medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

A Scoping Review of Privacy and Utility Metrics in Medical Synthetic Data

Bayrem Kaabachi[1,*], Jérémie Despraz[1], Thierry Meurers[2], Karen Otte[2], Mehmed Halilovic[2], Bogdan Kulynych[1], 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

The use of synthetic data is a promising solution to facilitate the sharing and reuse of health-related data beyond its initial collection while addressing privacy concerns. However, there is still no consensus on a standardized approach for systematically evaluating the privacy and utility of synthetic data, impeding its broader adoption. In this work, we present a comprehensive review and systematization of current methods for evaluating synthetic health-related data, focusing on both privacy and utility aspects. Our findings suggest that there are a variety of methods for assessing the utility of synthetic data, but no consensus on which method is optimal in which scenario. Moreover, we found that most studies included in this review do not evaluate the privacy protection provided by synthetic data, and those that do often significantly underestimate the risks.

Introduction

Access to high-quality data plays a crucial role in medical research and practice, particularly with the growing integration of Artificial Intelligence (AI) and Machine Learning (ML). These technologies contribute to advancements in areas like precision medicine[1], where personalized treatments depend on comprehensive and diverse datasets. Thus, establishing safe and reliable procedures for secondary data access is important to ensure these innovations are applied ethically, securely, and effectively.

Due to privacy concerns, however, 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 re-identification risk threshold[5]. This approach employs data masking and transformation techniques to reduce privacy risks. Nonetheless, even in cases where a sufficient protection level can be achieved, anonymizing high-dimensional data often comes with a severe deterioration of the utility of the anonymized dataset,[6] which can render it nearly unusable for research in the worst case.

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, including recent advancements such as *Generative Adversarial Networks*[8] (GANs), *diffusion models*[9], and *Large Language Models*[10] (LLMs). These new methods generate sample that closely resemble real data, which could reduce privacy risks compared to the direct sharing of original data, and increase utility compared to anonymization.

In the medical domain in particular, several studies[11–13] have used synthetic data to replicate case studies originally performed on real 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.

Although these results seem promising for the future of privacy-preserving data sharing in medical environments, more recent studies have pointed out risks associated with over-reliance on synthetic data as a "silver bullet" solution[14]. In particular, a malicious adversary could infer information about presence or absence of certain records in the original data, as well as infer values of sensitive attributes of known records by having access to the procedure for generating the synthetic data.[15] This is due to the tendency of machine learning and statistical models to overfit on their training data and memorize information about individuals in the dataset[16]. Moreover, due to the black-box nature of most synthetic data generation methods such as GANs, it is difficult to predict which useful information 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 needs to be taken

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

when generating and sharing synthetic data.

The potential risks associated with synthetic data usage highlighted in recent studies[14,17,18] raise the question of whether research priorities exhibit a stronger emphasis on utility over privacy considerations. Compared to anonymized data, for which there is extensive literature[19] describing different kinds of attacks and the corresponding privacy protection mechanisms, synthetic data has not yet been as thoroughly scrutinized. This prompted us to conduct this review in the hope of providing an informed and unbiased answer to that question.

A few surveys in the field have examined various aspects of synthetic data generation.[20,21] Figueira et al.[20] provided an extensive description of multiple generation methods, and Hernandez et al.[21] 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 pressing issues with synthetic data as we place a greater emphasis on the evaluation process and the privacy-utility trade-offs by having a systematic look at how synthetic data is evaluated across 73 studies. In a concurrent work, Vallevik et al.[22] propose a taxonomy that is similar to ours in terms of fidelity, utility, and fairness. Our work, however, offers a different approach to privacy by conducting a critical analysis and comparing to the work in the Computer Science literature. We thus reach different conclusions, as we show next.

A recent series of open-source solutions such as Synthetic Data Vault,[23] Table Evaluator,[24] synthcity[25] and TAPAS[15] enable researchers to 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[14,17,26] in the literature impedes the development of a unified understanding of synthetic data in healthcare.

To get a better understanding of the current landscape in healthcare-related synthetic data generation, we initiated this scoping review specifically targeting evaluation methodologies, aiming to provide a rigorous and quantitative analysis of the suitability of synthetic data evaluation methods. To do so, we have structured our analysis around answering the following two research questions:

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

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

Synthetic medical data aims to protect patient privacy while retaining useful information. Our investigation cuts to the heart of the matter: Can practitioners trust this data to protect patient privacy and accelerate healthcare research? By investigating these two research questions, we expose the pitfalls, and provide recommendations for trustworthy synthetic data in medicine.

Results

In this review, after reconciling methods that were semantically the same, we found that there were 17 methods used to assess utility and 5 methods used to assess privacy. Fig. 2 gives an overview of the overall landscape of utility and privacy evaluation methods used in all the publications we selected. We include the full results of the scoping review as supplementary materials.

We reviewed articles published from 2018 to July 2024, a period that saw the rise of generative AI technologies, including the early enthusiasm in GANs and the adoption of LLMs. This growing interest is evident in our corpus as we have only two eligible publications in 2018–2019, and 21 in 2023. See Fig. 1 for details.

Additionally, we found that most articles used cross-sectional data, making up 73% (53/73). Only 25% (18/73) used temporal longitudinal data, possibly as it is 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 shown in Fig. 5, unstructured data such as images or text were not considered during this review.

We found that the privacy aspect of synthetic data was the main incentive behind most selected papers, with 82% (60/73) intending to use synthetic data for private data sharing scenarios. The other 8% (6/73) used it for data augmentation purposes and to answer either data scarcity or class imbalance problems. The remaining 10% (7/73) studied the potential of synthetic data in both scenarios.

Different methods were used to create synthetic data. As we show in Fig. 2, out of all 49 synthetic data generation methods used in our corpus, 33% (16/49) are GANs. The rest, 67% (33/49), are a mix of other methods, including statistical modeling and methods implemented by specialized software such as Synthpop[28] R package or the MDClone[29] commercial platform.

Our findings indicate that the current landscape lacks a unified approach, as we identified 49 different ways to refer to utility and fairness metrics, and 22 different ways to discuss privacy which complicates the comparison and synthesis of existing evidence. We document the variability of those metrics in Supplementary Figure 1. By applying the taxonomy we proposed in Table 3 we were able to derive a trend towards *broad utility* evaluations which was noted in 153 instances (by an instance, we

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

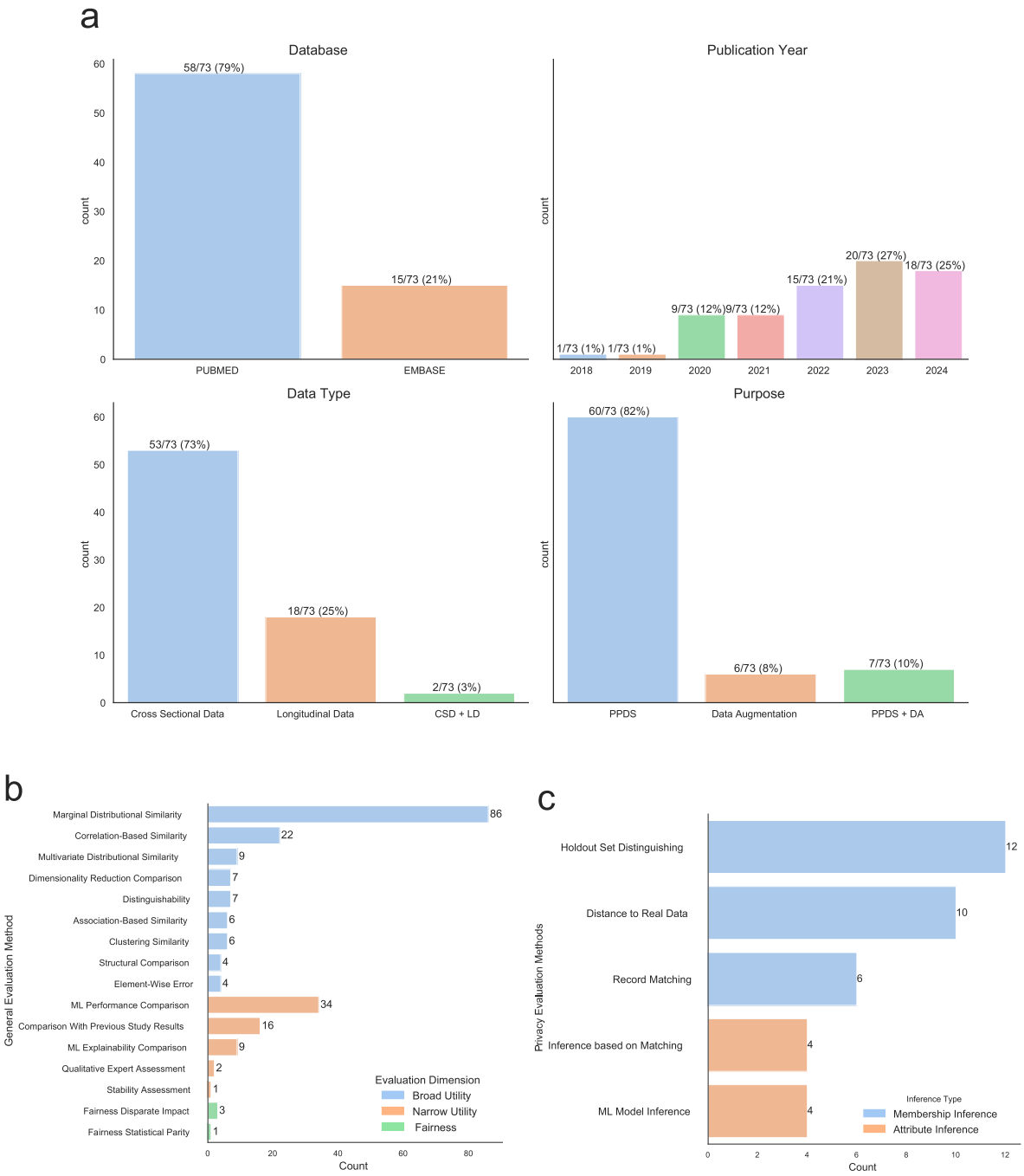

Figure 1. Scoping Review Results.
(a) Visual overview of included works across various metrics. The figure depicts four dimensions: Database, Data Type, Purpose, and Publication Year. PPDS refers to Privacy Preserving Data Sharing. **(b)** Summary of the performance-related methods used in the included works. This includes a breakdown of categories such as Broad Utility, Narrow Utility and Fairness. **(c)** Summary of the performance-related methods used in the included papers. We categorized methods as Membership Inference or Attribute Inference.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

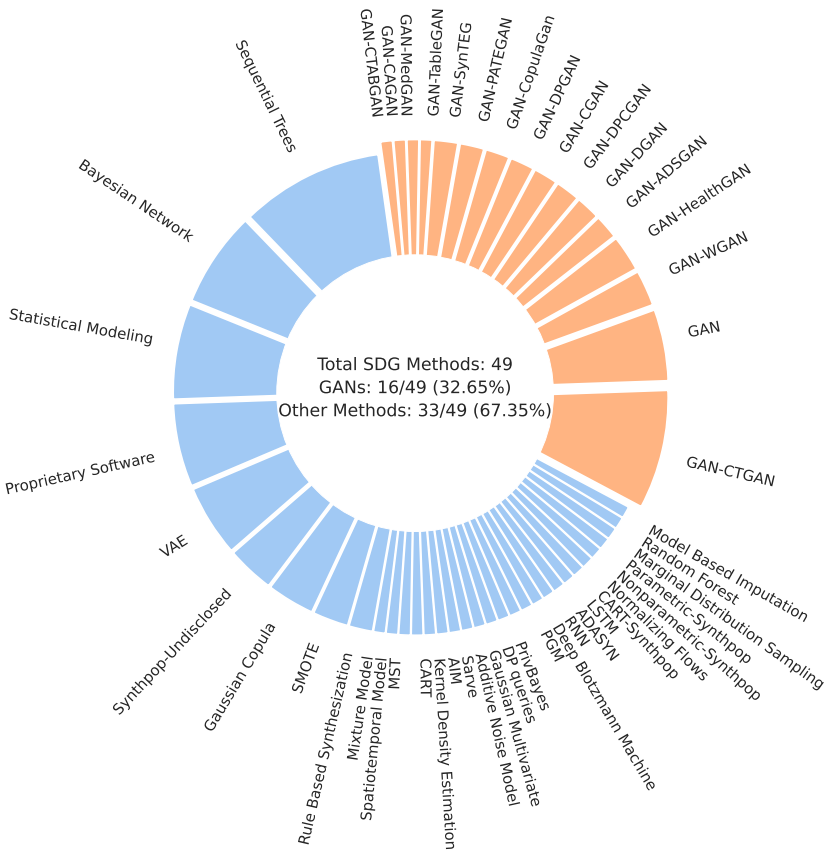

Figure 2. Synthetic data generation methods.
We found 49 different synthetic data generation (SDG) methods, split into two main categories: GANs (Generative Adversarial Networks) and other techniques. GANs, shown in orange, make up 32.65% of the total, with 16 different methods. The remaining 67.35%, in blue, includes various approaches like Bayesian Networks, VAEs, and proprietary software.

refer to evaluation of a specific metric, e.g., one paper can evaluate multiple metrics which all are classified as *broad utility*). *Narrow utility* is represented in 63 instances, whereas *fairness* is significantly less represented with only three instances of use. Among the works that evaluated the privacy risks of synthetic data, *membership inference* risk was the most common type, appearing in 28 instances, whereas *attribute inference* appeared in 9 instances. The specific methods used for privacy evaluation varied: 12 instances involved *holdout set distinguishing*, nine used *distance to real data*, seven employed *record matching*, five relied on *inference based on matching*, and four utilized *ML model inference*. Another notable finding is that privacy evaluations are not as often employed as utility evaluations. 95% of the studies (70/73) included utility evaluations while only 46% (31/67) of the studies claiming to employ synthetic data for preserving privacy, i.e., those that should evaluate privacy, conducted any privacy evaluation. We found that most of the studies have utilized synthetic data "as is", assuming inherent privacy benefits without empirical verification.

Figure 3. Number of works that evaluate privacy and use methods that provide privacy guarantees.
Gap between the intended privacy focus of studies and the actual privacy evaluation. Only 46% (31/67) of the studies claiming to employ synthetic data for preserving privacy conducted any privacy evaluation.

In the next section, we provide a discussion of salient issues that we have identified during the analysis of these research questions, and propose concrete steps forward to rectify these issues.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

The proposed taxonomy enables practitioners and researchers to mitigate the issue of the lack of consensus by ensuring a comprehensive evaluation within all dimensions from Table 3, covering *broad utility, narrow utility* (if synthetic data is released for a specific task), *fairness,* and *privacy.* For instance, some works in our corpus have evaluated synthetic data using multiple metrics within the same category, e.g., broad utility, yet used no metrics in other categories. Evaluating synthetic data generators or the released synthetic data across all of these dimensions provides a clearer picture of their trustworthiness.

We found that the privacy aspect of synthetic data evaluation has mostly revolved around using similarity-based metrics. It is notable that some *privacy* evaluation methods, such as *distance to real data*, can be directly at odds with equivalent metrics used for evaluating *utility*. Synthetic data is sometimes evaluated using these similarity-based metrics for both its privacy and utility even within the same study,[30] which can lead to conflicting results and complicate the interpretation of the privacy-utility trade-off. This dichotomy highlights a challenge in harmonizing the definitions of privacy and utility in synthetic data evaluation.

The fact that most works that use synthetic data for the purpose of preserving privacy do not evaluate the residual privacy risks (see Fig. 3) poses significant concern, especially with public synthetic data releases. Practitioners may inadvertently assume that synthetic data they are generating is privacy-preserving by default. This may lead to the uninformed sharing of sensitive data, potentially resulting in data breaches in addition to ethical and legal complications.

Moreover, most of the studies that evaluate privacy, have employed similarity-based metrics. The prior work has recently argued and empirically demonstrated that the reliance on similarity-based metrics for privacy evaluation is inadequate for two reasons. First, such metrics do not reflect the privacy guarantees faithfully, e.g., there can exist successful inference attacks even if synthetic data is dissimilar from the original dataset.[14,31] Second, the publication of similarity-based metrics on its own can lead to novel privacy risks such as reconstruction attacks that leverage the reported metrics.[31] The popularity of similarity-based metrics in our review suggests that many evaluations may offer a false sense of security regarding the privacy-preserving capabilities of synthetic data.[14] This contrasts sharply with more sophisticated attacks discussed in the Computer Science literature, such as shadow model attacks,[14,15] which employ advanced techniques to assess privacy risks in a more principled way.

To ensure privacy, 11% (8/73) of the reviewed works have used differentially private[32] synthetic data generators. Differential privacy (DP) is a well-established principled approach to ensuring provable privacy guarantees through controlled injection of random noise in the process of building the generative model. Although DP provides strong guarantees, they are significantly stronger than what is necessarily needed for practical privacy protection, which results in DP oftentimes significantly hurting utility, consistency, and fairness.[33–35] Recent years, however, have seen significant progress in making the DP methods, including synthetic data generation, effective and feasible. In particular, there exists a family of works for generating synthetic data based on k-way marginals with provable guarantees both in terms of privacy and utility.[36–38] Such methods were recently showed to be superior in terms of utility and fairness[39] compared to other private methods based on GANs, and even, in some cases, to non-private methods.[15] Despite this, such methods see almost no usage in the corpus we have reviewed compared to, e.g., less efficient methods based on GANs.

The level of privacy in DP is usually parameterized by the parameter ε, which is often criticized as non-interpretable to the practitioners.[33] Fortunately, recent works provide operational interpretations for the level of privacy provided by DP, e.g., via success of reconstruction attacks,[40] and argue that even if the formal privacy guarantees are weak, in practice, DP methods still provide strong resilience against practical inference attacks.[41]

Even though DP provides privacy guarantees in theory, recent studies show that practical implementations violate these guarantees due to software bugs or improper usage,[14,42] with a recent line of works being developed specifically to auditing the privacy guarantees afforded by DP methods.[43] Therefore, even with theoretical guarantees, it is still important to evaluate privacy in DP synthetic data generation.

In conclusion, this review offers a detailed insight into the present research landscape of synthetic health data's utility and privacy. The need for standardized evaluation measures stands out as a major takeaway where we believe 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. This need is increasingly apparent as international initiatives such as IEEE's Industry Connections activity[44] and Horizon Europe's call for synthetic data[45] confirm the urgency of creating clear guidelines for the development of reliable frameworks in the field. Our intention with this review is to not only shed light on these challenges, but also to inspire a collaborative effort in formulating best practices that make these techniques more accessible and understandable.

One significant concern raised throughout our work is the need for robust privacy evaluations. As the healthcare sector houses sensitive information, ensuring that synthetic data does not inadvertently lead to data leaks or result in a loss of trust is crucial. This is especially true when it comes to generative models such as GANs as their inherent complexity and lack of transparency can lead to misinformed usage where without a proper evaluation, either the privacy risks are higher than expected, or their utility is insufficient.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Key takeaways from various challenges in synthetic data generation.

Problem	Key takeaway
Overcoming lack of consensus	Evaluations of synthetic data generators and synthetic data releases should cover different dimensions: broad utility/statistical fidelity, narrow utility (if synthetic data is released for a specific task), fairness, and privacy.
Conflicting metrics	Privacy and utility metrics that rely on similarity of synthetic records to real data, such as *nearest neighbour distance ratio*, should be used cautiously. As different studies use equivalent similarity-based metrics for measuring both utility and privacy, the usage of such metrics complicates the interpretation of the privacy-utility trade-off.
Principled privacy evaluation	If the purpose of synthetic data is to preserve privacy of the original data, practitioners and researchers should rigorously evaluate the associated privacy risks using modern techniques,[14,15] avoiding similarity-based metrics such as *distance to the closest record* or *nearest neighbour distance ratio*.
Ensuring provable privacy guarantees	Differential privacy (DP) is a well-established formal theory for provably guaranteeing a given level of data privacy, including for synthetic data. Although DP oftentimes can hurt utility and fairness, recent methods such as those based on k-way marginals[36] have significantly improved on its privacy-utility trade-off, making private methods a compelling candidate for synthetic data generation, especially when releasing synthetic data publicly. Even if DP is used, however, privacy risk evaluation should still be performed.

The integration of synthetic data in healthcare demands caution. Although it is promising, especially when using principled and provable utility and privacy-preserving methods,[36] its potential must not be overstated. Rigorous, unbiased evaluation is crucial before implementation. Our review highlights key gaps: a lack of consensus on performance metrics, including conflicting metrics, and an absence of standardized practices for ensuring privacy guarantees. Given these shortcomings, we caution against trusting synthetic data in high-risk scenarios where false positives, missed findings, or privacy breaches could cause harm. This includes both releases for specific purposes such as medical research or decision-making, as well as public data releases. Before adopting new methods introduced in the literature or implemented in software, even those with strong guarantees, institutions should emphasize robust technical and organizational safeguards to ensure comprehensive privacy protection.

Methods

For this scoping review, we adopted the protocol from *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*[46] (PRISMA). PRISMA stands as a recognized guideline, commonly adopted for laying out systematic reviews and meta-analyses. According to this guideline, we conduct the review by defining research questions, setting unambiguous inclusion and exclusion parameters, and detailing methods for searching, choosing, and charting data from chosen documents. We provide an overview of the procedure in Fig. 4.

Search strategy and selection criteria To identify relevant studies, we conducted a comprehensive search across two bibliographic databases and repositories spanning the period from January 2018 to July 2024. The databases and repositories included PubMed and Embase, which are focused on healthcare and medical research. By using these biomedical databases, we could identify 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. Full-text articles were obtained for those meeting the inclusion criteria described in Fig. 5. 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. In order to capture actionable insights on the trustworthiness of synthetic data in medicine, we designed the queries to find publications that evaluate the utility or privacy aspects of synthetic data. The queries used for each database are listed in Table 2 and were last run on July 1st, 2024.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

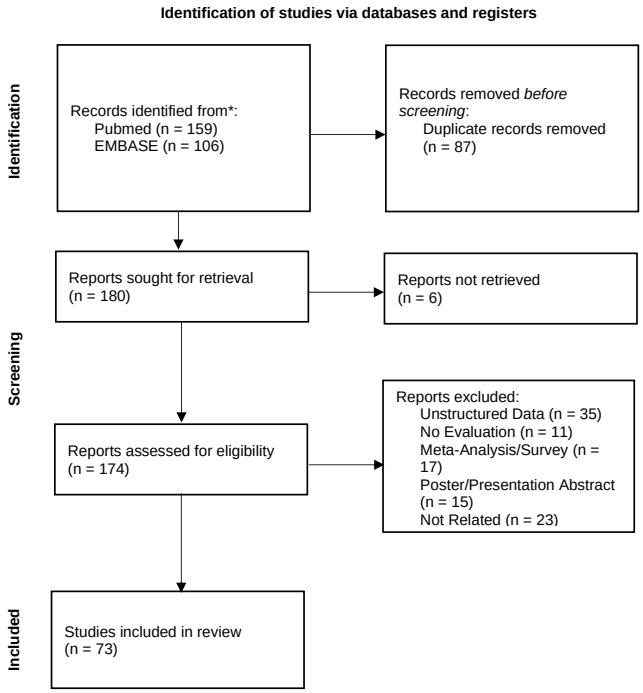

Figure 4. PRISMA flow diagram for the scoping review process.
Identification, screening, and inclusion process of studies for the scoping review. Following the PRISMA-SCR guidelines, 174 reports were assessed for eligibility and 73 of them were included in the final review.

Table 2. Queries by database.

Database	Query
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 [2018-2024]/py

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Inclusion criteria

- Publications describing research that uses synthetic data generation methods and evaluates their outputs

Exclusion criteria

- Surveys and systematic/scoping reviews
- Documents in languages other than English
- Publications with no assessment of the generated output, i.e., no evaluations of the utility/privacy aspects
- Publications that use unstructured data, i.e., images/text
- Poster abstracts

Figure 5. Eligibility criteria.
A list of inclusion and exclusion criteria used to select the studies from our initial database sample.

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 Supplementary Table 1, was collaboratively designed by the research team to delineate the specific variables to be extracted from the selected publications.

To standardize the data-charting process and ensure a unified treatment, we developed a taxonomy of evaluation methods suitable for the corpus of collected eligible publications, described next.

Taxonomy: Performance-Related Measures The proposed taxonomy classifies performance-related evaluation methods into three key dimensions: *broad utility*, *narrow utility*, and *fairness*. *Broad utility* (also referred to as statistical fidelity, or simply fidelity in the literature[47]) encompasses methods that we classify as *univariate similarity, bivariate similarity, multivariate similarity,* or *longitudinal similarity*. These methods are designed to capture specific aspects of data utility, ranging from straightforward one-dimensional comparisons to more complex analyses involving multiple variables and temporal patterns. This dimension is particularly valuable for making direct comparisons between different generative methods, ensuring that synthetic data can be effectively generalized across various applications and datasets. In contrast, *narrow utility* focuses on the performance of synthetic data in specific tasks or contexts. It evaluates how well the data serves particular purposes, such as improving model accuracy for a specific prediction task or supporting a specific type of statistical analysis. The *fairness* dimension examines how well synthetic data provides equitable treatment across different groups. This evaluation dimension is important as the standard measures of utility may not capture group level performance[48] which can in turn perpetuate harmful societal biases through the use of synthetic data. We include the extended taxonomy with detailed descriptions of each family of methods and specific examples in Supplementary Table 2.

Taxonomy: Privacy We divide the taxonomy for privacy evaluation methods into two main categories: *membership inference* and *attribute inference*. In the *membership inference* category, we include methods which study how effectively synthetic data can prevent the identification of whether specific individuals were part of the original dataset. Based on the literature we have reviewed, this category can be subdivided into three commonly used methods: record matching, distinguishing between synthetic records and real records from a holdout set, and various techniques for computing similarity between synthetic and real data records. We classify record matching as membership inference, which is consistent with prior approaches.[49] The *attribute inference* category addresses the risk of deducing sensitive information about individuals from synthetic data. This includes techniques like attribute inference based on record matching, which relies on conditioning on partial matches to predict specific attributes by comparing synthetic data with real data records. Another technique, *inference based on classification/regression models*, assesses how accurately private attributes can be inferred using predictive modeling approaches. As before, we provide a detailed description of the taxonomy items in Supplementary Table 2.

Data Availability

The comprehensive raw dataset is included as a supplementary file.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3. Synthetic data evaluation taxonomy.

Evaluation Dimensions	Family of Methods	General Evaluation Method
Broad Utility (Fidelity)	Univariate Similarity	Element-wise error 1-way marginals distributional similarity
	Bivariate Similarity	Correlation-based similarity Association-based similarity 2-way marginals distributional similarity
	Multivariate Similarity	Dimensionality reduction comparison Clustering similarity Distinguishability Multivariate distributional similarity
	Longitudinal Similarity	Correlation-based similarity Structural comparison
Narrow Utility	Replication of Predictive Models Performance	ML performance comparison ML explainability comparison
	Replication of Descriptive Statistics	Confidence interval overlap
	Expert Assessment	Qualitative expert assessment
Fairness	Statistical Parity of Generated Data	Difference in descriptive statistics between subgroups
	Disparate Impact	Difference in performance for a task between subgroups
Privacy	Membership Inference	Record matching Hold-out set distinguishing Distance to real data
	Attribute Inference	Inference based on record matching Inference based on classification/regression models

Code Availability

The code utilized for data analysis is available upon request.

Acknowledgements

Not Applicable.

Author Contributions

Ba.K., J.D. and J.L.R. conceived the scoping review design and objectives. Ba. 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, Bo.K. and F.P. assisted in interpreting the findings and shaping the discussion. All authors collaborated in structuring the manuscript's narrative, Ba.K. and Bo.K. wrote the manuscript and all authors read, edited, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

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.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 (2014).
9. Zhang, H. *et al.* Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The Twelfth International Conference on Learning Representations* (2024).
10. Brown, T. *et al.* Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (2020).
11. 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).
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, DOI: 10.1101/2022.11.22.517524 (2022).
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. Houssiau, F. *et al.* TAPAS: a toolbox for adversarial privacy auditing of synthetic data. *CoRR* **abs/2211.06550**, DOI: 10.48550/ARXIV.2211.06550 (2022). 2211.06550.
16. 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).
17. 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).
18. Arthur, L. *et al.* On the challenges of deploying privacy-preserving synthetic data in the enterprise. *CoRR* **abs/2307.04208**, DOI: 10.48550/ARXIV.2307.04208 (2023). 2307.04208.
19. Wagner, I. & Eckhoff, D. Technical privacy metrics: A systematic survey. *ACM Comput. Surv.* **51**, 1–38, DOI: 10.1145/3168389 (2019). 1512.00327.
20. Figueira, A. & Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **10**, 2733, DOI: 10.3390/math10152733 (2022).
21. 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).
22. Vallevik, V. B. *et al.* Can i trust my fake data–a comprehensive quality assessment framework for synthetic tabular data in healthcare. *Int. J. Med. Informatics* 105413 (2024).
23. The Synthetic Data Vault. Put synthetic data to work! https://sdv.dev/.
24. Brenninkmeijer, B. Table Evaluator. https://github.com/Baukebrenninkmeijer/table-evaluator (2023).
25. Qian, Z., Cebere, B.-C. & van der Schaar, M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, DOI: 10.48550/ARXIV.2301.07573 (2023).
26. 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).
27. Zhang, Z., Yan, C., Lasko, T. A., Sun, J. & Malin, B. A. SynTEG: a framework for temporal structured electronic health data simulation, DOI: 10.1093/JAMIA/OCAA262 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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. 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).

30. Hernadez, M., Epelde, G., Alberdi, A., Cilla, R. & Rankin, D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimension. *Methods Inf. Medicine* **62**, e19–e38, DOI: 10.1055/s-0042-1760247 (2023).

31. Ganev, G. & Cristofaro, E. D. On the inadequacy of similarity-based privacy metrics: Reconstruction attacks against "truly anonymous synthetic data". *CoRR* **abs/2312.05114**, DOI: 10.48550/ARXIV.2312.05114 (2023). 2312.05114.

32. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Foundations Trends Theor. Comput. Sci.* **9**, 211–407, DOI: 10.1561/0400000042 (2013).

33. Bambauer, J., Muralidhar, K. & Sarathy, R. Fool's gold: an illustrated critique of differential privacy. *Vand. J. Ent. & Tech. L.* **16**, 701 (2013).

34. Kulynych, B., Hsu, H., Troncoso, C. & Calmon, F. P. Arbitrary decisions are a hidden cost of differentially private training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1609–1623 (2023).

35. Bagdasaryan, E., Poursaeed, O. & Shmatikov, V. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, vol. 32 (2019).

36. McKenna, R., Miklau, G. & Sheldon, D. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *J. Priv. Confidentiality* **11** (2021).

37. McKenna, R., Mullins, B., Sheldon, D. & Miklau, G. AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.* **15**, 2599–2612 (2022).

38. Vietri, G. *et al.* Private synthetic data for multitask learning and marginal queries. In *Advances in Neural Information Processing Systems*, vol. 35, 18282–18295 (2022).

39. Pereira, M. *et al.* Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *PLOS ONE* **19**, e0297271, DOI: 10.1371/journal.pone.0297271 (2024).

40. Kulynych, B., Gómez, J. F., Kaissis, G., du Pin Calmon, F. & Troncoso, C. Attack-aware noise calibration for differential privacy. *CoRR* **abs/2407.02191**, DOI: 10.48550/ARXIV.2407.02191 (2024). 2407.02191.

41. Ziller, A. *et al.* Reconciling privacy and accuracy in AI for medical imaging. *Nat. Mach. Intell.* 1–11 (2024).

42. Ganev, G., Annamalai, M. S. M. S. & Cristofaro, E. D. The elusive pursuit of replicating PATE-GAN: benchmarking, auditing, debugging. *CoRR* **abs/2406.13985**, DOI: 10.48550/ARXIV.2406.13985 (2024). 2406.13985.

43. Nasr, M. *et al.* Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, 1631–1648 (2023).

44. Synthetic data activity by IEEE standards association. https://standards.ieee.org/industry-connections/synthetic-data/.

45. Maximising the potential of synthetic data generation in healthcare applications – European Commission call for proposals. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-ju-ihi-2023-05-04.

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

47. Jordon, J. *et al.* Synthetic data – what, why and how? *CoRR* **abs/2205.03257**, DOI: 10.48550/ARXIV.2205.03257 (2022). 2205.03257.

48. 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).

49. 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).

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

List of Figures

1 Scoping Review Results. **(a)** Visual overview of included works across various metrics. The figure depicts four dimensions: Database, Data Type, Purpose, and Publication Year. PPDS refers to Privacy Preserving Data Sharing. **(b)** Summary of the performance-related methods used in the included works. This includes a breakdown of categories such as Broad Utility, Narrow Utility and Fairness. **(c)** Summary of the performance-related methods used in the included papers. We categorized methods as Membership Inference or Attribute Inference. . . . 3
2 Synthetic data generation methods. We found 49 different synthetic data generation (SDG) methods, split into two main categories: GANs (Generative Adversarial Networks) and other techniques. GANs, shown in orange, make up 32.65% of the total, with 16 different methods. The remaining 67.35%, in blue, includes various approaches like Bayesian Networks, VAEs, and proprietary software. . . . 4
3 Number of works that evaluate privacy and use methods that provide privacy guarantees. Gap between the intended privacy focus of studies and the actual privacy evaluation. Only 46% (31/67) of the studies claiming to employ synthetic data for preserving privacy conducted any privacy evaluation. . . . 4
4 PRISMA flow diagram for the scoping review process. Identification, screening, and inclusion process of studies for the scoping review. Following the PRISMA-SCR guidelines, 174 reports were assessed for eligibility and 73 of them were included in the final review. . . . 7
5 Eligibility criteria. A list of inclusion and exclusion criteria used to select the studies from our initial database sample. 8

List of Tables

1 Key takeaways from various challenges in synthetic data generation. . . . 6
2 Queries by database. . . . 7
3 Synthetic data evaluation taxonomy. . . . 9

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary File

1 Database Search Strategy

Table 1. 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..2024]
Database	Database or retrieval tool	[PubMed, Embase]
General Utility Method	General Utility Metric category	Values in Table 2.
Utility Method	Specific Utility Metric used	Values in Table 2.
General Privacy Method	General Privacy Metric category	Values in Table 2.
Privacy Method	Specific Privacy Metric used	Values in Table 2.
Privacy Type	Type of privacy involved	[Membership Inference, Attribute inference]
Differential Privacy	Use of differential privacy	[Y,N]
SDG Method	Synthetic Data Generation Method used	Free text

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2 Synthetic Data Evaluation Extended Taxonomy

Fidelity: Refers to the accuracy with which synthetic data replicates the statistical properties and relationships of the original real data. Fidelity assessment includes various methods to determine how closely the synthetic data matches the real data across different statistical dimensions.

- **Univariate Similarity**
  - **Element-Wise Error**: Measures the difference between corresponding elements in synthetic and real datasets. Common metrics include Mean Squared Error (MSE), which calculates the average of the squared differences between corresponding data points.[1–3]
  - **Marginal Distributional Similarity**: Compares the distribution of individual variables between synthetic and real datasets to ensure that each variable's distribution in the synthetic data matches the real data. It features a multitude of different methods such as statistical tests (Mann Whitney U-test[4], T-Tests[5,6], Chi-Squared tests[5] ...),distance Between Probabilities (Wassestein Distance[7]), divergence computation (Kullback-Leibler Divergence, Hellinger Distance) and visual comparisons of marginal distributions.
- **Bivariate Similarity**
  - **Correlation-based Similarity**: Assesses the preservation of relationships between pairs of variables in synthetic data compared to real data. This method includes comparing correlation coefficients (e.g., Pearson or Spearman correlations).
  - **Association-based Similarity**: Evaluates the strength and direction of associations between variables, ensuring that the relationships in the synthetic data reflect those in the real data. This method includes Tau Statistic Comparison and Association Matrices comparison.
  - **2-Way Marginals Distributional Similarity**: Examines the joint distribution of two variables, assessing whether the synthetic data captures the correct dependencies between these variables as seen in the real data. This was usually present in the form of visual comparison of joint probabilities.
- **Multivariate Similarity**
  - **Dimensionality Reduction Comparison**: Uses techniques like Principal Component Analysis (PCA) to compare the principal components of synthetic and real data, revealing similarities or differences in underlying data structures.
  - **Clustering Similarity**: Assesses whether clusters identified in the real data are preserved in the synthetic data, using clustering algorithms to evaluate the consistency of groupings. For example, Emam et al.[7] merge synthetic data and real data, then perform a clustering algorithm. The metric is then extracted by comparing the number of synthetic samples in each cluster to the number of real samples.
  - **Distinguishability**: Measures how easily one can distinguish between synthetic and real data, often using classification tasks to determine the success rate of distinguishing the two datasets. Usually this comes in the form of ML models trained to do classification, mirroring what a Discriminator would do in a GAN architecture. Beigi et al.[8] train a classifier to distinguish between synthetic and real data then report its performance measure.
  - **Multivariate Distributional Similarity**: Examines the similarity of multivariate distributions between synthetic and real datasets, ensuring that complex interactions and dependencies among variables are accurately replicated. This is usually done with n-way marginals comparisons.
- **Longitudinal Similarity**
  - **Correlation-Based Similarity**: Evaluates the consistency of temporal correlations in longitudinal data, ensuring that trends and patterns over time are preserved in the synthetic data. This method includes comparing correlation coefficients between two time series (e.g., Pearson or Spearman correlations) or a comparison of autocorrelations.
  - **Structural Comparison**: Compares structural characteristics of data, such as trends, cycles, or other temporal features, ensuring that synthetic data accurately reflects the time-dependent structures in the real data. This method includes a comparison of directional symmetries[9] or simply visual comparisons of time series.

Utility: Measures the practical usefulness of synthetic data, particularly in replicating real-world scenarios and supporting decision-making processes. Utility evaluation methods ensure that synthetic data can effectively substitute real data in analytical applications.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Evaluation Dimensions	Family of Methods	General Evaluation Method	Instances
Broad Utility (Fidelity)	Univariate Similarity	Element-wise error Marginal distributional similarity	MSE Marginal Distributions Visual Comparison; Kullback-Leibler Divergence; Descriptive Statistics Comparison; Wilcoxon Signed-Rank Test; T-Test; Kolmogorov-Smirnov Test; Mann Whitney U-test; Chi-Squared Test; Hellinger Distance; Wasserstein Distance; Jensen-Shannon Divergence; Distance Between Probabilities
	Bivariate Similarity	Correlation-based similarity Association-based similarity 2-way marginals distributional similarity	Correlation Coefficient; Visual Comparison of Correlation Matrices; Covariance Matrix Comparison Log Odds Ratio; Association Matrices Comparison; Tau Statistic
	Multivariate Similarity	Dimensionality reduction comparison Clustering similarity Distinguishability Multivariate distributional similarity	PCA Visual Comparison; t-SNE Visual Comparison; FAMD Visual Comparison; Principal Components Comparison Cluster Analysis Distinguishability Performance Maximum Mean Discrepancy; Random Projection Normality Tests; Precision; Recall; Density; Coverage; Mardia Normality Tests
	Longitudinal Similarity	Correlation-based similarity Structural comparison	Autocorrelation Comparison Visual Inspection of Distances; Directional Symmetry; Transition Matrices Comparison
Narrow Utility (Utility)	Replication of Predictive Models Performance	ML performance comparison ML explainability comparison	ML Classification Performance; ML Regression Performance; ML RL Agent Comparison; ML Forecasting Performance Comparison ML Feature Importance Comparison; ML Classification Rules Comparison; ML Feature Importance
	Replication of Descriptive Statistics	Confidence interval overlap	Domain Specific Metric; Replication of Studies; Comparison to Previous Study Statistics; Hazard Ratio CI Comparison; Survival Analysis Comparison
	Expert Assessment	Qualitative expert assessment	Domain Expert Assessment
Fairness	Statistical Parity of Generated Data	Difference in descriptive statistics between subgroups	Fairness Statistical Parity of Generated Data
	Disparate Impact	Difference in performance for a task between subgroups	Fairness Disparate Impact
Privacy	Membership Inference	Record matching Hold-out set distinguishing Distance to real data	Exact Record Match; Partial Record Match Holdout Set Distance; Hypothesis Test Based on Holdout Set Distance; ε -Identifiability to Holdout Set; Distance to Closest Record; Nearest Neighbor Distance Ratio; Minimum Hamming Distance between Real Data and Synthetic Data; Distance to Real Data; Hamming Distance and Euclidean Distance Threshold; Cosine Distance to Real Data with Threshold; F1 Score of Model Using Distances of Synthetic Targets To Real Data; Nearest Neighbor Distance Threshold; Quantiles over Euclidean Distances to Real Data; Distance to Closest Ratio
	Attribute Inference	Inference based on record matching Inference based on classification/regression models	Partial Matching Reconstruction; CRLProxy Classification/Regression Model Attribute Prediction; Difference Between Expected and Observed Values

Table 2. Evaluation Dimensions and Corresponding Methods with Instances

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

- **Replication of Predictive Models Performance**
  - **ML Performance Comparison**: Assesses the performance of machine learning models trained on synthetic data compared to those trained on real data. This includes evaluating metrics like accuracy, precision, recall, and F1 score. This mostly entails classification tasks, but there have been works that compared regression performance or event reinforcement learning agents behaviour[10].
  - **ML Explainability Comparison**: Examines whether the feature importance and interpretability of models trained on synthetic data align with those trained on real data, using methods like SHAP values or feature importance scores.[11]
- **Replication of Descriptive Statistics**
  - **Comparison With Previous Study Results**: This method usually entains comparing whether the confidence intervals of key statistics in synthetic data overlap with those in real data, indicating that the synthetic data can support similar statistical analyses.[12] It also includes works that perform survival analyses on synthetic data and compare them to previously demonstrated results on real data[13].
- **Expert Assessment**
  - **Qualitative Expert Assessment**: Involves domain experts reviewing the synthetic data for its relevance, quality, and utility for specific applications, providing subjective evaluations to complement quantitative assessments.

Fairness: Evaluates whether synthetic data introduces or mitigates biases present in the original data. Fairness assessment is crucial to ensure equitable treatment across different demographic groups and to avoid perpetuating existing inequalities or introducing new biases.

- **Statistical Parity of Generated Data**
  - **Difference in Descriptive Statistics between Subgroups**: Assesses whether synthetic data maintains Statistical Parity of Generated Data across different demographic groups, ensuring equal representation and avoiding bias.[9]
- **Disparate Impact**
  - **Difference in Performance for a Task Between Subgroups**: Evaluates the differences in performance metrics (usually difference in True Positives) for machine learning models or other analytical tasks between different demographic groups in synthetic data, highlighting potential biases.[9]

Privacy: Concerns the ability of synthetic data to protect sensitive information. Privacy evaluation focuses on assessing whether synthetic data can inadvertently reveal information about individuals in the original dataset. The primary risks involve inference attacks, such as *membership inference*, where an attacker determines if an individual's data was part of the original dataset, and *attribute inference*, where sensitive attributes of individuals are deduced. These risks are particularly significant because, unlike exact data matching, synthetic data often retains statistical patterns and relationships from the original dataset, making it challenging to completely anonymize the data and prevent inference without compromising data utility. Prior works have also used adjacent terms to assess synthetic data privacy, such as *membership disclosure*. Emam et al.[14] link the two notions of inference and disclosure by describing inference as a broader type of disclosure risk.

- **Membership Inference**
  - **Record Matching**: Determines if synthetic data records can be matched to real data records, posing a risk of re-identification. This is also called Hit Rate and can take the form of a simple comparison between two records. Another way to do matching is "Partial Matching"[15], this is more relevant for cases of partial synthesis where not all attributes have been synthesized.
  - **Holdout Set Distinguishing**: Tests whether synthetic data can be distinguished from a holdout set not used in its generation, assessing overfitting and potential privacy risks.[16] Usually it has taken the form of computing a ratio such as Distance to Closest Ratio (DCR), performing hypothesis tests based on these distances, just reporting the itself. Another metric used is called CRLProxy which introduced the notion of a "trained distance" where the authors apply "a measure to calculate the distance between the representation of a known record and the representation of the synthetic records"[17].

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

  - **Distance to Real Data**: Uses distance metrics to measure the similarity between synthetic and real data, helping to evaluate the risk of re-identification. It has mostly taken the from of Nearest Neighbor Distance Ratio (NNDR)[18], and relied on multiple threshold based metrics such as epsilon-Identifiability to Real Data[19], threshold over quantiles, Nearest Neighbor Distance threshold, cosine distance[20], hamming distance and euclidean distance threshold[21].

- **Attribute Inference**
  - **Inference Based on Record Matching**: Assesses the risk of inferring sensitive attributes from synthetic data by matching it with real data records. Zhou et al.[22] for example perform partial matching in order to infer new information about sensitive records.
  - **Inference Based on Classification/Regression Models**: Evaluates the ability of models trained on synthetic data to predict sensitive attributes, assessing the risk of privacy breaches. Hernadez et al.[4] for example simulate attacks by providing an attacker a subset of the features. The attacker then uses ML models trained on their prior knowledge to infer the rest of the attributes.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

3 Extended Frequencies

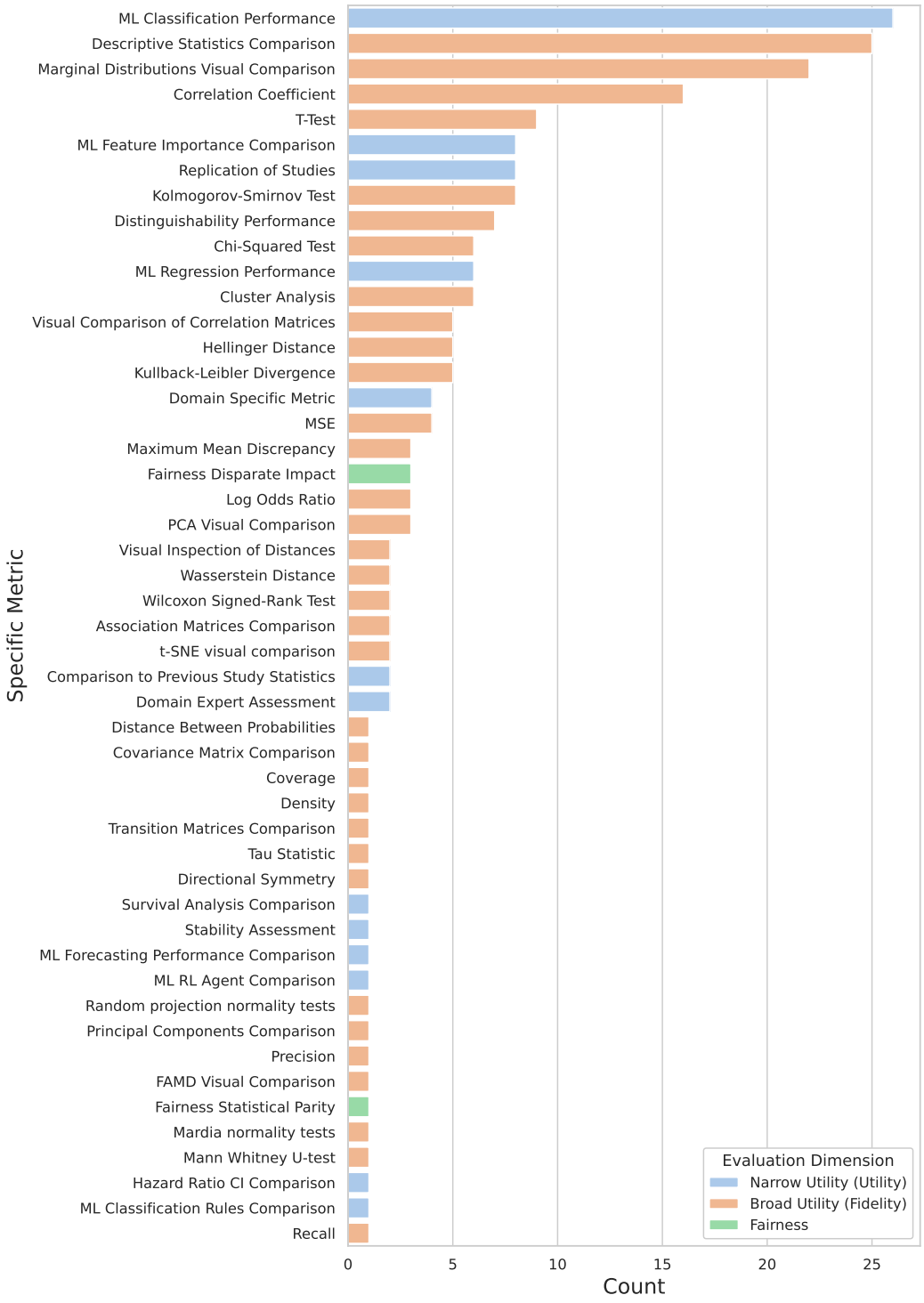

Figure 1. Utility metrics frequencies. Most works evaluated synthetic data by looking at ML Classification Performance, Descriptive Statistics Comparisons and Marginal Distributions Visual Comparison.

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

4 Additional Results

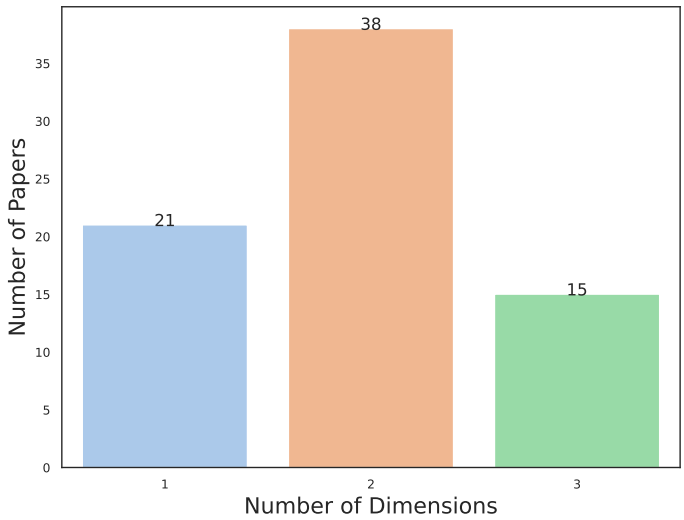

Figure 2. The number of utility evaluation dimensions. Utility evaluations are divided between Univariate Similarity, Bivariate Similarity, Multivariate Similarity and Domain Specific Similarity.

References

1. Karimian Sichani, E., Smith, A., El Emam, K. & Mosquera, L. Creating High-Quality Synthetic Health Data: Framework for Model Development and Validation. *JMIR Form. Res.* **8**, e53241, DOI: 10.2196/53241 (2024).
2. Varsou, D.-D. *et al.* In silico assessment of nanoparticle toxicity powered by the Enalos Cloud Platform: Integrating automated machine learning and synthetic data for enhanced nanosafety evaluation. *Comput. Struct. Biotechnol. J.* **25**, 47–60, DOI: 10.1016/j.csbj.2024.03.020 (2024).
3. 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).
4. Hernadez, M., Epelde, G., Alberdi, A., Cilla, R. & Rankin, D. Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions. *Methods Inf. Medicine* **62**, e19–e38, DOI: 10.1055/s-0042-1760247 (2023).
5. Greenberg, J. K. *et al.* Leveraging Artificial Intelligence and Synthetic Data Derivatives for Spine Surgery Research. *Glob. Spine J.* **13**, 2409–2421, DOI: 10.1177/21925682221085535 (2023).
6. Guillaudeux, M. *et al.* Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digit. Medicine* **6**, 37, DOI: 10.1038/s41746-023-00771-5 (2023).
7. 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).
8. Beigi, M., Shafquat, A., Mezey, J. & Aptekar, J. Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis. *AMIA Annu. Symp. Proc.* **2022**, 231–240 (2023).
9. 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).
10. Kuo, N. I.-H. *et al.* Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. *J. Biomed. Informatics* **144**, 104436, DOI: 10.1016/j.jbi.2023.104436 (2023).
11. Qian, Z. *et al.* Synthetic data for privacy-preserving clinical risk prediction, DOI: 10.1101/2023.05.18.23290114 (2023).

medRxiv preprint doi: https://doi.org/10.1101/2023.11.28.23299124; this version posted August 28, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

12. 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).

13. 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).

14. 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).

15. 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).

16. 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).

17. 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).

18. D'Amico, S. *et al.* Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. *JCO Clin. Cancer Informatics* e2300021, DOI: 10.1200/CCI.23.00021 (2023).

19. 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).

20. Isasa, I. *et al.* Comparative assessment of synthetic time series generation approaches in healthcare: Leveraging patient metadata for accurate data synthesis. *BMC Med. Informatics Decis. Mak.* **24**, 27, DOI: 10.1186/s12911-024-02427-0 (2024).

21. Sun, C., Van Soest, J. & Dumontier, M. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *J. Biomed. Informatics* **143**, 104404, DOI: 10.1016/j.jbi.2023.104404 (2023).

22. 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).