scoping review and the creation of a common ontology that encompasses all the methods and metrics used to assess synthetic data. We follow the PRISMA-ScR methodology in order to perform data collection and knowledge synthesis.
Results: We include 92 studies in the scoping review. We analyze and classify them according to the proposed ontology. We found 48 different methods to evaluate the residual statistical utility of synthetic data and 9 methods that are used to evaluate the residual privacy risks. Moreover, we observe that there is currently no consensus among researchers regarding neither individual metrics nor family of metrics for evaluating the privacy and utility of synthetic data. Our findings on the privacy of synthetic data show that there is an alarming tendency to trust the safety of synthetic data without properly evaluating it.
Conclusion: Although the use of synthetic data in healthcare promises to offer an easy and hassle-free alternative to real data, the lack of consensus in terms of evaluation hinders the adoption of this new technology. We believe that, by raising awareness and providing a comprehensive taxonomy on evaluation methods that takes into account the current state of literature, our work can foster the development and adoption of uniform approaches and consequently facilitate the use of synthetic data in the medical domain.

Introduction

Access to high-quality data is critical for impactful medical research and practice, especially with the rise of *Artificial Intelligence* (AI) and *Machine Learning* (ML), as it drives progress and innovation in fields such as Precision Medicine[1] where establishing safe, fast, and reliable procedures to access data for secondary use has become essential.

Yet, due to privacy concerns, access to medical data is usually highly restricted[2] and subject to safeguards specified in data protection laws, such as the United States *Health Insurance Portability and Accountability Act* (HIPAA)[3] and the European Union *General Data Protection Regulation (GDPR)*[4]. A common approach used to share highly sensitive data under these regulatory frameworks is data anonymization below an acceptance threshold[5]. This approach employs data masking and transformation techniques to reduce re-identification risks.

Nonetheless, even in cases where a sufficient protection level can be achieved, anonymizing high-dimensional data often comes with a severe hit[6] to the utility of the anonymized dataset which can render it nearly unusable for research.

A promising solution to this data-sharing problem is synthetic data, which has been described by Chen et al.[7] as a technique that "will undoubtedly soon be used to solve pressing problems in healthcare". The main idea behind it is to generate artificial data that mimics the statistical properties of real patient data. This data synthesis process can be achieved using multiple algorithms but the main breakthrough in these last few years has been the use of *Generative Adversarial Networks* (GANs)[8]. GANs work by employing two neural networks: one creates fake samples, and the other assesses how close they are to real data. These networks collaborate to refine the generated samples until they closely resemble real data, making it a valuable tool for generating realistic artificial information. Since all samples are generated artificially, the probability that a synthetic sample would match a real one is usually very small.

As a result, synthetic data has garnered considerable coverage, even beyond specialized sources, and this broader recognition has led to bold predictions claiming that "By 2024, 60% of the data utilized for AI and analytics projects will be synthetically generated"[9].

In the medical domain in particular, several studies[10–13] have used synthetic data to replicate case studies originally performed on health-related data. These results highlight the potential benefits of synthetic data in the medical context and give strong arguments for the use of synthetic data as an alternative to strictly regulated personal data.

However, while these results seem promising for the future of privacy-preserving data sharing in medical environments, more recent studies have pointed out several risks associated with over-reliance on synthetic data as a "silver bullet" solution[14]. For instance, individual records that are part of the synthesized data could have a strong impact on the synthetic data generated, allowing a malicious adversary to infer the presence of individuals in the original dataset[14]. This especially relates to the tendency of machine learning models to overfit on training data and memorize leaks about individuals in the dataset[15]. Generally speaking, a synthetic dataset that most closely mimics the original dataset is likely also to be most useful, but at the same time, provide less privacy protection. On the other hand, a synthetic dataset that is very different from the original data will provide strong protection, but likely less utility. Due to the black-box nature of GANs, it is difficult to predict which data utility is lost in the training-and-generation process and which sensitive information might be contained in the generated data. As a consequence, Stadler et al.[14] argue that a cautious approach has to be taken when generating and sharing synthetic data.

The potential risks associated with synthetic data usage highlighted in recent studies[14,16,17] raise the question of whether research priorities in the synthetic data domain exhibit a stronger emphasis on utility over privacy considerations. Compared to anonymized data, where we can find an extensive literature[18] describing all kind of attacks and privacy protection mechanisms that can be applied, synthetic data has not yet been as thoroughly scrutinized. This prompted us to conduct this review in hopes that we would provide an informed and unbiased answer to that question.

A few surveys in the field have examined various aspects of synthetic data generation[19,20]. Figueira et al.[19] provide an extensive description of multiple generation methods while Hernandez et al.[20] explored evaluation methods and compared them to determine the best-performing ones. In contrast to these prior studies, our approach differs in how we identify the obstacles hindering the adoption of synthetic data as we place a greater emphasis on the evaluation process and the privacy-utility trade-off dilemma by having a systematic look at how synthetic data is evaluated across 92 studies.

In parallel, open-source solutions such as Synthetic Data Vault[21], Table Evaluator[22] and TAPAS[23] have been developed and publicly released to help researchers create and measure the quality of synthetic data. These platforms offer a selection of evaluation metrics and methods for assessing both utility and privacy, streamlining the evaluation process. However, these open-source tools present their own challenges as they each employ their own nomenclatures and terminologies, adding to the complexity of achieving a harmonized perspective on synthetic data within the healthcare domain. This, coupled with the presence of contradictory perspectives[14,16,24,25] in the literature complicates the development of a unified understanding of synthetic data in healthcare.

As a result, to get a better understanding of the current landscape in healthcare-related synthetic data generation, we initiated this scoping review to specifically target evaluation methodologies, aiming to provide a rigorous and quantitative analysis of the suitability of synthetic data evaluation methods. To do so, we have been guided by the following research questions:

1. Is there consensus within the community on how to evaluate the privacy and utility of synthetic data?
2. Is privacy and utility given the same importance when assessing synthetic data?

Methods

For this scoping review, we adopted the protocol from *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA[26]). PRISMA stands as a recognized guideline, commonly adopted for laying out systematic reviews and meta-analyses. Its framework is designed to bring clarity and consistency to the process. Specifically, PRISMA emphasizes the importance of clearly defining the research question, setting unambiguous inclusion and exclusion parameters, and detailing methods for searching, choosing, and gathering data from chosen documents.

To identify pertinent studies for this scoping review on synthetic data evaluation methods, we conducted a comprehensive search across multiple bibliographic databases and repositories spanning the period from January 2018 to December 2022. The databases and repositories described in Figure 1 included IEEExplore and the ACM Digital Library, which are primary repositories for computer science literature, and PubMed and Embase, which are focused on healthcare and medical research. Full-text articles were obtained for those meeting the inclusion criteria described in Table 1.

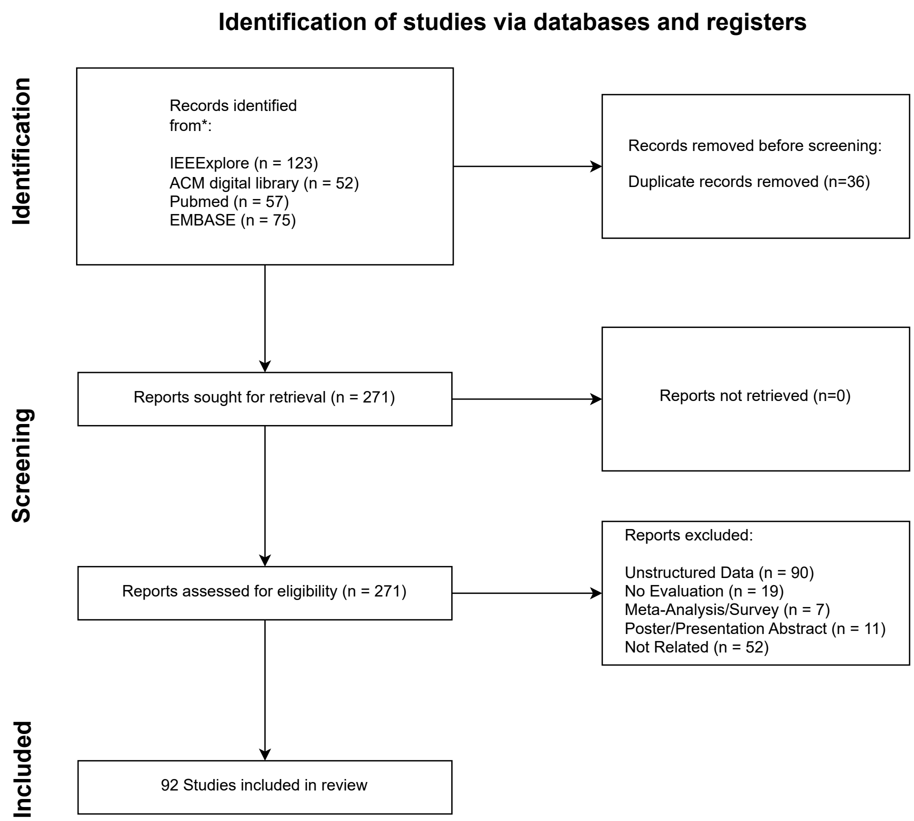

Figure 1. PRISMA flow diagram for the scoping review process.

The rationale for including computer science databases like IEEExplore and the ACM Digital Library was to capture more technologically advanced and innovative synthetic data generation and evaluation methods as these databases often contain articles about new techniques that have the potential to push the field forward.

Conversely, the inclusion of healthcare-specific databases like PubMed and Embase aimed to identify studies that might offer more grounded, practical, and clinically relevant evaluation methods. These databases are more likely to include studies that have considered the unique constraints and requirements of healthcare settings, thus ensuring that the synthetic data methods under review would be applicable in real-world medical contexts.

The search strategies for each database were developed at an early stage of the research and were then refined through team discussions and preliminary analysis of the results. We specifically designed the queries to focus on publications that evaluate the utility or privacy aspects of synthetic data. This was done to capture articles that provide actionable insights into the quality and safety of synthetic data methods, rather than merely describing new techniques. The queries used for each database are listed in Table 4 Appendix A and were last run on August 14, 2023.

Another consideration in query design was the avoidance of false positives, such as publications discussing synthetic compounds or materials rather than synthetic data. To this end, we included both "Title" and "Abstract" as fields for our queries, ensuring that the primary focus of the identified publications was indeed on synthetic data and its evaluation metrics for utility or privacy. We also removed such articles manually, should they have still appeared in the final selection of papers.

Any discrepancies in study selection were resolved through discussion and consensus between two of the authors. A data-charting form, illustrated in Table 5 Appendix A, was collaboratively designed by the research team to delineate the specific variables to be extracted from the selected publications. Upon selection, the information described in the data-charting form was extracted from each study.

The challenge in the data charting process was the standardization of properties as it ensured consistency in the extraction of information from the selected publications and enabled a quantitative evaluation of the studies under consideration.

To meet the requirement for standardization, we have created a taxonomy of evaluation methods. The subsequent section will present this taxonomy, concentrating on the facets of privacy and utility and their representation in the existing literature.

Table 1. Eligibility criteria

Inclusion criteria	Exclusion criteria
Publications describing research that uses synthetic data generation methods and evaluates their outputs.	Surveys and systematic/scoping reviews.
Papers published between 01.01.2018 and 31.12.2022.	Documents in languages other than English.
Publications describing work that focuses on structured data i.e. no images/text problems.	No assessment of the generated output i.e. no look at the utility/privacy aspect of the generated data.
–	Do not contain structured data.
–	Poster abstracts.

Taxonomy - Synthetic Data Utility

The taxonomy of utility methods shown in Figure 2 is first organized into several statistical categories: "Univariate Similarity", "Bivariate Similarity", and "Multivariate Similarity". A structured approach to evaluation streamlines the understanding of similarities between synthetic and real data across multiple dimensions, as it enables direct comparison for various generative methods. We also included a category for methods related to longitudinal data due to their unique nature of analyzing temporal patterns and trends. Another included category, "Domain Specific Similarity," evaluates how synthetic data performs in specific research areas, like replicating study results or using metrics particular to that domain.

A limitation of this approach involves the potential overlap in utility categories. For instance, performing a machine learning classification task on both synthetic and real data could fall under both "Domain Specific Similarity" as a "Replication of Studies" and "Multivariate Similarity" as "Classification Performance." Since it is challenging to discern the original intentions of the authors of the publications we examined, we opted to classify methods as "Replication of Studies" when any ambiguity arose to avoid conflict.

Appendix B contains a comprehensive description of each item in this taxonomy.

Taxonomy - Synthetic Data Privacy

The taxonomy of privacy methods shown in Figure 2 is organized into two categories: "Dataset Evaluation Methods" and "Model Evaluation methods".

Dataset evaluation methods assess the privacy protection provided by the synthetic data itself. The primary goal is to determine how well the synthetic dataset safeguards sensitive information and preserves privacy, especially when compared to the original real data. This evaluation is important when the primary concern is the privacy of the data itself, such as when a hospital wants to release a dataset to the public or when the focus is on releasing a single dataset. Often, these methods utilize distance metrics for comparison. They either contrast the generated synthetic dataset directly with the original one involved in the generation process or with a holdout dataset drawn from the same population.

Model evaluation methods shift the focus to assessing the privacy-preserving algorithm or method used to generate synthetic data which makes it possible to understand how well the chosen mechanism protects privacy across various scenarios, and it often involves computing an estimated upper bound on the privacy risk posed by synthetic data generation mechanisms. A common usage entails evaluating multiple outputs of the generation mechanism to establish what can be described as a "worst-case scenario". This evaluation is crucial when the emphasis is on the performance and robustness of the privacy-preserving mechanism itself. It helps fairly compare techniques between each other as both are evaluated in terms of the upper bounds of privacy risk. An example of this is the use of shadow models[15]. These models involve the creation of multiple replicas that mirror the behavior of the primary synthetic data model. Though this approach might be resource-intensive, it establishes a robust evaluation framework simulating a black-box attack scenario, ensuring a holistic privacy risk assessment.

Appendix C contains a comprehensive description of each item in this taxonomy.

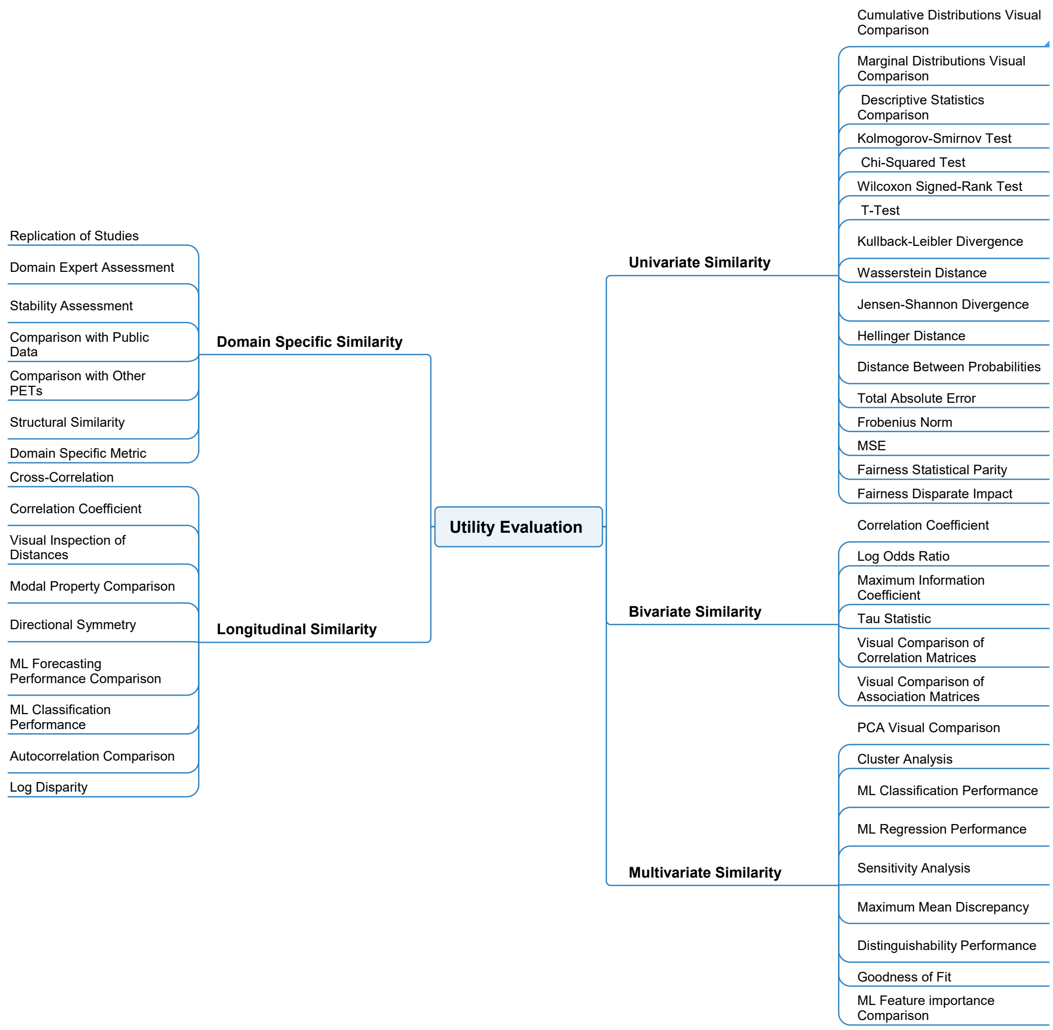

Figure 2. Taxonomy of synthetic data utility evaluation

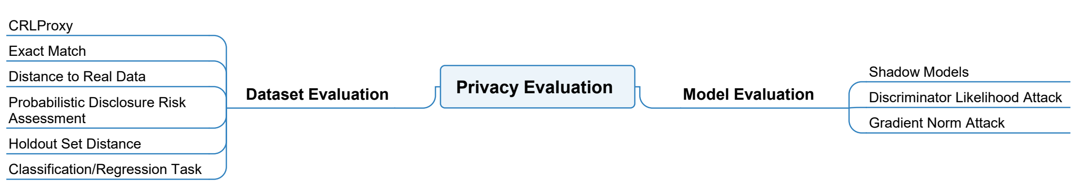

Figure 3. Taxonomy of synthetic data privacy evaluation

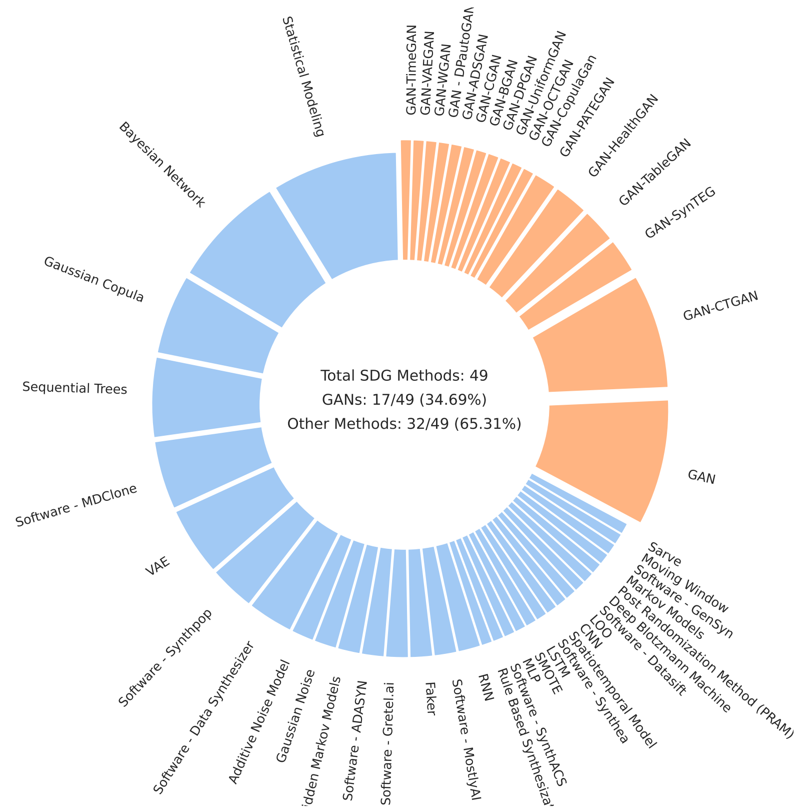

Figure 4. Synthetic Data Generation Methods

Results

General Results

In this review, we found that, after reconciling methods that were semantically the same but named differently under a unique definition, there were 48 methods used to assess utility and 9 methods used to assess privacy. Figure 7 gives an overview of the overall landscape of utility and privacy evaluation methods used in all the publications we selected. The full result of the scoping review can be found in Table 2 and in Table 3.

We reviewed articles published from 2018 to 2022, a timeframe encompassing the surge and ascendance of generative AI technologies, including the early enthusiasm for GANs and the advent of Large Language Models (LLMs). Based on Figure 9 Appendix D, we found that only 4.35% (4/92) were from 2018 and 10.87% (9/92) from 2019. By 2022, this percentage had jumped to 43.48% (40/92) which suggests a rising interest in synthetic data over time.

Additionally, we found that most articles used cross-sectional data, making up 70% (64/92) of the total. Only 26% (24/92) used temporal longitudinal data, possibly as it is usually harder to synthesize[27]. For this type of tabular data, the difficulty comes in maintaining relationships not just between columns which are reflected in the correlations between variables but also between rows which represent the temporal consistency of the data. As explained in Table 1, unstructured data were not considered during this review.

Different methods were used to create synthetic data. About 35% (17/49) of the articles used GANs. The rest, 65% (32/92), used a mix of other methods, including statistical modeling and specialized software like Synthpop[28] R package or the MDClone[29] platform.

Synthetic Data Utility

In our eligibility criteria, we specifically focused on works that evaluated the output of their Synthetic Data Generation method. Of these, 94% (86/92) evaluated the utility of synthetic data.

Among the 48 utility evaluation methods, we identified 17 that were for univariate similarity, 9 for longitudinal similarity, another 9 dedicated to multivariate similarity, 8 specific to domain-related similarity, and 6 specific to bivariate similarity.

Three methods stood out as the most commonly used. Multivariate Similarity: ML Classification Performance was the

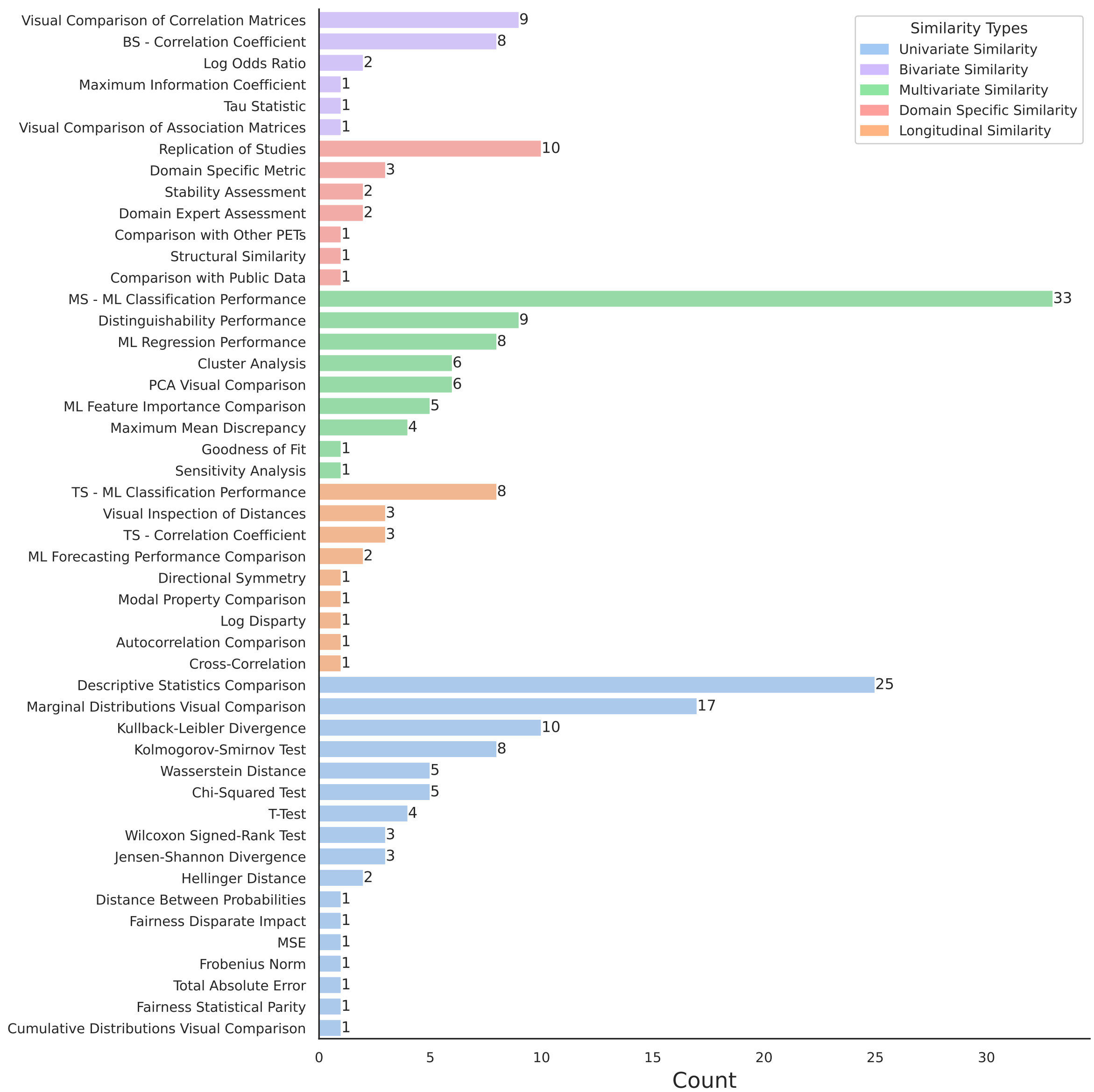

Figure 5. Synthetic Data Utility

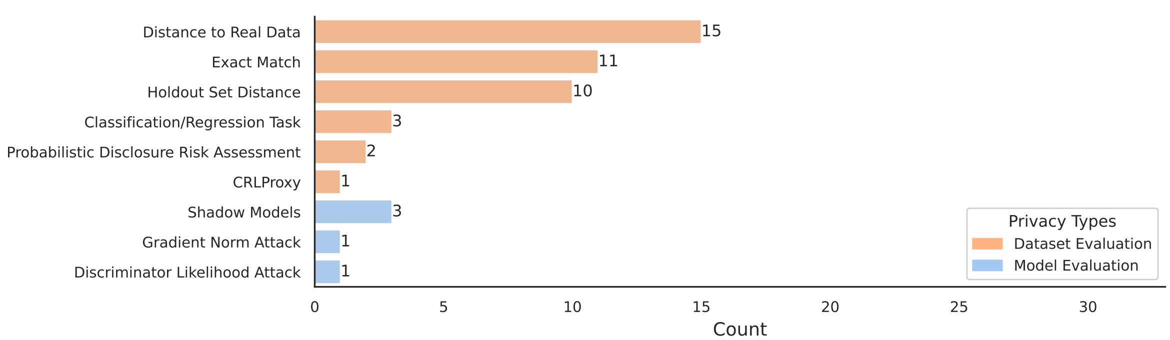

Figure 6. Synthetic Data Privacy

Figure 7. Synthetic Data Utility and Privacy Landscape

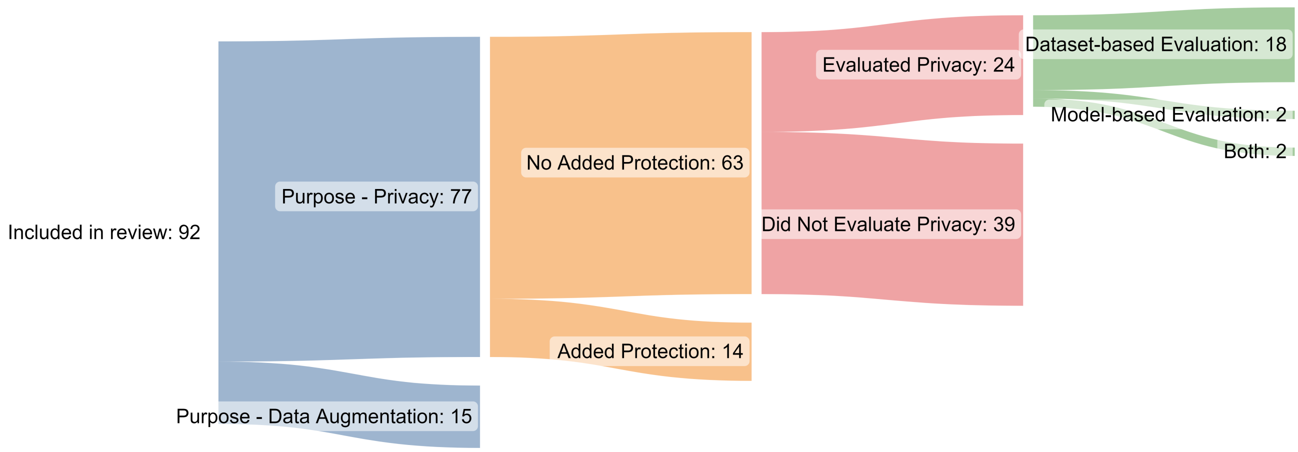

Figure 8. Sankey Diagram of how and whether privacy of synthetic data is evaluated

predominant method, applied in 33 instances. Univariate Similarity: Descriptive Statistics Comparison was used in 25 cases and Univariate Similarity: Marginal Distributions Visual Comparison was employed 17 times.

Synthetic Data Privacy

Figure 8 shows that the privacy aspect of synthetic data was the main incentive behind most selected papers as 80% (74/92) of them intended to use synthetic data for private data sharing scenarios. The other 16% (15/92) used it for data augmentation purposes and to answer either data scarcity problems or class imbalance. The remaining 4% (3/92) studied the potential of synthetic data in both scenarios.

Of these 77 studies that aimed to use synthetic data for privacy preservation, 15 applied either differential privacy techniques or introduced an extra masking layer to the data. This additional layer mask functions by adding a secondary level of data alteration, which further conceals the original records, ensuring that the actual data remains protected and less traceable to individual sources Of the 63 studies that remain, only 38% (24/63) included at least one privacy evaluation method. This implies that while the need for privacy in the assessed works was apparent, the evaluation part of the equation did not follow and the privacy of synthetic data has often been blindly trusted.

From the papers that did include a privacy evaluation, 84% (20/24) mainly relied on dataset-based evaluation. A smaller number, 8% (2/24), focused on based on the model or mechanism itself, such as those that exploit GAN architectures[30] or those that involve a shadow modeling process[31–33] and another 8& (2/24) performed both evaluations.

Discussion

RQ1: Is there consensus within the community on how to evaluate the privacy and utility of synthetic data?

Our findings shown in the previous section indicate that there is currently no consensus among researchers on standardized metrics for evaluating the privacy and utility of synthetic data. The use of a wide variety of metrics across studies makes it challenging to compare and synthesize the existing evidence.

As research in this area continues to grow, it is becoming more and more difficult to choose the best "solution" to generate high fidelity and high privacy synthetic data, as it is not possible to compare available solutions directly and fairly. This overall confusion around how to know if the up-and-coming new synthetic data generation method is truly fit for adoption renders these state-of-the-art techniques challenging to utilize in real-world environments which highlights the need for a standardized set of evaluation metrics to facilitate meaningful

This is even more apparent when it comes to privacy evaluations, as these are also bound by legal constraints. To the best of our knowledge, there is no clear legal text on how synthetic data privacy risk should be assessed. Although there have been recent attempts to map synthetic data metrics to existing GDPR definitions such as "singling out", "linkage" and "inference" by Giomi et al.[119], there is no confirmation yet about compliance.

In summary, we conclude that the field of synthetic data evaluation is still nascent. We anticipate that as both the technology matures and legal frameworks adapt, methods for evaluating synthetic data should converge into a more standardized and trustworthy approach. This is important, especially in critical sectors like healthcare, where stakeholder trust is indispensable.

RQ2: Is privacy and utility given the same importance when assessing synthetic data?

Our findings in Figure 7 and 8 clearly show that privacy evaluations are often not as thorough as utility evaluations.

While the utility of synthetic data has been a major focus, privacy evaluation is often quite limited and incomplete as there is a clear discrepancy between how many times methods that evaluate datasets are applied in the literature compared to methods that evaluate the mechanisms.

The under-evaluation of privacy in the use of synthetic data is particularly evident in this review. More than half of the studies claiming to employ synthetic data for its privacy-preserving attributes and that "should" evaluate privacy[1] did not conduct any formal privacy evaluation. Instead, they utilized synthetic data "as is" assuming inherent privacy benefits without empirical verification.

This oversight poses significant concerns, especially in the realm of software solutions that generate synthetic data. Users may inadvertently assume that the synthetic data they are generating is privacy-preserving by default. This may lead to the uninformed sharing of synthetic data, potentially resulting in personal data breaches in addition to ethical and legal complications.

The lax approach toward privacy evaluation, combined with assumptions about synthetic data's privacy-preserving capabilities, exposes a critical gap in current research and practice. It highlights the need for a more balanced approach in evaluating both utility and privacy in synthetic data generation methods.

Factors influencing the selection of evaluation methods

In evaluating the utility and privacy of synthetic data, a diverse range of approaches are evident. The literature reveals 48 distinct methods for utility evaluation, while the methods for privacy evaluation are fewer in comparison. The choice of these evaluation techniques can be attributed to multiple intertwined factors:

Research objectives play a pivotal role in method selection. For utility evaluation, when synthetic data is used for data augmentation, metrics largely gravitate towards machine learning tasks and traditional benchmarking[65,66]. Conversely, when synthetic data serves as a "proxy" for real data[11,39,55], the metrics are more focused on specific attributes over a broader assessment. In privacy considerations, the choice often falls between membership inference and attribute inference based on research goals. Membership inference, for example, is selected for its direct assessment of data leakage and as a precursor to examining the feasibility of intricate inference attacks.

The complexity of implementation is another crucial determinant. Simpler methods, such as univariate similarity comparisons or correlation matrix analyses for utility, and distance-based metrics for privacy, are favored due to their ease of implementation using standard software packages. Conversely, more intricate methods like log cluster metrics or shadow models require additional considerations like unsupervised learning or the training of multiple models.

Interpretability is also central to method choice. Evaluation techniques that allow for visual comparisons are often more attractive, especially when presenting to stakeholders. While some methods, like exact match attacks in privacy evaluation, offer clear interpretability, others demand more detailed interpretation due to their intricacies.

Lastly, the structure and type of data, as well as model generalizability, affect the selection process. Time-series data, for example, demands different utility metrics than cross-sectional data. Moreover, some attack methods are custom-designed for specific synthetic data generation techniques, such as GANs, where the discriminator could be utilized to quantify a risk factor[30], limiting their generalizability across various data generation techniques.

Limitations

This scoping review, while comprehensive, is not without limitations, as it is possible that some relevant studies or methods were not captured in our analysis. During the charting process and the development of our taxonomy, certain decisions had to be made that could potentially introduce subjectivity or limit the granularity of our evaluation. This is especially apparent when interpreting diverse metrics across multiple papers and attempting to consolidate them under a unified terminology.

For instance, a limitation pertains to the categorization of "domain-specific similarity" metrics as it became evident that the approaches under this category often have a scope or meaning that diverges from other metrics. This umbrella term might encompass various methods that differ significantly in their granularity and specific objectives. The decision to bucket these diverse metrics under "domain-specific similarity" was made to streamline the taxonomy, but we acknowledge that it might not be the most precise fit for each situation.

[1]We define a work that 'should evaluate' as one which asserts that synthetic data served as a privacy-preserving tool, without implementing any added protections like Differential Privacy.

In addition, we found that the terms "fidelity" and "utility" are sometimes used interchangeably, yet some research[120] argues that they should be considered as distinct metrics. Fidelity largely pertains to the statistical similarity between synthetic and original data, while utility focuses on the functional usefulness of the synthetic dataset for specific tasks. This distinction, while not directly addressed in this review, was still reflected in the construction of the taxonomy under the term "Domain Specific Similarity".

Conclusion

This review offers a detailed insight into the present research landscape of synthetic health data's utility and privacy revealing both its potential and pitfalls. The urgent requirement for standardized evaluation measures stands out as a major point where we think that having uniform metrics can offer a level playing field, allowing different synthetic data generation methods to be compared in a consistent and meaningful manner.

One significant concern raised throughout this work is the need for robust privacy evaluations. As the healthcare sector houses sensitive information, ensuring that synthetic data doesn't inadvertently lead to data leaks or result in a loss of trust is paramount. This is especially true when it comes to GANs as their inherent complexity and lack of transparency can either act as roadblocks by deterring many from adopting them or lead to misinformed usage due to a lack of awareness. The pressing need for standardized and secure synthetic data in healthcare is increasingly when international initiatives such as the IEEE's Industry Connections activity[121] and the Horizon Europe[122] call for synthetic data confirm the urgency of creating clear guidelines for the safe and the developing of reliable frameworks in the field. Thus, our intention with this review is not just to shed light on these challenges but also to inspire a collaborative effort in formulating best practices that make these techniques more accessible and understandable.

The journey of integrating synthetic data into healthcare environments should be treaded with caution. The allure of its capabilities should be tempered with a balanced view, avoiding over-promotion. Any evaluation or implementation should be approached methodically, ensuring the results are both valid and unbiased.

References

1. Johnson, K. B. *et al.* Precision Medicine, AI, and the Future of Personalized Health Care. *Clin. Transl. Sci.* **14**, 86–93, DOI: 10.1111/cts.12884 (2021).
2. Xiang, D. & Cai, W. Privacy Protection and Secondary Use of Health Data: Strategies and Methods. *BioMed Res. Int.* **2021**, 6967166, DOI: 10.1155/2021/6967166 (2021).
3. Privacy | HHS.gov. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.
4. General Data Protection Regulation (GDPR) – Official Legal Text. https://gdpr-info.eu/.
5. EMA. External guidance on the implementation of European Medicines Agency policy publication clinical data for medicinal products human use. https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data (2018).
6. Aggarwal, C. C. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, 901–909 (VLDB Endowment, Trondheim, Norway, 2005).
7. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5**, 493–497, DOI: 10.1038/s41551-021-00751-8 (2021).
8. Goodfellow, I. *et al.* Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc., 2014).
9. Castellanos, S. Fake It to Make It: Companies Beef Up AI Models With Synthetic Data. *Wall Str. J.* (2021).
10. Wang, Z., Myles, P. & Tucker, A. Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility & Patient Privacy. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 126–131, DOI: 10.1109/CBMS.2019.00036 (IEEE, Cordoba, Spain, 2019).
11. Azizi, Z. *et al.* SEX, GENDER AND CARDIOVASCULAR HEALTH, AN ANALYSIS OF SYNTHETIC DATA FROM A POPULATION BASED STUDY. *J. Am. Coll. Cardiol.* **77**, 3258, DOI: 10.1016/S0735-1097(21)04612-X (2021).
12. Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. & El Emam, K. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* **11**, e043497, DOI: 10.1136/bmjopen-2020-043497 (2021).

13. Cockrell, C., Schobel-McHugh, S., Lisboa, F., Vodovotz, Y. & An, G. Generating synthetic data with a mechanism-based Critical Illness Digital Twin: Demonstration for Post Traumatic Acute Respiratory Distress Syndrome. Preprint, Systems Biology (2022). DOI: 10.1101/2022.11.22.517524.

14. Stadler, T., Oprisanu, B. & Troncoso, C. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, 1451–1468 (USENIX Association, Boston, MA, 2022).

15. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18, DOI: 10.1109/SP.2017.41 (IEEE, San Jose, CA, USA, 2017).

16. Appenzeller, A., Leitner, M., Philipp, P., Krempel, E. & Beyerer, J. Privacy and Utility of Private Synthetic Data for Medical Data Analyses. *Appl. Sci.* **12**, 12320, DOI: 10.3390/app122312320 (2022).

17. Arthur, L. *et al.* On the Challenges of Deploying Privacy-Preserving Synthetic Data in the Enterprise. (2023).

18. Wagner, I. & Eckhoff, D. Technical Privacy Metrics: A Systematic Survey. *ACM Comput. Surv.* **51**, 1–38, DOI: 10.1145/3168389 (2019). 1512.00327.

19. Figueira, A. & Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **10**, 2733, DOI: 10.3390/math10152733 (2022).

20. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R. & Rankin, D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* **493**, 28–45, DOI: 10.1016/j.neucom.2022.04.053 (2022).

21. The Synthetic Data Vault. Put synthetic data to work! https://sdv.dev/.

22. Brenninkmeijer, B. Table Evaluator (2023).

23. Houssiau, F. *et al.* TAPAS: A Toolbox for Adversarial Privacy Auditing of Synthetic Data (2022). 2211.06550.

24. Platzer, M. & Reutterer, T. Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data. *Front. Big Data* **4**, 679939, DOI: 10.3389/fdata.2021.679939 (2021).

25. Hittmeir, M., Ekelhart, A. & Mayer, R. Utility and Privacy Assessments of Synthetic Data for Regression Tasks. In *2019 IEEE International Conference on Big Data (Big Data)*, 5763–5772, DOI: 10.1109/BigData47090.2019.9005476 (IEEE, Los Angeles, CA, USA, 2019).

26. PRISMA. https://prisma-statement.org/Extensions/ScopingReviews.

27. SynTEG: A framework for temporal structured electronic health data simulation | Journal of the American Medical Informatics Association | Oxford Academic. https://academic.oup.com/jamia/article/28/3/596/6024632.

28. Nowok, B., Raab, G. M. & Dibben, C. Synthpop: Bespoke Creation of Synthetic Data in R. *J. Stat. Softw.* **74**, 1–26, DOI: 10.18637/jss.v074.i11 (2016).

29. MDClone - The World's Most Powerful Healthcare Data Platform. https://www.mdclone.com.

30. Del Grosso, G., Pichler, G. & Piantanida, P. Privacy-Preserving Synthetic Smart Meters Data. In *2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 1–5, DOI: 10.1109/ISGT49243.2021.9372157 (IEEE, Washington, DC, USA, 2021).

31. Maeda, W., Higuchi, Y., Minami, K. & Morikawa, I. Membership Inference Countermeasure With A Partially Synthetic Data Approach. In *2022 4th International Conference on Data Intelligence and Security (ICDIS)*, 374–381, DOI: 10.1109/ICDIS55630.2022.00063 (IEEE, Shenzhen, China, 2022).

32. Kaabachi, B., Despraz, J., Meurers, T., Prasser, F. & Raisaro, J. L. Generation and Evaluation of Synthetic Data in a University Hospital Setting. In Séroussi, B. *et al.* (eds.) *Studies in Health Technology and Informatics*, DOI: 10.3233/SHTI220420 (IOS Press, 2022).

33. Dwork, C. & Roth, A. The Algorithmic Foundations of Differential Privacy. *Foundations Trends Theor. Comput. Sci.* **9**, 211–407, DOI: 10.1561/0400000042 (2013).

34. Dankar, F. K., Ibrahim, M. K. & Ismail, L. A Multi-Dimensional Evaluation of Synthetic Data Generators. *IEEE Access* **10**, 11147–11158, DOI: 10.1109/ACCESS.2022.3144765 (2022).

35. Liu, J., Qu, F., Hong, X. & Zhang, H. A Small-Sample Wind Turbine Fault Detection Method With Synthetic Fault Data Using Generative Adversarial Nets. *IEEE Transactions on Ind. Informatics* **15**, 3877–3888, DOI: 10.1109/TII.2018.2885365 (2019).

36. Alkurd, R., AbuAlhaol, I. & Yanikomeroglu, H. A Synthetic User Behavior Dataset Design for Data-Driven AI-Based Personalized Wireless Networks. In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 1–6, DOI: 10.1109/ICCW.2019.8756804 (IEEE, Shanghai, China, 2019).

37. Chundawat, V. S., Tarun, A. K., Mandal, M., Lahoti, M. & Narang, P. A Universal Metric for Robust Evaluation of Synthetic Tabular Data. *IEEE Transactions on Artif. Intell.* 1–11, DOI: 10.1109/TAI.2022.3229289 (2022).

38. Hyeong, J., Kim, J., Park, N. & Jajodia, S. An Empirical Study on the Membership Inference Attack against Tabular Data Synthesis Models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4064–4068, DOI: 10.1145/3511808.3557546 (ACM, Atlanta GA USA, 2022).

39. Reiner Benaim, A. *et al.* Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med. Informatics* **8**, e16492, DOI: 10.2196/16492 (2020).

40. Kaur, D. *et al.* Application of Bayesian networks to generate synthetic health data. *J. Am. Med. Informatics Assoc.* **28**, 801–811, DOI: 10.1093/jamia/ocaa303 (2021).

41. Yale, A. *et al.* Assessing privacy and quality of synthetic health data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, 1–4, DOI: 10.1145/3359115.3359124 (ACM, Pittsburgh Pennsylvania, 2019).

42. Javaid, U. *et al.* Blockchain based Secure Group Data Collaboration in Cloud with Differentially Private Synthetic Data and Trusted Execution Environment. In *2022 IEEE International Conference on Big Data (Big Data)*, 3919–3927, DOI: 10.1109/BigData55660.2022.10021011 (IEEE, Osaka, Japan, 2022).

43. Meeker, D., Kallem, C., Heras, Y., Garcia, S. & Thompson, C. Case report: Evaluation of an open-source synthetic data platform for simulation studies. *JAMIA Open* **5**, ooac067, DOI: 10.1093/jamiaopen/ooac067 (2022).

44. Mounir, M., Karsmakers, P. & Waterschoot, T. V. CNN-based Note Onset Detection using Synthetic Data Augmentation. In *2020 28th European Signal Processing Conference (EUSIPCO)*, 171–175, DOI: 10.23919/Eusipco47968.2020.9287621 (IEEE, Amsterdam, Netherlands, 2021).

45. Park, N. *et al.* Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.* **11**, 1071–1083, DOI: 10.14778/3231751.3231757 (2018).

46. Zhou, N., Wang, L., Marino, S., Zhao, Y. & Dinov, I. D. DataSifter II: Partially synthetic data sharing of sensitive information containing time-varying correlated observations. *J. Algorithms & Comput. Technol.* **16**, 174830262110653, DOI: 10.1177/17483026211065379 (2022).

47. Thomas, J. A. *et al.* Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: Results from analyzing >1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). Preprint, Health Informatics (2021). DOI: 10.1101/2021.07.06.21259051.

48. Thomas, J. A. *et al.* Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: Results from analyzing >1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). *J. Am. Med. Informatics Assoc.* **29**, 1350–1365, DOI: 10.1093/jamia/ocac045 (2022).

49. Tantipongpipat, U. T., Waites, C., Boob, D., Siva, A. A. & Cummings, R. Differentially Private Synthetic Mixed-Type Data Generation For Unsupervised Learning. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1–9, DOI: 10.1109/IISA52424.2021.9555521 (IEEE, Chania Crete, Greece, 2021).

50. Esmaili, A. & Farzi, S. Effective synthetic data generation for fake user detection. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 1–5, DOI: 10.1109/CSICC52343.2021.9420570 (IEEE, Tehran, Iran, 2021).

51. Hittmeir, M., Mayer, R. & Ekelhart, A. Efficient Bayesian Network Construction for Increased Privacy on Synthetic Data. In *2022 IEEE International Conference on Big Data (Big Data)*, 5721–5730, DOI: 10.1109/BigData55660.2022.10020936 (IEEE, Osaka, Japan, 2022).

52. Lu, P.-H., Wang, P.-C. & Yu, C.-M. Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, 1–6, DOI: 10.1145/3326467.3326474 (ACM, Seoul Republic of Korea, 2019).

53. Visani, G. *et al.* Enabling Synthetic Data adoption in regulated domains. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10, DOI: 10.1109/DSAA54385.2022.10032356 (IEEE, Shenzhen, China, 2022).

54. Ickin, S., Vandikas, K., Moradi, F., Taghia, J. & Hu, W. Ensemble-based Synthetic Data Synthesis for Federated QoE Modeling. In *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, 72–76, DOI: 10.1109/NetSoft48620.2020.9165379 (IEEE, Ghent, Belgium, 2020).

55. El Emam, K., Mosquera, L. & Bass, J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *J. Med. Internet Res.* **22**, e23139, DOI: 10.2196/23139 (2020).

56. El Emam, K., Mosquera, L., Jonker, E. & Sood, H. Evaluating the utility of synthetic COVID-19 case data. *JAMIA Open* **4**, ooab012, DOI: 10.1093/jamiaopen/ooab012 (2021).

57. Tai, B.-C., Li, S.-C., Huang, Y. & Wang, P.-C. Examining the Utility of Differentially Private Synthetic Data Generated using Variational Autoencoder with TensorFlow Privacy. In *2022 IEEE 27th Pacific Rim International Symposium on Dependable Computing (PRDC)*, 236–241, DOI: 10.1109/PRDC55274.2022.00038 (IEEE, Beijing, China, 2022).

58. Thorve, S., Vullikanti, A., Mortveit, H. S., Swarup, S. & Marathe, M. V. Fidelity and diversity metrics for validating hierarchical synthetic data: Application to residential energy demand. In *2022 IEEE International Conference on Big Data (Big Data)*, 1377–1382, DOI: 10.1109/BigData55660.2022.10020837 (IEEE, Osaka, Japan, 2022).

59. Flanagan, B., Majumdar, R. & Ogata, H. Fine Grain Synthetic Educational Data: Challenges and Limitations of Collaborative Learning Analytics. *IEEE Access* **10**, 26230–26241, DOI: 10.1109/ACCESS.2022.3156073 (2022).

60. Vaden, K. I., Gebregziabher, M., Dyslexia Data Consortium & Eckert, M. A. Fully synthetic neuroimaging data for replication and exploration. *NeuroImage* **223**, 117284, DOI: 10.1016/j.neuroimage.2020.117284 (2020).

61. Helfer, S., Kümmel, M., Bathelt, F. & Sedlmayr, M. Generating Enriched Synthetic German Hospital Claims Data – A Use Case Driven Approach. In Röhrig, R. *et al.* (eds.) *Studies in Health Technology and Informatics*, DOI: 10.3233/SHTI210051 (IOS Press, 2021).

62. Shi, J., Wang, D., Tesei, G. & Norgeot, B. Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments. *Front. Artif. Intell.* **5**, 918813, DOI: 10.3389/frai.2022.918813 (2022).

63. Tucker, A., Wang, Z., Rotalinti, Y. & Myles, P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digit. Medicine* **3**, 147, DOI: 10.1038/s41746-020-00353-9 (2020).

64. Smith, A., Lambert, P. C. & Rutherford, M. J. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med. Res. Methodol.* **22**, 176, DOI: 10.1186/s12874-022-01654-1 (2022).

65. Struye, J., Lemic, F. & Famaey, J. Generating Realistic Synthetic Head Rotation Data for Extended Reality using Deep Learning. In *Proceedings of the 1st Workshop on Interactive eXtended Reality*, 19–28, DOI: 10.1145/3552483.3556458 (ACM, Lisboa Portugal, 2022).

66. Strelcenia, E. & Prakoonwit, S. Generating Synthetic Data for Credit Card Fraud Detection Using GANs. In *2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT)*, 42–47, DOI: 10.1109/CAIT56099.2022.10072179 (IEEE, Quzhou, China, 2022).

67. Berke, A., Doorley, R., Larson, K. & Moro, E. Generating synthetic mobility data for a realistic population with RNNs to improve utility and privacy. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 964–967, DOI: 10.1145/3477314.3507230 (ACM, Virtual Event, 2022).

68. Ponge, J. *et al.* Generating Synthetic Populations Based On German Census Data. In *2021 Winter Simulation Conference (WSC)*, 1–12, DOI: 10.1109/WSC52266.2021.9715369 (IEEE, Phoenix, AZ, USA, 2021).

69. Yale, A. *et al.* Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* **416**, 244–255, DOI: 10.1016/j.neucom.2019.12.136 (2020).

70. Goncalves, A. *et al.* Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* **20**, 108, DOI: 10.1186/s12874-020-00977-1 (2020).

71. Zhang, C., Kuppannagari, S. R., Kannan, R. & Prasanna, V. K. Generative Adversarial Network for Synthetic Time Series Data Generation in Smart Grids. In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 1–6, DOI: 10.1109/SmartGridComm.2018.8587464 (IEEE, Aalborg, 2018).

72. Zheng, X., Wang, B., Kalathil, D. & Xie, L. Generative Adversarial Networks-Based Synthetic PMU Data Creation for Improved Event Classification. *IEEE Open Access J. Power Energy* **8**, 68–76, DOI: 10.1109/OAJPE.2021.3061648 (2021).

73. Gujar, S. *et al.* GenEthos: A Synthetic Data Generation System With Bias Detection And Mitigation. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–6, DOI: 10.1109/IC3SIS54991.2022.9885653 (IEEE, Kochi, India, 2022).

74. Acharya, A., Sikdar, S., Das, S. & Rangwala, H. GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources. In *2022 IEEE International Conference on Big Data (Big Data)*, 685–692, DOI: 10.1109/BigData55660.2022.10021001 (IEEE, Osaka, Japan, 2022).

75. Platzer, M. & Reutterer, T. Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data. *Front. Big Data* **4**, 679939, DOI: 10.3389/fdata.2021.679939 (2021).

76. Ge, C., Mohapatra, S., He, X. & Ilyas, I. F. Kamino: Constraint-aware differentially private data synthesis. *Proc. VLDB Endow.* **14**, 1886–1899, DOI: 10.14778/3467861.3467876 (2021).

77. Zhang, Z., Yan, C. & Malin, B. A. Keeping synthetic patients on track: Feedback mechanisms to mitigate performance drift in longitudinal health data simulation. *J. Am. Med. Informatics Assoc.* **29**, 1890–1898, DOI: 10.1093/jamia/ocac131 (2022).

78. Idehen, I., Jang, W. & Overbye, T. J. Large-Scale Generation and Validation of Synthetic PMU Data. *IEEE Transactions on Smart Grid* **11**, 4290–4298, DOI: 10.1109/TSG.2020.2977349 (2020).

79. Greenberg, J. K. *et al.* Leveraging Artificial Intelligence and Synthetic Data Derivatives for Spine Surgery Research. *Glob. Spine J.* 219256822210855, DOI: 10.1177/21925682221085535 (2022).

80. Jiang, Y., Mosquera, L., Jiang, B., Kong, L. & El Emam, K. Measuring re-identification risk using a synthetic estimator to enable data sharing. *PLOS ONE* **17**, e0269097, DOI: 10.1371/journal.pone.0269097 (2022).

81. Zhang, Z., Yan, C. & Malin, B. A. Membership inference attacks against synthetic health data. *J. Biomed. Informatics* **125**, 103977, DOI: 10.1016/j.jbi.2021.103977 (2022).

82. Iantovics, L. B. & Enăchescu, C. Method for Data Quality Assessment of Synthetic Industrial Data. *Sensors* **22**, 1608, DOI: 10.3390/s22041608 (2022).

83. Dietz, K., Gray, N., Seufert, M. & Hossfeld, T. ML-based Performance Prediction of SDN using Simulated Data from Real and Synthetic Networks. In *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*, 1–7, DOI: 10.1109/NOMS54207.2022.9789916 (IEEE, Budapest, Hungary, 2022).

84. Shouryadhar, K., Kiran Rao, P. & Chatterjee, S. Multilevel Ensemble Method to Identify Risks in Chronic Kidney Disease Using Hybrid Synthetic Data. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–6, DOI: 10.1109/ICCCNT54827.2022.9984346 (IEEE, Kharagpur, India, 2022).

85. Hittmeir, M., Ekelhart, A. & Mayer, R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 1–6, DOI: 10.1145/3339252.3339281 (ACM, Canterbury CA United Kingdom, 2019).

86. Fang, K., Mugunthan, V., Ramkumar, V. & Kagal, L. Overcoming Challenges of Synthetic Data Generation. In *2022 IEEE International Conference on Big Data (Big Data)*, 262–270, DOI: 10.1109/BigData55660.2022.10020479 (IEEE, Osaka, Japan, 2022).

87. Bird, J. J., Faria, D. R., Premebida, C., Ekart, A. & Ayrosa, P. P. S. Overcoming Data Scarcity in Speaker Identification: Dataset Augmentation with Synthetic MFCCs via Character-level RNN. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 146–151, DOI: 10.1109/ICARSC49921.2020.9096166 (IEEE, Ponta Delgada, Portugal, 2020).

88. Mosquera, L. PCN429 THE GENERATION OF SYNTHETIC CLINICAL TRIAL DATA. *Value Heal.* **22**, S519, DOI: 10.1016/j.jval.2019.09.622 (2019).

89. M, G. H., Shenoy, P. D. & R, V. K. Performance Analysis of Real and Synthetic Data using Supervised ML Algorithms for Prediction of Chronic Kidney Disease. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 1–6, DOI: 10.1109/CONECCT55679.2022.9865722 (IEEE, Bangalore, India, 2022).

90. Rashidi, H. H. *et al.* Prediction of tuberculosis using an automated machine learning platform for models trained on synthetic data. *J. Pathol. Informatics* **13**, 100172, DOI: 10.4103/jpi.jpi_75_21 (2022).

91. Zhang, F. & Zhang, D. Privacy-aware synthesis of sensing data based on learning model at metropolitan scale: Poster abstract. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 428–429, DOI: 10.1145/3356250.3361957 (ACM, New York New York, 2019).

92. Liu, F. *et al.* Privacy-Preserving Synthetic Data Generation for Recommendation Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1379–1389, DOI: 10.1145/3477495.3532044 (ACM, Madrid Spain, 2022).

93. Esser, A. & Rinderknecht, S. Process for the Validation of Using Synthetic Driving Cycles Based on Naturalistic Driving Data Sets. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–6, DOI: 10.1109/ITSC45102.2020.9294369 (IEEE, Rhodes, Greece, 2020).

94. Fan, J. *et al.* Relational data synthesis using generative adversarial networks: A design space exploration. *Proc. VLDB Endow.* **13**, 1962–1975, DOI: 10.14778/3407790.3407802 (2020).

95. Rankin, D. *et al.* Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Med. Informatics* **8**, e18910, DOI: 10.2196/18910 (2020).

96. Varma, G., Chauhan, R. & Singh, D. Sarve: Synthetic data and local differential privacy for private frequency estimation. *Cybersecurity* **5**, 26, DOI: 10.1186/s42400-022-00129-6 (2022).

97. El Emam, K. Seven Ways to Evaluate the Utility of Synthetic Data. *IEEE Secur. & Priv.* **18**, 56–59, DOI: 10.1109/MSEC.2020.2992821 (2020).

98. Foraker, R. E. *et al.* Spot the difference: Comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* **3**, 557–566, DOI: 10.1093/jamiaopen/ooaa060 (2021).

99. Kothare, A., Chaube, S., Moharir, Y., Bajodia, G. & Dongre, S. SynGen: Synthetic Data Generation. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, 1–4, DOI: 10.1109/ICCICA52458.2021.9697232 (IEEE, Nagpur, India, 2021).

100. Benarous, M., Toch, E. & Ben-gal, I. Synthesis of Longitudinal Human Location Sequences: Balancing Utility and Privacy. *ACM Transactions on Knowl. Discov. from Data* **16**, 1–27, DOI: 10.1145/3529260 (2022).

101. Imtiaz, S., Arsalan, M., Vlassov, V. & Sadre, R. Synthetic and Private Smart Health Care Data Generation using GANs. In *2021 International Conference on Computer Communications and Networks (ICCCN)*, 1–7, DOI: 10.1109/ICCCN52240.2021.9522203 (IEEE, Athens, Greece, 2021).

102. Yue, Y., Li, Y., Yi, K. & Wu, Z. Synthetic Data Approach for Classification and Regression. In *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, 1–8, DOI: 10.1109/ASAP.2018.8445094 (IEEE, Milan, 2018).

103. Wilchek, M. & Wang, Y. Synthetic Differential Privacy Data Generation for Revealing Bias Modelling Risks. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 1574–1580, DOI: 10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00211 (IEEE, New York City, NY, USA, 2021).

104. Ooko, S. O., Mukanyiligira, D., Munyampundu, J. P. & Nsenga, J. Synthetic Exhaled Breath Data-Based Edge AI Model for the Prediction of Chronic Obstructive Pulmonary Disease. In *2021 International Conference on Computing and Communications Applications and Technologies (I3CAT)*, 1–6, DOI: 10.1109/I3CAT53310.2021.9629420 (IEEE, Ipswich, United Kingdom, 2021).

105. Nußberger, J., Boesel, F., Lenz, S., Binder, H. & Hess, M. Synthetic observations from deep generative models and binary omics data with limited sample size. *Briefings Bioinforma.* **22**, bbaa226, DOI: 10.1093/bib/bbaa226 (2021).

106. Behjati, R., Arisholm, E., Bedregal, M. & Tan, C. Synthetic Test Data Generation Using Recurrent Neural Networks: A Position Paper. In *2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)*, 22–27, DOI: 10.1109/RAISE.2019.00012 (IEEE, Montreal, QC, Canada, 2019).

107. Kiran Rao, P. & Chatterjee, S. TabNet to Identify Risks in Chronic Kidney Disease Using GAN's Synthetic Data. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 209–215, DOI: 10.1109/ICTACS56270.2022.9988284 (IEEE, Tashkent, Uzbekistan, 2022).

108. Bhanot, K., Qi, M., Erickson, J. S., Guyon, I. & Bennett, K. P. The Problem of Fairness in Synthetic Healthcare Data. *Entropy* **23**, 1165, DOI: 10.3390/e23091165 (2021).

109. Guo, A. *et al.* The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation. *Front. Digit. Heal.* **2**, 576945, DOI: 10.3389/fdgth.2020.576945 (2020).

110. Holmes, M. & Theodorakopoulos, G. Towards using differentially private synthetic data for machine learning in collaborative data science projects. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 1–6, DOI: 10.1145/3407023.3407024 (ACM, Virtual Event Ireland, 2020).

111. Quick, H. & Waller, L. A. Using spatiotemporal models to generate synthetic data for public use. *Spatial Spatio-temporal Epidemiol.* **27**, 37–45, DOI: 10.1016/j.sste.2018.08.004 (2018).

112. Nabati, M., Navidan, H., Shahbazian, R., Ghorashi, S. A. & Windridge, D. Using Synthetic Data to Enhance the Accuracy of Fingerprint-Based Localization: A Deep Learning Approach. *IEEE Sensors Lett.* **4**, 1–4, DOI: 10.1109/LSENS.2020.2971555 (2020).

113. Grund, S., Lüdtke, O. & Robitzsch, A. Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychol. Methods* DOI: 10.1037/met0000526 (2022).

114. Resnick, D. M., Cox, C. S. & Mirel, L. B. Using synthetic data to replace linkage derived elements: A case study. *Heal. Serv. Outcomes Res. Methodol.* **21**, 389–406, DOI: 10.1007/s10742-021-00241-z (2021).

115. Hittmeir, M., Ekelhart, A. & Mayer, R. Utility and Privacy Assessments of Synthetic Data for Regression Tasks. In *2019 IEEE International Conference on Big Data (Big Data)*, 5763–5772, DOI: 10.1109/BigData47090.2019.9005476 (IEEE, Los Angeles, CA, USA, 2019).

116. El Emam, K., Mosquera, L., Fang, X. & El-Hussuna, A. Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. *JMIR Med. Informatics* **10**, e35734, DOI: 10.2196/35734 (2022).

117. El Emam, K., Mosquera, L. & Fang, X. Validating a membership disclosure metric for synthetic health data. *JAMIA Open* **5**, ooac083, DOI: 10.1093/jamiaopen/ooac083 (2022).

118. Razghandi, M., Zhou, H., Erol-Kantarci, M. & Turgut, D. Variational Autoencoder Generative Adversarial Network for Synthetic Data Generation in Smart Home. In *ICC 2022 - IEEE International Conference on Communications*, 4781–4786, DOI: 10.1109/ICC45855.2022.9839249 (IEEE, Seoul, Korea, Republic of, 2022).

119. Giomi, M., Boenisch, F., Wehmeyer, C. & Tasnádi, B. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *Proc. on Priv. Enhancing Technol.* (2023).

120. Jordon, J. *et al.* Synthetic Data - what, why and how? .

121. Synthetic Data.

122. Funding & tenders. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-ju-ihi-2023-05-04?tenders=false&programmePart=&callIdentifier=HORIZON-JU-IHI-2023-05.

Author contributions statement

B.K., J.D. and J.L.R. conceived the scoping review design and objectives. B.K. conducted database searches and screened potential articles for inclusion. J.L.R., T.M. and F.P. provided methodological guidance and critically reviewed the protocol. T.M., K.O., M.H and F.P. assisted in interpreting the findings and shaping the discussion. All authors collaborated in structuring the manuscript's narrative, B.K. wrote the manuscript and all authors read, edited, and approved the final manuscript.

Additional information

Competing interests: The authors declare no competing interests.

Table 2. Scoping Review Results

	Utility					Privacy	
	Univariate Similarity	Bivariate Similarity	Multivariate Similarity	Longitudinal Similarity	Domain Specific Similarity	Model Evaluation	Dataset Evaluation
34	X	X	X				
35	X	X		X			
36				X			
37	X	X	X				
38							X
39					X		
40	X	X	X				
41	X						X
42			X		X		
12	X	X	X		X		
43					X		
44				X			
45	X		X			X	X
46				X			X
47	X			X			X
48	X			X			
49	X		X				
50			X				
51		X	X				X
52		X	X				X
53	X	X	X				X
54	X		X				
55							X
56			X		X		X
57	X		X				
58			X				
59			X				
60	X	X					
10	X	X	X				X
61		X					
62	X	X					X
63	X		X				X
64	X				X		
65	X		X	X			
66			X				
13					X		
67	X				X		X
68	X						
69	X		X		X		X
32	X		X			X	
70	X	X	X				X
71			X	X			
72				X			
73	X						
74	X	X	X				
75	X						X

Table 3. Scoping Review Results

	Utility					Privacy	
	Univariate Similarity	Bivariate Similarity	Multivariate Similarity	Longitudinal Similarity	Domain Specific Similarity	Model Evaluation	Dataset Evaluation
76	X		X				
77			X		X		
78				X	X		
79	X		X				
80							X
81							X
31			X			X	
82					X		
83				X			
84			X				
85	X	X	X				X
86	X	X	X				X
87				X			
88	X		X				
89			X				
90	X		X				
91	X						
92			X				X
30						X	X
93					X		
94			X				X
95		X	X				
96	X						
97					X		
11	X		X				
98	X		X				
99			X				
100	X		X				X
101	X						
102			X				
103		X	X				
104			X				X
105		X					
106	X						
107	X		X				
108		X		X			
109	X		X	X			
110	X		X				
111					X		
112			X				
113			X				X
114	X				X		
115		X	X				X
116	X		X				
117							X
118	X		X				

Appendices

A Database Search Strategy

Table 4. Queries by database

Database	Query
IEEExplore	("Document Title":synthetic data) AND ("Abstract":utility OR "Abstract":privacy OR "Abstract":evaluation OR "Abstract":metric)
ACM DL	Abstract:(utility OR privacy OR evaluation OR metric) AND Title:(synthetic AND data)
PubMed	synthetic[Title] AND data[Title]) AND (utility[Title/Abstract] OR privacy[Title/Abstract] OR evaluation[Title/Abstract] OR metric[Title/Abstract]
Embase	synthetic:ti AND data:ti AND (utility:ab OR privacy:ab OR evaluation:ab OR metric:ab) AND [2018-2022]/py

Table 5. Data items used in full-text charting

Title	Description	Possible Values
DOI	Digital Object Identifier	Free text
Document Title	Title of publication	Free text
Authors	First Author of publication	Free text
Publication Year	Year of publication	[2018..2022]
Database	Database or retrieval tool	[IEEExplore, ACM, PubMed, Embase]
Broad Utility Metric	General Utility Metric category	Values in Figure 2.
Utility Metric	Specific Utility Metric used	Values in Figure 2.
Broad Privacy Metric	General Privacy Metric category	Values in Figure 3.
Privacy Metric	Specific Privacy Metric used	Values in Figure 3.
Privacy Type	Type of privacy involved	[Membership Inference, Attribute inference]
Additional Noise Layer	Use of differential privacy and/or added noise to the output	[Y,N]
Adversary Knowledge	Knowledge of adversary	[Full Knowledge, Partial Knowledge]
SDG Method	Synthetic Data Generation Method used	Free text

B Utility Evaluation Methods

- **Cumulative Distributions Visual Comparison**: Evaluates the visual similarity between the cumulative distribution functions of synthetic and original datasets.
- **Marginal Distributions Visual Comparison**: Compares the marginal distributions of individual variables in synthetic and original datasets through visual inspection.
- **Descriptive Statistics Comparison**: Measures the agreement between summary statistics such as mean, median, and standard deviation for synthetic and original datasets.
- **Kolmogorov-Smirnov Test**: Uses the Kolmogorov-Smirnov test to statistically assess the difference between the empirical distribution functions of synthetic and original datasets.
- **Chi-Squared Test**: Utilizes the Chi-squared test to examine if synthetic and original datasets differ significantly in terms of their categorical variables.
- **Wilcoxon Signed-Rank Test**: Applies the Wilcoxon Signed-Rank test to compare two related samples, in this case, synthetic and original datasets, to assess whether their population mean ranks differ.
- **T-Test**: Uses the T-Test to compare the means of synthetic and original datasets and assess if they come from populations with equal means.
- **Kullback-Leibler Divergence**: Quantifies how much one distribution diverges from another, measuring the difference between synthetic and original datasets.
- **Wasserstein Distance**: Utilizes the Wasserstein distance metric to quantify the dissimilarity between the synthetic and original distributions.
- **Hellinger Distance**: Measures the Hellinger distance to evaluate the similarity between the synthetic and original datasets' distributions.
- **Distance Between Probabilities**: Calculates the difference between probabilities associated with various states or events in synthetic and original datasets.
- **Correlation Coefficient**: Quantifies how strongly pairs of variables in the synthetic and original datasets are linearly related.
- **Log Odds Ratio**: Measures the log odds ratio to evaluate associations between categorical variables in synthetic and original datasets.
- **Maximum Information Coefficient**: Utilizes the Maximum Information Coefficient to capture a wide range of associations between variables.
- **Tau Statistic**: Applies the Tau statistic to assess the strength of the relationship between two variables in synthetic and original datasets.
- **Visual Comparison of Correlation Matrices**: Visually compares the correlation matrices of synthetic and original datasets to assess bivariate similarity.
- **PCA Visual Comparison**: Employs Principal Component Analysis (PCA) for a visual comparison of the main components in synthetic and original datasets.
- **Cluster Analysis**: Uses cluster analysis to evaluate how closely the synthetic dataset replicates the natural groupings present in the original dataset.
- **ML Classification Performance**: Evaluates the performance of machine learning classification models trained on synthetic data.
- **ML Regression Performance**: Measures the performance of machine learning regression models when trained on synthetic data.

- **Sensitivity Analysis**: Conducts a sensitivity analysis to evaluate how small changes in the synthetic dataset affect outcomes.
- **Maximum Mean Discrepancy**: Utilizes Maximum Mean Discrepancy to measure the difference between the synthetic and original datasets' distributions.
- **Replication of Studies**: Assesses the utility of synthetic data by attempting to replicate the findings of studies based on the original dataset.
- **Domain Expert Assessment**: Involves a domain expert's qualitative assessment to validate the utility of synthetic data.
- **Distinguishability Performance**: Measures how well the synthetic dataset can be distinguished from the original dataset.
- **Stability Assessment**: Evaluates the stability of conclusions drawn from synthetic data when subjected to perturbations.
- **Comparison with Public Data**: Compares the synthetic dataset with publicly available data in the same domain to assess its utility.
- **Comparison with Other PETs**: Compares the utility of synthetic data to other Privacy-Enhancing Technologies (PETs).
- **Structural Similarity**: Measures the similarity in structural attributes between the synthetic and original datasets.
- **Cross-Correlation**: Measures the relationship between two time-series data sets.
- **Correlation Coefficient**: Quantifies how strongly time-dependent variables in the synthetic dataset correlate with those in the original dataset.
- **Visual Inspection of Distances**: Uses visual methods to compare the distances between elements in the synthetic and original time-series datasets.
- **Modal Property Comparison**: Compares the properties of modes in both the synthetic and original time-series datasets.
- **Directional Symmetry**: Assesses whether the synthetic data maintains the same directional changes over time as the original data.
- **ML Forecasting Performance Comparison**: Compares the performance of machine learning forecasting models trained on synthetic versus original time-series data.
- **ML Classification Performance**: Measures the performance of machine learning classifiers when trained on synthetic time-series data versus original time-series data.
- **Domain Specific Metric**: Utilizes a customized metric particularly relevant to the specific field.
- **Total Absolute Error**: Measures the total absolute error between the synthetic and original datasets.
- **Frobenius Norm**: Uses the Frobenius norm to measure the difference between the synthetic and original datasets.
- **Visual Comparison of Association Matrices**: Uses visual methods to compare association matrices derived from the synthetic and original datasets.
- **Mean Square Error (MSE)**: Measures the mean square error between the synthetic and original datasets.
- **Cross-Correlation**: Measures the relationship between two time-series data sets.
- **Correlation Coefficient**: Quantifies how strongly time-dependent variables in the synthetic dataset correlate with those in the original dataset.
- **Visual Inspection of Distances**: Uses visual methods to compare the distances between elements in the synthetic and original time-series datasets.
- **Modal Property Comparison**: Compares the properties of modes in both the synthetic and original time-series datasets.
- **Directional Symmetry**: Assesses whether the synthetic data maintains the same directional changes over time as the original data.

- **ML Forecasting Performance Comparison**: Compares the performance of machine learning forecasting models trained on synthetic versus original time-series data.
- **ML Classification Performance**: Measures the performance of machine learning classifiers when trained on synthetic time-series data versus original time-series data.
- **Domain Specific Metric**: Utilizes a customized metric particularly relevant to the specific field.
- **Total Absolute Error**: Measures the total absolute error between the synthetic and original datasets.
- **Frobenius Norm**: Uses the Frobenius norm to measure the difference between the synthetic and original datasets.
- **Visual Comparison of Association Matrices**: Uses visual methods to compare association matrices derived from the synthetic and original datasets.
- **Mean Square Error (MSE)**: Measures the mean square error between the synthetic and original datasets.

C Privacy Evaluation Methods

- **Exact Match**: Identifies exact matches between synthetic and real data records, often referred to as the hit rate. This method assesses the risk of individual record re-identification, effectively acting as a direct measure of data leakage.
- **Shadow Models**: Involves the generation of multiple models that replicate the behavior of the primary synthetic data model. While this method can be computationally expensive, it creates a robust evaluation framework that mimics a black-box attack scenario, thereby offering a comprehensive privacy risk assessment.
- **Classification/Regression Task**: Utilizes machine learning models trained to either classify or regress on attributes of the synthetic data. Upon training, these models are subsequently evaluated on real-world data to gauge how well their predictions generalize, serving as an indirect measure of the privacy level of the synthetic data.
- **Discriminator Likelihood Attack**: This specialized technique targets Generative Adversarial Networks (GANs) by relying on the characteristics of the discriminator component. The focus is on evaluating how well the discriminator distinguishes between real and synthetic data, thereby serving as a proxy for privacy risk.
- **CRLProxy**: Zhang et al.[81] adopt contrastive representation learning approach, supplemented with proxy-based augmentations, to shift the synthetic model's focus from weak-level to strong-level features.
- **Probabilistic Disclosure Risk Assessment**: Trains an estimator of re-identification probability that is based on synthetic data generation methods. Jiang et al.[80] for example use this metric to give the probability that a random record selected from a microdata sample can be correctly matched to a record (or individual) in the population from which the sample comes from. Zhou et al.[46] compute the statistical disclosure risk for every time point in a longitudinal record.
- **Gradient Norm Attack**: Leverages the gradient norms of synthetic data models as an attack vector to exploit potential overfitting vulnerabilities. The intention behind this method is to expose weak spots where the synthetic data might reveal too much about the original dataset, thereby compromising privacy. Notable example of its implementations can be seen in the works of Del Grosso et al.[30].
- **Distance to Real Data**: This method calculates the mathematical distance between synthetic and actual data points.
- **Holdout Set Distance**: Extends the distance measurement by incorporating a holdout set—data not used during the training process.

D Additional Results

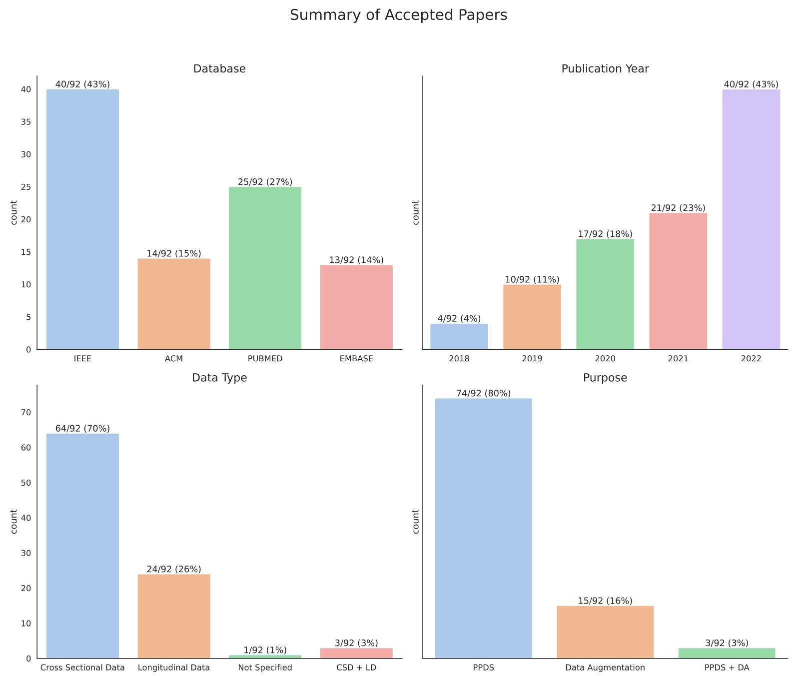

Figure 9. Visual overview of included papers across various metrics. The figure depicts four dimensions— Database, Data Type, Purpose, and Publication Year. PPDS refers to Privacy Preserving Data Sharing while IEEE refers to IEEExplore database.