Information Processing and Management 63 (2026) 104715

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Synthetic data generation: A tertiary study

Navid Nobani [a,b,*], Giovanni Officioso [a], Filippo Pallucchini [a,b], Giancarlo Sperlì [c], Fabio Mercorio [a,b]

[a] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
[b] CRISP Research Centre Univ. of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, Milan, 20126, Italy
[c] Department of Electrical and Information Technology, University of Naples Federico II, Naples, Italy

ARTICLE INFO

Keywords:
Synthetic data generation
Tertiary study
Survey of surveys
Machine learning
Data privacy
Evaluation metrics

ABSTRACT

Synthetic Data Generation (SDG) is expanding rapidly, yet existing surveys differ widely in scope and methodological quality. This tertiary study systematically searched four major scholarly databases (2015-2025) and, after PRISMA screening and DARE-4 appraisal,[1] identified 17 eligible secondary studies. The evidence reveals a strong concentration in healthcare (58.8% of surveys), limited coverage of non-health domains, and inconsistent reporting of evaluation protocols (e.g., incomplete specification of metrics, data splits, baselines, or evaluation scripts). Fidelity and downstream utility dominate assessment practices, whereas privacy and diversity remain underexamined. Only 4 of 17 surveys provide any reproducibility artefacts. By consolidating these findings, we propose a compact, domain-agnostic evaluation baseline and highlight structural gaps in transparency, domain breadth, and methodological consistency. The study offers actionable guidance for strengthening reproducibility and broadening the evidential foundations of SDG research.

1. Introduction

Synthetic data refers to data generated artificially, rather than collected from real-world measurements (Andreini et al., 2026; Hellwig et al., 2025; Hernández-Ferrándiz et al., 2025; Rubin, 1993). The idea has roots in long-standing statistical practice, for example, simulation-based analysis and data perturbation for disclosure control (Reiter, 2005), and in computer graphics and vision, where simulators and renderers have been used to create controlled datasets (Chen et al., 2025a; Dosovitskiy et al., 2017; Singh et al., 2024). What was once a niche tool has become central to modern artificial intelligence. Synthetic data now supports model pretraining, augments scarce labels, stress-tests systems under rare or unsafe conditions (Nikolenko et al., 2021), and enables experimental designs that would be impractical or unethical with real subjects (van Breugel et al., 2024; Dhinakaran et al., 2025; Feng et al., 2025).

Early work on synthetic data for tabular and time-dependent settings largely relied on classical generative mechanisms. These include parametric models fitted to known families, copulas to capture cross-feature dependencies, Bayesian networks for structured variables (Xu et al., 2019), agent-based and micro-simulation frameworks for populations (Steinbacher et al., 2021), and domain-specific simulators for signals and events. Such approaches offer interpretability and clear assumptions, and they align well with the needs of statistical disclosure control in official statistics and healthcare registries (Nowok et al., 2016; Shi et al., 2026).

* Corresponding author.
E-mail addresses: navid.nobani@unimib.it (N. Nobani), g.officioso@campus.unimib.it (G. Officioso), filippo.pallucchini@unimib.it (F. Pallucchini), giancarlo.sperli@unina.it (G. Sperlì), fabio.mercorio@unimib.it (F. Mercorio).

https://doi.org/10.1016/j.ipm.2026.104715
Received 20 November 2025; Received in revised form 25 February 2026; Accepted 27 February 2026
Available online 16 March 2026

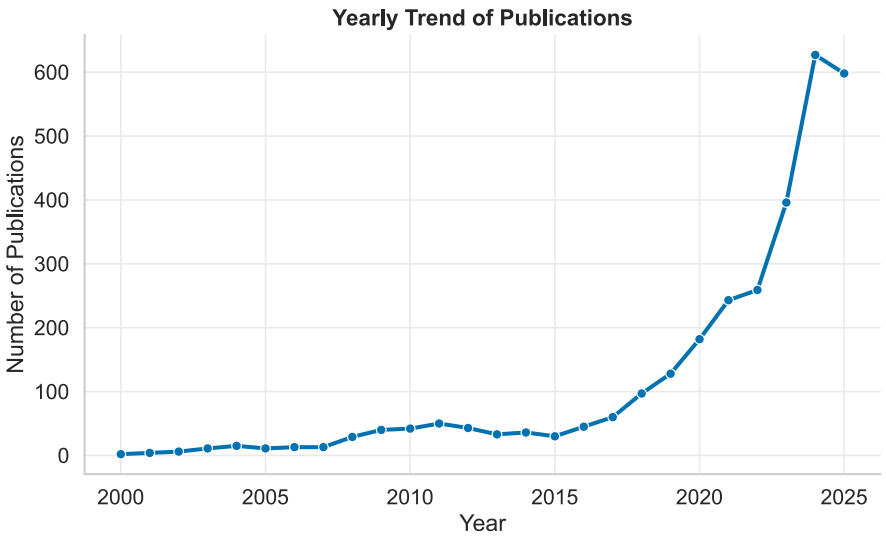

Fig. 1. Yearly trend of "synthetic data generation" as of 10 October 2025.

The landscape changed with the advent of powerful deep generative models. Variational autoencoders, generative adversarial networks (Chatterjee et al., 2025; Goodfellow et al., 2014; RezvaniNejad & Yameqani, 2025), auto-regressive transformers, normalizing flows, and, more recently, diffusion models (Ho et al., 2020), provide expressive mechanisms to learn complex data distributions across modalities (Chen et al., 2026; Wang et al., 2026b). These models have pushed synthetic data into the mainstream of the AI pipeline. In vision, audio, and language, they enable scale and diversity. For tabular and time-series data, they have matured from proof-of-concept tools into viable components for imputation, augmentation, and what-if analysis (Camino et al., 2018). The result is a steady shift from hand-crafted simulators to learned generators that can capture high-order dependencies without explicit structural specification (Gayathri et al., 2024; Huang et al., 2024; Nadăs et al., 2025).

1.1. Scope and focus

In this work, we explicitly target SDG methods. Our objective is to survey how synthetic data is produced across domains and modelling families. We do not cover synthetic data as a general topic beyond generation, and we do not review in depth adjacent areas such as stand-alone evaluation frameworks, governance and policy, data sharing infrastructures, or purely synthetic-data applications that lack a generation focus. Where we discuss evaluation, privacy, fairness, and robustness, we do so only insofar as they are necessary to understand and assess SDG methods.

Progress in SDG has occurred shoulder to shoulder with an expanding set of concerns. As models become more expressive, the risk of leakage from training records increases. Empirical attacks such as membership and attribute inference (Shokri et al., 2017), record linkage, and reconstruction have highlighted the possibility that synthetic datasets can reveal information about individuals (Stadler et al., 2022) if generation is not carefully controlled. Alongside privacy, additional issues have emerged. Utility is not guaranteed: a dataset that looks realistic may fail to support downstream tasks. Distributional coverage, mode collapse, and spurious correlations can mislead evaluation. Fairness and bias are also live concerns, since generators trained on imbalanced or historically biased data may reproduce or amplify those biases unless evaluation and safeguards are in place.

These opportunities and risks have made evaluation central to the field. Utility is often measured through task-based protocols such as train-on-synthetic, test-on-real (Alaa et al., 2022), complemented by fidelity, diversity, and calibration metrics. Privacy assessments range from empirical audits to formal guarantees under differential privacy (Abowd, 2018). There is growing interest in robustness and safety testing, for example, how synthetic data behaves under shifts (Taori et al., 2020) and how it affects decisions in high-stakes applications.

The literature that surveys SDG has expanded quickly (See Fig. 1), yet it is fragmented across domains and modalities. Many reviews focus on specific sectors such as healthcare (Baowaly et al., 2019) or on a single family of models, while others emphasise evaluation, privacy, or governance. This fragmentation makes it difficult for practitioners and researchers to obtain a consolidated view of methods, use-cases, risks, and reporting practices.

In this paper, we present a tertiary study, i.e., a survey of surveys on SDG. Our aim is to consolidate the secondary literature, chart what has been covered, and identify where gaps remain. We examine which application domains have been most frequently targeted, how publication activity and venue types have evolved, which institutions are contributing, and how transparency and traceability practices are reported. Methodologically, we adhere to a structured review process with a PRISMA (Moher et al., 2009) flow and a final quality gate based on DARE-4 criteria to ensure that only methodologically sound secondary studies enter our synthesis. We

make no claims beyond the evidence available in the included sources, and we highlight limitations where information is missing or ambiguous.

Contributions. Our contributions are threefold. First, we assemble and quality-appraise the body of SDG surveys to provide a consolidated map of the area. Second, we analyse domain coverage, publication and venue trends, and institutional participation to surface both concentration and under-explored regions. Third, we provide a structured taxonomy of SDG evaluation metrics by dividing the metrics into privacy, fidelity, diversity and utility categories. Rather than claiming novelty in isolation, our contribution lies in consolidating and systematising dispersed practices across surveys into a compact, domain-agnostic reference framework.

Structure of the article. The remainder of this article is structured as follows. Section 2 reviews the existing literature on tertiary studies and meta-surveys, outlining how this work complements prior research in machine learning and data privacy. Section 3 presents the review methodology, including the research objectives and questions, the multi-stage search strategy, the inclusion and exclusion criteria, and the quality assessment conducted using the DARE-4 framework. Section 4 reports the results derived from the selected studies, providing quantitative and qualitative insights into domain coverage, publication trends, and transparency practices. Section 5 discusses the main findings, highlights their implications, and integrates the discussion strands into a unified reflection on methodological heterogeneity and research gaps. Section 6 outlines the limitations of the study and suggests potential avenues for improvement in subsequent tertiary reviews. Finally, Section 7 concludes the study by summarising the key outcomes and presenting recommendations for future research on SDG.

2. Related work

In recent years, the consolidation of knowledge in rapidly evolving fields has seen the emergence not only of primary studies and secondary reviews (systematic literature reviews, mapping studies) but also of tertiary studies (Kurdi, 2025): reviews that systematically compile data and insights from various secondary sources, including systematic literature reviews, surveys and mapping studies on a particular subject (Kotti et al., 2023). This methodological approach has garnered significant attention across various scientific fields as it enables researchers to identify applicable methods and research questions while revealing patterns across large numbers of primary studies (Kotti et al., 2023; Kurdi, 2025). Through systematic collection, assessment, analysis, and categorisation of existing secondary research, tertiary studies facilitate a comprehensive synthesis of knowledge that would be difficult to achieve through individual secondary studies alone (Hernandez et al., 2022; Kaabachi et al., 2025; Lautrup et al., 2024).

Several prior works have proposed taxonomies that are closely related to ours, particularly in the context of evaluation. For instance, Kaabachi et al. (2025) introduce a multidimensional framework that organises synthetic data methods and metrics around utility, privacy, and fairness. Similarly, Stenger et al. (2024) propose an extensive taxonomy of evaluation metrics for synthetic time-series data, providing a detailed treatment of temporal fidelity, privacy risk, and downstream utility.

These contributions are complementary to the present study. While existing taxonomies primarily focus on evaluation metrics and evaluation frameworks for synthetic data, our work adopts a broader, tertiary perspective by organising how synthetic data generation is surveyed across domains, model families, and methodological orientations, with evaluation treated as one analytical dimension among others. This distinction allows us to position our taxonomy not as a replacement for prior frameworks, but as a unifying layer that connects evaluation practices to the wider ecosystem of SDG survey research.

2.1. Tertiary studies and meta-reviews in computer science

To contextualise our contribution within the contemporary academic landscape, we examine existing tertiary studies in computer science and data privacy. One notable example is the work by Egger et al. (2021) that proposes a meta-survey that considers review articles on deep learning across disciplines, quantifies bibliometric indicators (such as number of references and citations), analyses structural commonalities (for example, domain taxonomy or tasks), and extracts overarching challenges common to the surveys. It thereby treats survey articles as objects of analysis and surfaces cross-domain lessons that individual reviews cannot easily provide. Similarly, van Mourik et al. (2024) in their work analyse a corpus of 40 systematic reviews on Explainable AI (XAI), categorising them by characteristics such as model dependency, input-data types, output explanation types and mapping coverage of methods. They map each included review along a multi-dimensional grid and thereby reveal combinations of review focus that remain unexplored (i.e., meta-gaps). This approach highlights how tertiary studies can handle both what reviews cover (domain content) and how they cover it (methodological scope). In the software engineering domain, Kotti et al. (2023) conducted a comprehensive tertiary review on the use of machine learning in software engineering. They investigate secondary studies (SLRs-Systematic Literature Review, mapping studies, surveys) of machine-learning applications in software engineering, categorise them by research type (SLR vs mapping vs survey), knowledge area (e.g., defect prediction, effort estimation), and assess methodological rigour (for example, guideline adherence, quality assessment). By assessing not only domain coverage but also methodological quality of the survey works, they show how tertiary studies can critique the ecosystem of review work itself. In the same field, Kalibatiene and Jolanta (2025) enhance the understanding of the attributes that directly influence the SLR process in terms of time consumption. They performed a tertiary study that identified 138 secondary studies, mapped the possible influential attributes for SLR performance, extracted data from SLR reports and metadata, synthesised and analysed their influence, providing an overview of core trends related to those attributes over time. Beyond these domain-specific tertiary works, other umbrella-review style analyses have begun to appear: for example, a tertiary study of sentiment analysis reviews (Ligthart et al., 2021) treats survey works as data, applies an SLR protocol to the set of survey papers,

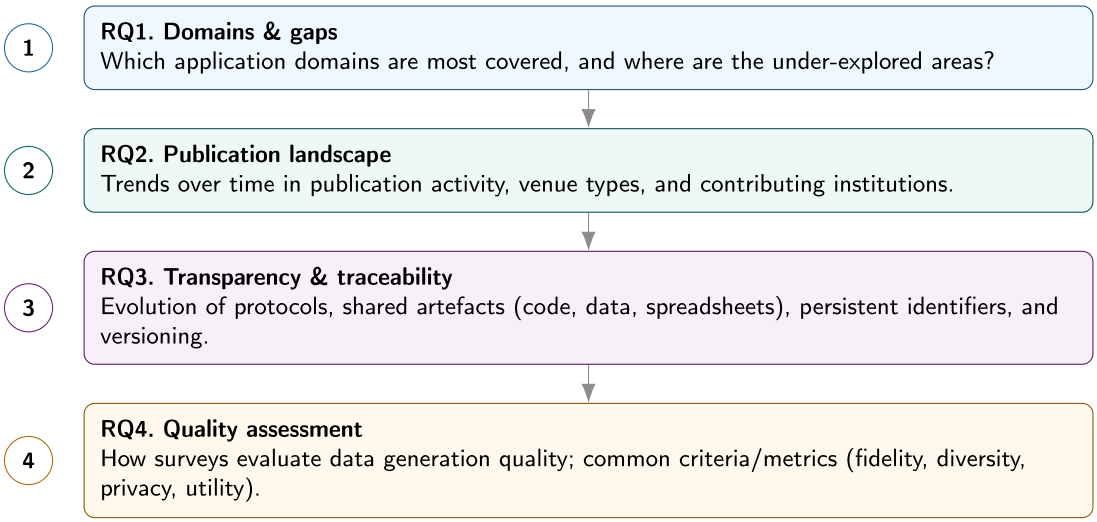

Fig. 2. Overview of the research questions guiding this tertiary study. Each question corresponds to one of the four research objectives defined in Section 3.1 and structures the analysis described in Section 3.2.

and identifies methodological commonalities and gaps in coverage. Finally, expanding the topic of computer science in the education field, some works (Bond et al., 2024; Bouguettaya et al., 2025) propose a comprehensive meta review to explore the scope and nature of Artificial Intelligence in education. Such works demonstrate that tertiary studies typically adopt the following pattern: (1) clear definition of the inclusion criteria for survey/review papers; (2) construction of a taxonomy or classification scheme for the included studies (often by domain, method, data type, year); (3) extraction of meta-metrics (number of papers, citations, references, distribution over years, venue types); (4) assessment of methodological aspects of the reviews (e.g., search strategy, inclusion/exclusion criteria, quality assessment of primary studies, synthesis method); (5) identification of coverage gaps (topics or combinations of methods/data unaddressed by reviews) and research agenda for future review work.

Tertiary studies in SDG. Beyond domain-specific reviews and methodological surveys that investigate particular facets of SDG, to our knowledge, the literature includes only one tertiary study explicitly centered on SDG. Rujas et al. (2025) conducted a scoping review of reviews following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-analyses extension for Scoping Reviews) framework. Their analysis covered 42 secondary studies published between 2014 and 2024, identified through PubMed, Scopus, and Web of Science. While their work provides a valuable synthesis of the healthcare-focused SDG literature, its scope remains confined to medical and clinical contexts involving human subjects.

In contrast, the present study offers a broader and domain-agnostic tertiary review that systematically maps the evolution, methodologies, evaluation frameworks, and reproducibility practices across the entire SDG research spectrum. By integrating perspectives from computer science, data privacy, and applied domains, this work extends beyond biomedical applications to capture the multidisciplinary nature and methodological diversity of SDG. The resulting synthesis establishes a comprehensive and comparable baseline for assessing how SDG surveys conceptualise, evaluate, and report synthetic data generation methods, thus facilitating greater transparency, interoperability, and reproducibility within the field.

3. Review methodology

3.1. Research objectives

The objective of this tertiary study is to provide a structured and quality-assessed synthesis of existing secondary literature on Synthetic Data Generation (SDG). Specifically, we aim to: (i) identify which domains and data modalities are most represented in current SDG surveys; (ii) characterise publication patterns, including venue types, temporal trends, and institutional contributions; (iii) assess the transparency and traceability practices adopted in prior surveys; and (iv) consolidate how SDG quality is evaluated across fidelity, utility, privacy, and diversity dimensions. These objectives collectively support a comprehensive and comparable overview of the SDG survey landscape and highlight gaps that should inform future research.

3.2. Research questions

To operationalise the above objectives, we define the following research questions, each mapping to a specific analytical dimension of the review. Together, they structure the data extraction, synthesis, and interpretation phases. Fig. 2 summarises the research questions.

Table 1
Search strings used for the automatic search across scholar databases.

Database	Search Strings
DBLP and Semantic Scholar	"synthetic data" generation survey; "synthetic data" generation review; "synthetic data generation" survey; "synthetic data generation" review; "synthetic data" survey; "synthetic data" review; synthetic data generation survey; synthetic data generation review.
Google Scholar and Crossref	"synthetic data" generation (survey OR review OR "systematic review" OR overview OR tutorial OR "state of the art").

By addressing these questions, we identify how SDG surveys have contributed to domain coverage, highlight under-represented areas, map the evolution of publishing patterns, and document the maturity of transparency and traceability. The current Section details how each question is operationalised from our extraction schema and quality filters.

3.3. Search strategy

The search strategy was conducted in three stages: (1) an automated search across four major scholarly databases, (2) automatic filtering of the retrieved records using predefined keywords, and (3) de-duplication based on publication titles. The search covered studies published between 2015 and 2025.

Automatic search. We queried four sources programmatically with custom Python scripts: Google Scholar, Crossref, Semantic Scholar, and DBLP. Scripts normalised core metadata (title, venue, year and DOI when available) and logged query strings and timestamps. We limited the publication window to 2015-2025 inclusive. The choice of 2015 as the starting year is motivated by the observed temporal trend in SDG research. As illustrated in Fig. 1, 2015 marks the earliest point from which the trend shows a consistent and sustained increase in publication activity. Because each source has its own query syntax, we used semantically aligned but source-specific strings. Table 1 presents all keywords used within different sources.

Automatic filtering. To ensure high precision in the initial retrieval phase, we implemented a conservative automatic pre-filtering step. Specifically, we retained only those records whose *titles* simultaneously contained the tokens `synthetic`, `data`, and `generation`, as well as at least one of the terms `survey` or `review`. This rule-based filtering approach aimed to prioritise relevance and reduce noise before any manual inspection. Such lexical filtering, though restrictive, is a well-established practice in large-scale literature mining for technical domains, where high recall is less critical than maintaining topical coherence at early stages ($n = 2{,}821$).

Deduplication. Following the automatic filtering, duplicate records were removed by matching identical titles across all queried sources. This ensured that multiple index entries pointing to the same publication were consolidated ($n = 95$).

3.4. Selection process

For all records deemed potentially eligible after preliminary filtering, we retrieved both the abstract and full text to assess compliance with the predefined inclusion criteria. Each record was evaluated systematically to ensure that only studies addressing synthetic data generation surveys or reviews were retained. The overall selection flow is summarised in the PRISMA diagram (Fig. 3), which provides a transparent account of the inclusion and exclusion steps. In summary: records screened ($n = 369$); records excluded during manual screening ($n = 257$); reports sought for retrieval ($n = 63$); reports not successfully retrieved ($n = 9$); and reports assessed for eligibility ($n = 54$). This multi-stage selection ensured a rigorous and reproducible filtering pipeline, combining automation for scalability with manual verification for semantic accuracy.

3.5. Quality appraisal with DARE-4

As a final inclusion gate, i.e., the last phase of the screening stage of the PRISMA framework, we applied the DARE-4 (Database of Abstracts of Reviews of Effects) checklist, presented in Table 2, to the remaining $n = 54$ papers. This structured quality appraisal tool evaluates secondary studies along four methodological dimensions, each scored as $\{1, 0.5, 0\}$, yielding total scores within the interval $[0, 4]$. Similar to Kotti et al. (2023), we adopted a decision threshold of 2, retaining 17 studies with DARE-4 scores ≥ 2 and excluding those below this threshold (see Table 3). The use of DARE-4 at the post–full-text stage ensured that only methodologically robust secondary studies contributed to the synthesis. We believe that this step improved the overall reliability of our findings while maintaining a manageable screening workload, thus aligning the review process with best practices in evidence-based computer science research.

Table 2
DARE-4 criteria applied for quality assessment of secondary studies.

QA Criterion	Assessment	Score	Description
Inclusion/Exclusion Criteria (IC/EC)	Yes	1	IC/EC explicitly stated and well-defined.
	Partial	0.5	IC/EC mentioned but only implicitly or partially described.
	No	0	No IC/EC criteria described.
Search Space	Yes	1	Search in four or more digital libraries with extra strategies.
	Partial	0.5	Search in three to four libraries, no extra strategies.
	No	0	One or two libraries searched, or very restricted scope.
Quality Assessment of Primary Studies	Yes	1	Quality criteria clearly defined and applied.
	Partial	0.5	Quality assessment discussed but not formalised.
	No	0	No quality assessment reported.
Information About Primary Studies	Yes	1	Complete and detailed information on primary studies.
	Partial	0.5	Only summarised information provided.
	No	0	Primary study details not reported.

Table 3
Overview of included secondary studies: year, type, DARE-4 quality score, number of primary studies, covered years, and availability of supplementary materials.

Study	Year	Type	QA score	Studies	Covered years	Supplementary Materials
Figueira and Vaz (2022)	2022	Review	2.5	99	2010–2022	No
Hernandez et al. (2022)	2022	Systematic Review	3.0	34	2016–2021	No
Perkonoja et al. (2023)	2023	Systematic Review	3.5	36	2016–2024	No
Goyal and Mahmoud (2024a)	2024	Systematic Review	3.5	77	2014–2024	No
Delussu et al. (2024)	2024	Review	2.5	100	2023–2024	No
Schieber et al. (2024)	2024	Systematic Review	2.0	34	2000–2023	No
Liu et al. (2024)	2024	Systematic Review	3.0	35	2019–2023	No
Shahul Hameed et al. (2024)	2024	Review	2.5	17	2021–2023	No
Lautrup et al. (2024)	2024	Systematic Review	2.5	47	2020–2023	GitHub
Rao et al. (2025)	2025	Scoping review	2.5	59	2020–2025	No
Chen et al. (2025b)	2025	Scoping review	2.5	48	2010–2024	GitHub
Ibrahim et al. (2025)	2025	Systematic Review	2.0	249	2021–2023	GitHub
Alismail and Lanquillon (2025)	2025	Systematic Review	3.0	36	2022–2024	GitHub
Loni et al. (2025)	2025	Scoping review	2.0	52	2018–2023	No
Hyrup et al. (2025)	2025	Systematic Review	2.0	32	2018–2023	No
Kaabachi et al. (2025)	2025	Scoping review	2.0	73	2018–2024	No
Rujas et al. (2025)	2025	Scoping review	2.0	42	2014–2024	No

Execution protocol and stopping rule. All searches were executed programmatically using identical query strings and parameters. Retrieval from Google Scholar was conducted via SerpApi[1] to minimise browser-level personalisation effects, and the execution environment and codebase were kept constant. Results were paginated in blocks of 20 with a safety cap of 100 pages. Pagination followed a saturation-based stopping criterion: retrieval ceased when no new records were returned or when additional pages yielded only duplicates, rather than after an arbitrary page limit.

3.6. Final dataset composition

The final dataset analysed in this study consists of the 17 secondary studies that passed the full screening and quality appraisal stages described above. These records constitute the complete evidence base for the tertiary synthesis.

For each included study, we extracted a structured set of metadata to enable systematic comparison and quantitative summaries. The extracted fields include authors and institutional affiliations; declared study type; years covered; venue types considered (conference, journal, preprint, other); number of primary papers analysed; application domains; data modalities; evaluation metrics; the availability of reproducibility or supplementary artefacts (e.g., code, data, spreadsheets, living tables); and whether limitations were explicitly stated.

Data extraction followed a fixed schema and was conducted by one author and verified by another. Any disagreements were resolved through discussion and inspection of the original source text, ensuring consistency and reliability of the extracted information.

The characteristics of all included studies are reported in Table 3, which serves as the dataset underlying all quantitative and qualitative analyses presented in Sections 4 and 5.

[1] https://serpapi.com/

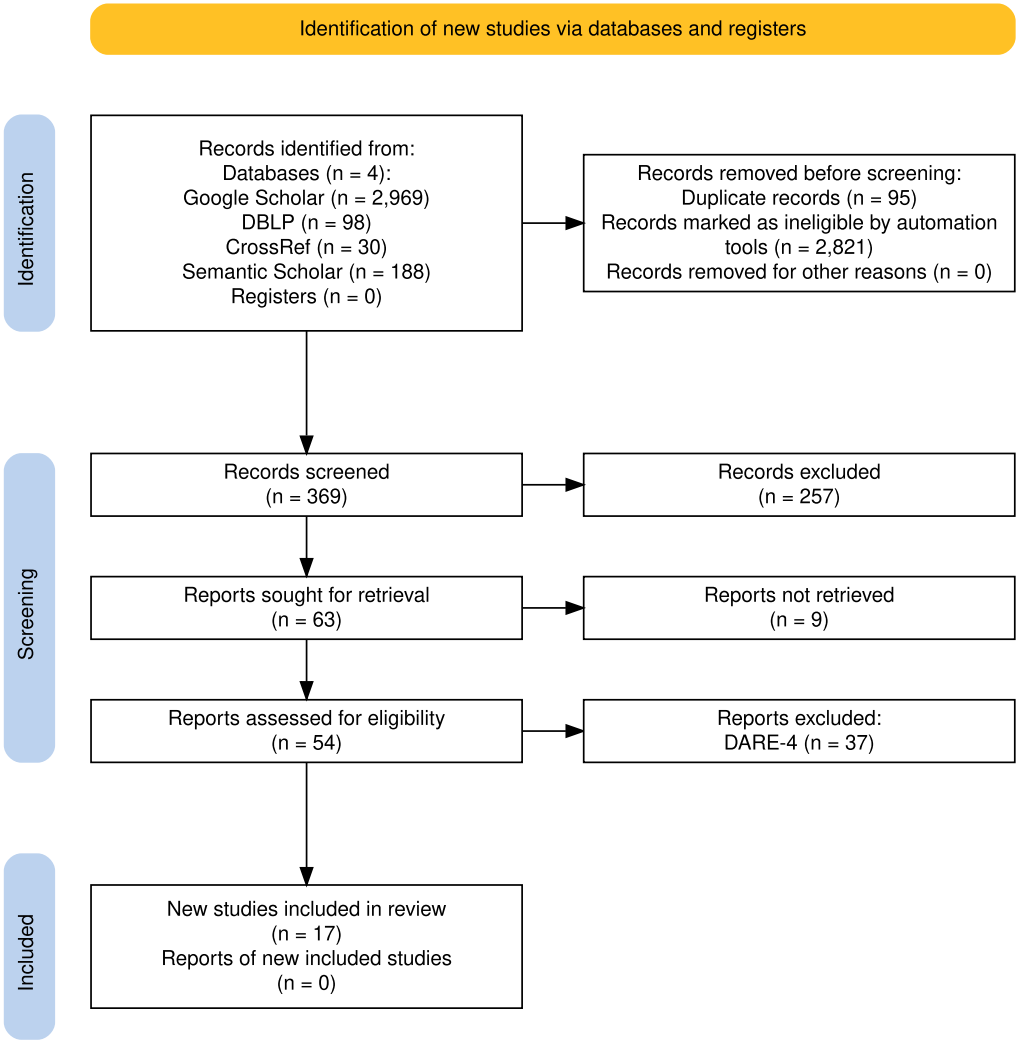

Fig. 3. PRISMA flow diagram for identification, screening, eligibility, quality appraisal, and inclusion in our survey of surveys on synthetic data generation. (Tool used to generate the figure: Haddaway et al., 2022).

3.7. Overlap of primary studies across included surveys

To assess the degree of convergence within the surveyed literature, we analysed the overlap of primary studies cited across the 17 included secondary studies. Specifically, for each survey, we extracted the set of foundational primary works explicitly discussed and then computed their frequency of occurrence across surveys. Table 4 reports the most recurrent primary studies, grouped by methodological family (e.g., GAN-based models, non-deep learning baselines, and other notable approaches), together with the number of surveys in which each study appears.

This analysis serves two purposes. First, it quantifies the extent to which the survey landscape is built upon a shared methodological core, as opposed to covering largely disjoint sets of primary contributions. Second, it identifies canonical reference models and benchmarks that function as de facto standards in synthetic data generation research. By highlighting highly recurrent studies (e.g., CTGAN, medGAN, SMOTE, PATE-GAN), we make explicit which primary methods structure the comparative narrative across surveys and, conversely, where fragmentation or limited cross-survey alignment persists.

4. Results

4.1. RQ1: targeted domains and modalities in SDG surveys

Methodological note. To distinguish between descriptive breadth of the literature and conclusions grounded in methodological quality, we adopt a dual perspective for RQ1. Domain and modality coverage is analysed on the **full screened set of 54 surveys**, while all subsequent research questions rely exclusively on the **DARE-4-filtered subset of 17 methodologically robust studies**.

Table 4
Highly Recurrent Primary Studies Across the Survey Corpus. This table illustrates the degree of overlap among the 17 included secondary studies by highlighting the most frequently analysed foundational methods and benchmarks.

Primary Study	Reference	Overlap	Surveys Analysing the Study
A. Generative Adversarial Networks (GANs)			
CTGAN	Xu et al. (2019)	9	Figueira and Vaz (2022), Goyal and Mahmoud (2024b), Hyrup et al. (2025), Ibrahim et al. (2025), Kaabachi et al. (2025), Lautrup et al. (2024), Liu et al. (2024), Perkonoja et al. (2023), Shahul Hameed et al. (2024)
medGAN	Choi et al. (2017)	8	Chen et al. (2025b), Figueira and Vaz (2022), Goyal and Mahmoud (2024b), Hernandez et al. (2022), Hyrup et al. (2025), Ibrahim et al. (2025), Kaabachi et al. (2025), Liu et al. (2024)
HealthGAN	Yale et al. (2020)	7	Chen et al. (2025b), Hernandez et al. (2022), Hyrup et al. (2025), Ibrahim et al. (2025), Kaabachi et al. (2025), Lautrup et al. (2024), Perkonoja et al. (2023)
B. Non-Deep Learning Methods			
SMOTE	Chawla et al. (2002)	5	Figueira and Vaz (2022), Goyal and Mahmoud (2024b), Ibrahim et al. (2025), Kaabachi et al. (2025), Shahul Hameed et al. (2024)
Synthpop	Nowok et al. (2016)	4	Hyrup et al. (2025), Kaabachi et al. (2025), Lautrup et al. (2024), Perkonoja et al. (2023)
Synthea	Walonoski et al. (2018)	3	Chen et al. (2025b), Hernandez et al. (2022), Perkonoja et al. (2023)
Datasynthesizer	Ping et al. (2017)	2	Goyal and Mahmoud (2024b), Lautrup et al. (2024)
C. Other Notable Recurrent Studies			
Privbayes	Zhang et al. (2017)	5	Figueira and Vaz (2022), Hernandez et al. (2022), Hyrup et al. (2025), Kaabachi et al. (2025), Lautrup et al. (2024)
PATE-GAN	Jordon et al. (2018)	5	Hernandez et al. (2022), Hyrup et al. (2025), Ibrahim et al. (2025), Kaabachi et al. (2025), Lautrup et al. (2024)
Promptehr	Wang and Sun (2022)	4	Chen et al. (2025b), Ibrahim et al. (2025), Loni et al. (2025), Perkonoja et al. (2023)

Notes. [1]Shahul Hameed et al. (2024) cite a method that uses CTGAN. [2]Ibrahim et al. (2025) cite a method that uses HealthGAN. [3]Goyal and Mahmoud (2024b), Ibrahim et al. (2025) cite the SMOTE without detailed secondary analysis.

Domain and modality coverage (full set, $n = 54$). Across the complete screened corpus, surveys remain heavily concentrated in **healthcare and biomedical contexts**, with a strong emphasis on electronic health records, longitudinal patient trajectories, and medical multi-modal data. Outside this dominant cluster, a more heterogeneous but still sparse landscape emerges. A limited number of surveys address **time-series synthesis**, **industrial and manufacturing data**, **cybersecurity**, **IoT systems**, and **social-science survey data**. These areas, however, appear only sporadically and lack the continuity observed in health-centred work.

In terms of modalities, **tabular data** and **computer vision** remain the most frequently targeted. Time-dependent data and non-medical multi-modal settings are present but marginal, typically confined to isolated surveys rather than sustained thematic lines. Notably, several reviews focusing on **time-series generation** belong to this broader screened set but fall outside the final quality-filtered corpus due to limited methodological reporting.

Comparison with the DARE-4-filtered subset ($n = 17$). Restricting the analysis to the quality-assessed studies sharpens rather than overturns these patterns. Healthcare and biomedicine become even more dominant, while non-health verticals and structured modalities beyond tabular and vision largely disappear. This indicates that the apparent lack of coverage in areas such as time-series, industrial systems, and cyber-physical domains is not solely an artefact of the DARE-4 filter, but reflects a genuine imbalance in the SDG survey landscape: these topics are present, yet tend to be addressed in surveys that report less systematic methodologies.

Implication. Overall, the dual analysis suggests that the field exhibits both a **descriptive breadth gap** and a **quality gap**. Some domains, such as time-series synthesis, do exist in the survey literature but are under-represented among methodologically rigorous reviews. Others, including finance, public administration, education, mobility, energy systems, and large-scale industrial analytics, remain thinly covered even at the descriptive level. This reinforces the need for future SDG surveys that both broaden domain scope and adhere to stronger standards of transparency and methodological rigour.

For the extended RQ1 analysis ($n = 54$), domain and modality labels were extracted using title, abstract, and high-level scope descriptions in the full text, following a lightweight descriptive coding procedure that was independently performed by two authors and reconciled through discussion.

Take-away. In answer to RQ1, SDG surveys most frequently target healthcare and biomedicine, with strong emphasis on tabular EHRs, longitudinal patient data, and medical imaging. Expanding the lens to the full screened set reveals additional, though still sparse, coverage of time-series and industrial applications. However, the core conclusion remains unchanged: domain diversification beyond healthcare and richer modality coverage in non-medical settings represent the clearest opportunities for future survey work. Table 5 reports the raw labels as recorded and their counts.

4.2. RQ2: most frequently targeted application domains, and under-explored gaps

Study types and publication outlet. Study type is reported for 17 records: *systematic review* (9), *review* (3), and *scoping review* (5). Publication outlet type is dominated by journals (14), with smaller contributions from arXiv (2) and conferences (1).

Scope of primary-literature windows. Almost all records clearly specify a year range for the primary studies they analyse; collectively, these span from **2000** to **2025**.

Table 5
Summary of domains, modalities, and cross-cutting foci in the SDG surveys. The upper block reports counts for the DARE-4-filtered subset ($n = 17$). The lower block highlights additional domains emerging in the full screened set ($n = 54$) that are not represented among the quality-filtered studies.

Category	Count (% of 17)
Healthcare & biomedicine (incl. EHRs, longitudinal, multi-modal)	11 (58.8%)
Tabular modality (incl. tabular EHRs)	3 (17.6%)
Computer vision modality	1 (5.9%)
Generic domain (domain-agnostic)	1 (5.9%)
Multi-modal medical data	1 (5.9%)
Bias mitigation (cross-cutting)	1 (5.9%)
Agentic generators (approach-focused)	1 (5.9%)
GAN-focused (model family)	1 (5.9%)
Object recognition (CV task)	1 (5.9%)
Additional domains observed in the full screened set ($n = 54$), present only outside the DARE-4-filtered set	
Time-series synthesis (temporal and sequential data)	
Manufacturing and industrial systems	
Cybersecurity and intrusion detection	
IoT and cyber-physical systems	
Social-science and survey data	

Notes. The health bucket aggregates *Healthcare*, *Biomedical research*, *Health records*, *Tabular health records*, *Longitudinal patient data*, and *Medical multi-modal data*. The tabular modality count includes *Tabular data* and *Tabular health records*.

Size of the surveyed primary corpora. Regarding the reported number of primary papers, the median is **47**, with a range of **17** to **249** and a mean of **62.9**.

Institutions. Across final records, we identified **30** distinct institutions. The most frequently represented are the *University of Southern Denmark* and the *University of Turku*, each appearing twice. The remaining affiliations are highly diverse, spanning universities and research centres across Europe, North America, and Asia, such as the *Technical University of Munich*, *University of Michigan*, *Maastricht University*, and *Universidad Politécnica de Madrid*.

Methodological approaches in existing surveys. Across domains, survey methodologies exhibit substantial variation in rigour and transparency. Some works adhere to systematic review protocols following established PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines for comprehensive literature identification and selection (Akpinar et al., 2024; Kaabachi et al., 2025; Perkonoja et al., 2023), while others rely on narrative or scoping methodologies such as PRISMA-ScR (PRISMA extension for Scoping Reviews) that provide broader overviews without systematic selection criteria (Chen et al., 2025b; Little et al., 2023). Quality assessment practices are likewise inconsistent: certain surveys adopt structured evaluation criteria based on frameworks from the Cochrane Handbook for Systematic Reviews of Interventions and GRADE (Grading of Recommendations, Assessment, Development and Evaluations) framework (Perkonoja et al., 2023), assessing risks of selection bias, reporting bias, and performance bias (Perkonoja et al., 2023), whereas many provide no explicit quality appraisal or rely on informal judgment of study relevance and validity (Lautrup et al., 2024). This methodological heterogeneity complicates cross-domain comparison and limits the synthesis of higher-level insights (Hernandez et al., 2022; Lautrup et al., 2024).

Transparency signals. Across the reviewed studies, transparency indicators such as the availability of reproducible artefacts and the inclusion of limitation statements were examined. Reported artefacts primarily include `GitHub` repositories, while many studies lack any supplementary materials. Similarly, only a few studies provide explicit limitation statements or specify concrete constraints, whereas most omit such information altogether.

4.3. RQ3: transparency and traceability practices

We assess transparency and traceability strictly from the fields captured during data extraction, without imputing missing information or aggregating unlabelled items.

Released artefacts (code, data, spreadsheets). Among the records where an explicit entry on artefacts was captured ($n = 17$), **4** include a link to a public repository (e.g., GitHub) and **13** report *no* artefacts. The artefact type (code vs. data vs. spreadsheet) is not consistently specified beyond the presence/absence of a public link, so we refrain from finer-grained categorisation.

Protocol availability. No record contains an explicit protocol link or registration identifier in the captured transparency fields. Consequently, we cannot quantify protocol availability or its change over time.

Persistent identifiers and versioning. Persistent identifiers (beyond standard bibliographic metadata) and versioning statements for released artefacts were not captured in a form that supports quantitative analysis; we therefore do not report figures for these aspects.

On temporal "evolution". Because transparency fields were not consistently time-annotated at the level required for longitudinal analysis, we do not claim trends. The evidence we *can* report is the cross-sectional snapshot above: limited public release of artefacts and no documented protocol links in the captured fields.

4.4. RQ4: quality assessment and evaluation metrics

The assignment of individual metrics to quality dimensions follows a hybrid procedure. Where surveyed reviews explicitly categorise metrics, we retain their conventions; where such categorisations are absent or inconsistent, we apply a harmonisation step based on the primary evaluation intent of each metric. The full mapping, including the rationale for each assignment and discussion of borderline cases (e.g., coverage ratio vs. recall), is provided in Appendix A.

This research question examines how SDG surveys frame and operationalise evaluation. Rather than providing a technical review of individual metrics, our focus is on understanding *how evaluation is used in practice* across the secondary literature: which quality dimensions are prioritised, how they are instantiated through metrics, and what this implies for comparability and cumulative knowledge building.

How evaluation is framed. Across the analysed surveys, evaluation is most commonly organised around a small set of high-level quality dimensions, typically including *utility*, *fidelity*, and *privacy*. A smaller subset of reviews also incorporates *diversity*, *fairness*, or *robustness*, although these dimensions are far less consistently operationalised. This shared conceptual vocabulary suggests an emerging consensus at the level of principles; however, the translation of these principles into concrete evaluation practices remains highly heterogeneous.

Operationalisation through metrics. Despite agreement on broad quality dimensions, surveys differ markedly in how these dimensions are instantiated. The same category (for example, utility) may be evaluated through predictive performance in one survey, statistical similarity in another, and downstream task effectiveness in a third. As a result, the meaning of core concepts such as *high utility* or *good fidelity* varies substantially across reviews. This diversity of operationalisations reflects the interdisciplinary nature of SDG, but it also complicates cross-study synthesis and weakens the interpretability of comparative claims at the tertiary level.

Imbalances among evaluation dimensions. A further pattern concerns the uneven attention devoted to different aspects of quality. Utility- and fidelity-oriented metrics dominate the landscape, while privacy is treated more selectively and diversity or fairness remain peripheral in most surveys. This imbalance suggests that, in practice, evaluation in SDG surveys continues to be driven primarily by performance-oriented considerations, even when ethical, legal, and societal implications are acknowledged at a conceptual level. The consequence is that some of the most critical motivations for using synthetic data, such as risk mitigation and bias reduction, are not yet matched by equally mature evaluation practices.

Comparability and reproducibility. From a tertiary perspective, one of the most consequential findings is the limited comparability of evaluation practices across SDG surveys. Even when reviews consider overlapping sets of primary studies, differences in metric choice, experimental design, and reporting standards often prevent meaningful aggregation of results. Moreover, although surveys typically list the metrics employed in the primary literature, fewer provide sufficient detail on evaluation protocols, parameter settings, or dataset splits to support reproducibility. This lack of standardisation constrains the extent to which evidence can be accumulated across studies and undermines the development of shared benchmarks for SDG methods.

Representative metrics and dominant patterns. Table 6 summarises a representative subset of the most frequently reported evaluation metrics, grouped by quality dimension. The purpose of this table is not to provide an exhaustive catalogue, but to illustrate dominant patterns in metric usage across surveys. In particular, it highlights the concentration of effort around utility and fidelity measures, the comparatively narrower set of privacy metrics, and the limited operationalisation of diversity- and fairness-related criteria. For completeness, detailed lists and descriptions of individual metrics are provided in Appendix A.

Implications for future SDG surveys. Taken together, these findings indicate that evaluation in SDG surveys currently plays a dual role: it serves as a necessary tool for assessing methodological progress, but it also functions as a source of fragmentation due to the absence of shared conventions. For future secondary and tertiary studies, this suggests a need to move beyond metric enumeration toward clearer rationales for metric selection, stronger alignment between quality dimensions and operational measures, and more transparent reporting of evaluation protocols. Such advances would strengthen the interpretability of survey findings and enhance the cumulative value of the SDG literature.

4.5. How SDG surveys frame and report generation methods

Across the analysed corpus of seventeen SDG surveys, generation methods are framed using three recurring organisational logics: (i) *architecture-driven* taxonomies that group approaches by model family, (ii) *modality-driven* taxonomies that start from the data type and then align suitable generators, and (iii) *workflow- and mechanism-driven* framings that emphasise how synthesis is operationalised in practice. Most surveys combine at least two of these perspectives, although the depth and technical specificity vary considerably.

Table 6
Top 20 evaluation metrics most frequently mentioned across the analysed surveys.

Metric	Mentions	Criterion
Accuracy	13	Utility
F1-score	13	Utility
AUROC	13	Utility
Recall (Sensitivity)	10	Utility
Wasserstein distance (EMD)	10	Fidelity
Jensen-Shannon divergence	9	Fidelity
KL divergence	9	Fidelity
Maximum Mean Discrepancy (MMD)	9	Fidelity
Precision	8	Utility
Inception Score	8	Diversity
FID (Fréchet Inception Distance)	7	Diversity
RMSE	6	Fidelity
Structural Similarity Index (SSIM)	6	Fidelity
ROUGE	6	Fidelity
MSE	5	Fidelity
BLEU	5	Fidelity
Perplexity	5	Fidelity
Kolmogorov-Smirnov test	4	Fidelity
Chi-square test	3	Fidelity
Cosine similarity	3	Fidelity

Architecture-driven and family-based organisation. A large share of the surveys adopt an explicit family-based structure in which generative approaches are grouped into classical/statistical models, GAN variants, VAEs, transformer-based generators, diffusion models, and, more recently, LLM-centred pipelines (Chen et al., 2025b; Delussu et al., 2024; Figueira & Vaz, 2022; Hernandez et al., 2022; Hyrup et al., 2025; Ibrahim et al., 2025; Kaabachi et al., 2025; Lautrup et al., 2024; Liu et al., 2024; Loni et al., 2025; Perkonoja et al., 2023). In tabular and EHR-oriented reviews, classical generators such as Bayesian networks, CART-based synthesis, copulas, and imputation-based methods are consistently treated as first-class baselines alongside GAN toolkits, with several surveys stressing that simpler models often outperform deep generators in small-sample or high-dependency regimes (Figueira & Vaz, 2022; Hyrup et al., 2025; Lautrup et al., 2024).

In healthcare-focused methodological surveys, GANs still constitute a central reference family, but recent work increasingly expands the taxonomy to include transformers and diffusion models as independent categories rather than peripheral extensions (Chen et al., 2025b; Ibrahim et al., 2025; Liu et al., 2024; Loni et al., 2025; Perkonoja et al., 2023). Longitudinal patient data reviews, in particular, present multi-family taxonomies that place GANs, VAEs, language models, probabilistic graphical models, and diffusion models on equal footing, reflecting a clear shift away from a purely GAN-centric view of SDG. Reviews centred on privacy and utility, by contrast, often retain a coarser distinction between *GAN-based* and *other* generators, with a long tail of statistical, causal, and proprietary tools(Kaabachi et al., 2025; Liu et al., 2024).

Modality-driven organisation and method-to-modality alignment. A second dominant pattern is to organise generation methods by modality and then articulate a mapping between data type and generator family. In healthcare surveys spanning structured and unstructured data, a common division separates structured data (cross-sectional tabular and time series) from unstructured data (images, text, video), with knowledge-driven, data-driven, and hybrid generators discussed within each category (Ibrahim et al., 2025; Rujas et al., 2025). Across this literature, several stable alignments emerge: GAN variants are predominantly associated with medical imaging and biomedical signals, RNN- or LSTM-based components (often embedded in GANs) with temporal sequences, and classical probabilistic models with mixed categorical and numerical tables (Chen et al., 2025b; Hernandez et al., 2022; Hyrup et al., 2025; Ibrahim et al., 2025; Perkonoja et al., 2023).

For clinical text, transformer-based language models and LLM-centred workflows are increasingly framed as the dominant paradigm, typically via prompting, instruction tuning, and hybrid pipelines rather than architectural novelty alone (Alismail & Lanquillon, 2025; Goyal & Mahmoud, 2024b; Ibrahim et al., 2025; Loni et al., 2025; Rao et al., 2025). In contrast, surveys in vision-centric and surveillance settings describe generation primarily through simulation and rendering pipelines, emphasising 3D modelling, game engines, and compositional techniques as much as neural generators, particularly when precise ground-truth annotations are required for downstream tasks such as pose estimation, tracking, and segmentation (Delussu et al., 2024; Schieber et al., 2024). A similar tool-centric framing appears in video surveillance reviews, which contrast graphics engines and compositional pipelines with GANs and diffusion-based generators, and explicitly connect each approach to the annotation types needed for training perception systems (Delussu et al., 2024).

Workflow- and mechanism-driven framings. Beyond architecture and modality, several surveys organise SDG through the lens of *workflow patterns*. LLM-focused reviews classify generation strategies by prompting, fine-tuning, and specialised architectures, and further distinguish single-model pipelines from multi-agent or cooperative workflows designed to increase controllability, faithfulness, or domain adaptation (Alismail & Lanquillon, 2025; Goyal & Mahmoud, 2024b; Rao et al., 2025). Agentic patterns are presented as a

general design paradigm for textual SDG, distinguishing traditional single-LLM baselines from orchestrated, sequential, and generator-evaluator loops, with the workflow rather than the model family treated as the primary analytical unit (Alismail & Lanquillon, 2025; Goyal & Mahmoud, 2024b).

In privacy-preserving healthcare contexts, surveys adopt a different mechanism-driven logic. Rather than organising methods by workflow or architecture, they classify generation techniques according to how privacy is implemented, such as noise-based, constraint-based, or intrinsic model-level guarantees (Liu et al., 2024). Other healthcare reviews adopt hybrid framings that combine conditional versus unconditional synthesis, data-type categorisation, and model clusters, explicitly linking these choices to regulatory and clinical validation concerns (Ibrahim et al., 2025).

What surveys report in practice and what is often missing. Across domains, surveys consistently report generation methods at the level of model family and named variants, accompanied by high-level architectural notes (for example, generator-discriminator structure, use of RNNs for sequences, U-Net backbones for diffusion) and, in simulation-heavy pipelines, detailed descriptions of tooling and rendering setups (Chen et al., 2025b; Delussu et al., 2024; Hernandez et al., 2022; Ibrahim et al., 2025; Perkonoja et al., 2023; Schieber et al., 2024). However, important implementation details are frequently under-specified. Training configurations, hyperparameter ranges, compute requirements, and convergence behaviour are often missing, which limits reproducibility and weakens cumulative evidence across studies (Chen et al., 2025b; Hernandez et al., 2022; Hyrup et al., 2025; Lautrup et al., 2024; Liu et al., 2024; Perkonoja et al., 2023).

Clinical scoping reviews that focus on motivations and case studies tend to prioritise application narratives over methodological depth, providing only light technical descriptions of generators (Loni et al., 2025; Rujas et al., 2025; Shahul Hameed et al., 2024). Even when modern families such as diffusion models or LLM-based synthesis are covered extensively, guidance on conditioning strategies, handling of discrete structured variables, and systematic benchmarking remains uneven across surveys (Chen et al., 2025b; Hyrup et al., 2025; Ibrahim et al., 2025; Liu et al., 2024).

From GAN-centric to multi-paradigm SDG. Taken together, the seventeen surveys reveal a clear evolution in how generation methods are framed. While GANs remain a central reference point, especially in tabular and imaging contexts, recent reviews increasingly position transformer-based generators, diffusion models, and agentic LLM workflows as first-class paradigms rather than marginal extensions (Chen et al., 2025b; Ibrahim et al., 2025; Liu et al., 2024; Loni et al., 2025; Perkonoja et al., 2023; Rao et al., 2025). In vision and surveillance, diffusion is emerging as a practical route to photorealistic synthesis, often complemented by simulation pipelines when precise labels are required (Delussu et al., 2024; Schieber et al., 2024).

Overall, the secondary literature suggests that future tertiary syntheses of SDG should explicitly separate three analytical dimensions: (i) *generator families* (classical/statistical, GAN, VAE, transformer/LM, diffusion, flows), (ii) *workflow patterns* (single-model vs. agentic or tool-augmented pipelines), and (iii) *modality constraints* (structured vs. unstructured, cross-sectional vs. longitudinal). These dimensions recur most consistently across how generation methods are framed in existing surveys and provide a coherent basis for comparative analysis in large-scale SDG reviews.

5. Discussion

This tertiary study consolidates the scattered secondary literature on SDG and surfaces a consistent pattern across domains, venues, transparency practices, and evaluation norms. In brief, health-centred surveys dominate, journals are the primary outlet family, transparency artefacts are sparse, and quality assessment remains skewed towards resemblance and task utility rather than privacy or diversity. These signals outline both the maturity achieved in specific verticals and the headroom for more balanced and reproducible practice across the field.

5.1. Findings across the research questions

Common objectives and contributions. Across the reviewed surveys, a unifying aim emerges: to map, evaluate, and contextualise the rapidly expanding field of SDG. Regardless of domain or methodology, all studies converge on the objective of advancing data accessibility and model performance while addressing privacy, bias, and fidelity concerns.

Several works (Alismail & Lanquillon, 2025; Figueira & Vaz, 2022; Goyal & Mahmoud, 2024a; Hernandez et al., 2022; Zhang et al., 2025) provide comprehensive overviews of SDG architectures, ranging from classical statistical techniques to modern deep generative models, particularly GANs, VAEs, and LLMs. These studies serve as conceptual foundations for newcomers by organising the literature around the evolution of SDG paradigms.

A second common contribution concerns *benchmarking and evaluation*. Papers such as Chen et al. (2025b), Lautrup et al. (2024) and Kaabachi et al. (2025) propose structured benchmarking protocols that assess model performance along the dimensions of fidelity, utility, and privacy. Their frameworks reveal a pervasive lack of consensus on evaluation metrics, underscoring the field's fragmentation.

Third, a strong emphasis on *privacy and ethical data use* characterises reviews focused on healthcare and biomedical contexts. Works such as Hyrup et al. (2025), Liu et al. (2024), Hernandez et al. (2022) highlight how differential privacy, federated learning, and adversarial regularisation can mitigate re-identification risks. Others, including Shahul Hameed et al. (2024), explore bias mitigation through synthetic sampling and causal modelling, demonstrating the potential of SDG to enhance fairness in data-driven decision-making.

Finally, domain-specific contributions are widespread. Reviews such as Ibrahim et al. (2025), Loni et al. (2025), Rao et al. (2025) and Rujas et al. (2025) concentrate on biomedical and clinical data, reflecting the critical role of SDG in privacy-sensitive domains. Meanwhile, Delussu et al. (2024) and Schieber et al. (2024) emphasise visual and surveillance applications, showcasing how SDG contributes to robustness and scalability in computer vision tasks.

Domains and modalities. The evidence indicates a strong concentration of surveys in healthcare and biomedicine, frequently organised around tabular EHRs, longitudinal patient data, and medical imaging, with comparatively few surveys addressing non-health verticals or structured modalities beyond tabular and vision. This concentration suggests that methodological lessons from health are abundant and transferable, yet also that important application areas such as finance, public administration, education, mobility, energy, manufacturing, and cyber security remain under-synthesised at the survey-of-surveys level.

Publication and institutional patterns. The included set is predominantly published in journals, with scattered contributions from conferences and preprints. Although many affiliations are represented, there is no clear clustering that would indicate an institutional hub for SDG surveys. This dispersion may reflect the cross-disciplinary nature of SDG but it also complicates sustained community standards and shared benchmarks, which typically coalesce around stable venues and consortia.

Transparency and traceability. Only a minority of surveys link public artefacts such as repositories, and explicit protocol registration or versioned research objects are largely absent in the captured fields. This limits re-analysis, hinders cumulative meta-research, and raises barriers to adoption for practitioners who rely on executable exemplars and verifiable extraction sheets. Concretely, the snapshot shows a small number of GitHub-linked artefacts and a larger block reporting none, with insufficient detail to track protocol availability over time.

Quality assessment practice. Metric reporting is extensive yet imbalanced. Fidelity measures and task-grounded utility dominate, while diversity and privacy auditing are less consistently included. Surveys frequently recommend multi-view resemblance checks and standard predictive metrics under Train-on-Synthetic and related protocols, with image-style proxies used when appropriate, but fewer studies pair these with formal guarantees or systematic attack-based audits. This asymmetry leaves residual risk for leakage and mode collapse in high-stakes settings.

5.2. Comparative frameworks and methodological orientations

The reviewed papers differ notably in their frameworks for organising and comparing SDG approaches.

From an *architectural perspective*, Alismail and Lanquillon (2025) contrast single-LLM pipelines with agentic, multi-LLM workflows, emphasising trade-offs between simplicity, quality, and computational cost. Similarly, Figueira and Vaz (2022) structure SDG models according to their underlying architectures, distinguishing traditional machine learning methods from deep generative approaches such as GANs and VAEs.

From a *data modality perspective*, studies like Loni et al. (2025), Hernandez et al. (2022) and Perkonoja et al. (2023) categorise SDG methods based on the type of data generated: tabular, textual, time-series, or longitudinal. These distinctions reveal that model suitability is highly dependent on data structure and temporal characteristics.

In contrast, *evaluation-driven frameworks* proposed by Chen et al. (2025b), Lautrup et al. (2024) and Kaabachi et al. (2025) and others classify methods by performance metrics, proposing systematic evaluation taxonomies. Notably, Kaabachi et al. (2025) extend this idea by introducing a multidimensional taxonomy that incorporates broad and narrow utility, fairness, and privacy, offering a holistic lens for SDG evaluation.

5.3. Emerging trends across studies

Methodological note. The trends discussed in this subsection are inferred from cross-sectional contrasts across surveys published at different points in time, rather than from longitudinal tracking of consistent evaluation or reporting practices. Accordingly, references to shifts or movements should be interpreted as changes in emphasis across survey cohorts, not as causal or continuous temporal evolution. Three cross-cutting trends can be observed across the analysed survey papers.

Shift from GANs to foundation models. While earlier surveys (Figueira & Vaz, 2022; Hernandez et al., 2022; Ruiz-Gándara et al., 2025; Zhang et al., 2026) emphasise GAN-based architectures as the de facto standard for SDG, recent works (Alismail & Lanquillon, 2025; Hellwig et al., 2025; Rao et al., 2025) identify a growing transition toward foundation models and agentic workflows that leverage LLMs for more adaptive and context-aware data generation.

Integration of privacy-preserving mechanisms. The synthesis of privacy-preserving approaches into generative pipelines marks a major methodological advance. Differential privacy, federated learning, and causal modelling (Cai et al., 2025; Hyrup et al., 2025; Liu et al., 2024; Shahul Hameed et al., 2024; Wang et al., 2026a) increasingly appear as integral design components rather than post-hoc safeguards.

Movement toward standardisation and benchmarking. Surveys by Chen et al. (2025b), Lautrup et al. (2024) and Kaabachi et al. (2025) converge on the urgent need for standard benchmarks and reproducible evaluation tools. The creation of open-source frameworks such as SynthEHRella (Chen et al., 2025b) exemplifies a practical step toward transparent, comparable, and collaborative SDG evaluation.

5.4. Divergent research orientations

Despite these convergences, substantial variation persists in scope and emphasis.

Technical versus Conceptual Orientation. Methodologically detailed reviews such as Figueira and Vaz (2022), Goyal and Mahmoud (2024a) and Chen et al. (2025b) delve deeply into model architectures and empirical performance. In contrast, works such as Rujas et al. (2025) and Rao et al. (2025) adopt broader conceptual perspectives, emphasising the societal motivations and translational impact of SDG.

Domain-Specific versus Cross-Domain Focus. Healthcare-centred surveys (Hyrup et al., 2025; Ibrahim et al., 2025; Loni et al., 2025) prioritise privacy and fairness as primary outcomes, while domain-general reviews (Delussu et al., 2024; Schieber et al., 2024) explore SDG as a technical tool for addressing data scarcity in visual and simulation-based contexts.

Practical Implementation versus Theoretical Synthesis. Some studies (Chen et al., 2025b; Lautrup et al., 2024) operationalise their findings through benchmarking software and comparative experiments, whereas others (Goyal & Mahmoud, 2024a; Rujas et al., 2025) remain primarily descriptive, calling for the establishment of unified methodologies and shared datasets.

5.5. Persistent gaps and open challenges

Despite rapid advances, the field remains constrained by several unresolved issues.

- **Lack of standardised metrics:** evaluation across fidelity, utility, and privacy remains fragmented, with little consensus on how to quantify or balance these dimensions consistently across domains (Kaabachi et al., 2025; Lautrup et al., 2024).
- **Limited multi-modal and temporal capabilities:** reviews such as Perkonoja et al. (2023), Loni et al. (2025) and Ibrahim et al. (2025) reveal that few SDG models handle multi-modal or longitudinal data effectively, despite their prevalence in healthcare and sensor-based applications.
- **Ethical and legal uncertainties:** as Alismail and Lanquillon (2025) and Liu et al. (2024) note, reliance on proprietary LLMs and external APIs raises unresolved legal and ethical concerns regarding ownership, consent, and reusability of synthetic data.
- **Computational sustainability:** studies such as Goyal and Mahmoud (2024a) and Alismail and Lanquillon (2025) draw attention to the environmental and financial costs of training large generative models, advocating for resource-efficient workflows.

5.6. Toward a common evaluation baseline

The results motivate a pragmatic baseline that future surveys can endorse to reduce ambiguity. A minimal, domain-agnostic battery would include: (i) fidelity via KS or Chi-square for marginals, a multivariate distance such as Wasserstein or MMD, one divergence (JS or KL), a dependency check such as correlation error or cosine similarity, and RMSE or MSE for key numerics; (ii) utility via AUROC and F1 under matched TSTR/TRTR-style protocols; (iii) diversity via coverage and collapse diagnostics and, in image-like pipelines, FID and Inception Score; and (iv) privacy via at least one membership or attribute inference audit plus a nearest-record or linkability check, and a differential privacy budget where applicable. Adopting this baseline does not preclude domain-specific measures, but it establishes a common floor for comparability.

Overall take-away SDG survey work has matured enough in specific verticals to support clear, actionable guidance, yet remains uneven in domain breadth, transparency, and the balance of evaluation criteria. By consolidating the state of practice and articulating a compact baseline for reporting and auditing, this meta-survey provides a path to more reproducible, comparable, and deployment-ready SDG research.

The consolidated implications are summarised in Table 7.

6. Limitations

Our tertiary analysis provides a consolidated view of surveys on synthetic data generation, yet several constraints affect the interpretation and generalisability of the results.

Scope and coverage. The corpus is bounded by our search string design, database selection, and the publication window adopted for this study. Relevant secondary studies using unconventional terminology, appearing in unindexed venues, or published just beyond our window may not have been captured. Grey literature and non-indexed reports are only partially represented.

Search and screening. We combined automated queries, keyword-based filtering, and deduplication at the title level. While this automation reduced effort, it can propagate false negatives when phrasing is unusual or metadata are incomplete. Title-level deduplication can also fail in the presence of errata or extended versions. Abstract and full-text screening depends on what is exposed by publishers and aggregators, which is not always uniform across venues.

Table 7
Comparative synthesis of common and distinctive contributions across SDG surveys.

Thematic Dimension	Common Focus Across Surveys	Distinctive Contributions
Objectives and Scope	All surveys share the goal of mapping SDG methods, assessing privacy-utility trade-offs, and identifying research gaps. They collectively highlight SDG's potential for privacy protection, fairness, and reproducibility across domains such as healthcare, vision, and NLP.	The focus varies: general SDG frameworks (Alismail & Lanquillon, 2025; Figueira & Vaz, 2022; Goyal & Mahmoud, 2024a) cover multiple architectures and domains, whereas others target healthcare (Hernandez et al., 2022; Ibrahim et al., 2025; Rujas et al., 2025) or vision (Delussu et al., 2024; Schieber et al., 2024). Some emphasise methodological breadth, others depth and benchmarking precision.
Methodological Framing	Most surveys organise SDG by model category (GAN, VAE, Diffusion, LLM) or task (augmentation, anonymisation). They commonly apply PRISMA or scoping methodologies to ensure systematic coverage.	Distinct frameworks appear: architectural (Alismail & Lanquillon, 2025), data-type (Loni et al., 2025; Perkonoja et al., 2023), and evaluation-oriented (Chen et al., 2025b; Lautrup et al., 2024). Some contrast traditional ML approaches with foundation models; others focus on domain-driven adaptations such as healthcare EHR synthesis or multimodal fusion.
Evaluation and Benchmarking	A shared concern is the absence of unified evaluation metrics for fidelity, utility, and privacy. All surveys underline the need for reproducible, comparable benchmarks and open-source tools.	Benchmarking frameworks (Chen et al., 2025b; Lautrup et al., 2024) formalise metrics and datasets, while others propose conceptual taxonomies (Kaabachi et al., 2025). Recent work stresses integrated evaluation pipelines combining privacy risk, data realism, and task-specific utility.
Privacy, Bias, and Fairness	Privacy preservation and bias mitigation are universally recognised priorities. Most surveys frame SDG as a privacy-enhancing alternative to data sharing while noting risks of re-identification and bias propagation.	Different emphases emerge: technical privacy (differential privacy, federated learning) in Liu et al. (2024) an Hyrup et al. (2025); fairness and rebalancing in Shahul Hameed et al. (2024); and ethical governance in Alismail and Lanquillon (2025). Some also discuss regulatory and IP concerns for foundation models.
Domain Contexts	Healthcare dominates due to privacy constraints and data scarcity, but vision and environmental domains are growing. Several surveys recognise SDG's cross-domain versatility for structured, textual, and visual data.	Healthcare reviews (Hyrup et al., 2025; Ibrahim et al., 2025; Loni et al., 2025) emphasize privacy and utility, while vision-oriented ones (Delussu et al., 2024; Schieber et al., 2024) stress realism and scene diversity. Emerging surveys (Rao et al., 2025) advocate multimodal and agentic SDG for clinical data integration.
Emerging Trends	Consensus on the shift from GAN-based to foundation-model and LLM-based SDG. Recent works also highlight agentic workflows, multimodal generation, and ethical-by-design frameworks.	Alismail and Lanquillon (2025) explores agentic critique-refinement loops; Chen et al. (2025b) and Lautrup et al. (2024) advance open benchmarking; Liu et al. (2024) emphasizes privacy-centered design. Some notes on sustainability and computational efficiency as rising priorities.
Gaps and Challenges	All surveys cite fragmented evaluation, limited multimodal benchmarks, and a lack of transparency in large SDG models. Ethical, legal, and computational barriers persist across domains.	Key gaps include standardized metrics (Kaabachi et al., 2025; Lautrup et al., 2024), longitudinal data generation (Perkonoja et al., 2023), and dependence on proprietary LLMs (Alismail & Lanquillon, 2025). Some also highlight reproducibility and energy-efficiency concerns.
Future Directions and Implications	There is strong alignment around developing standardised, interoperable SDG frameworks that balance fidelity, privacy, and fairness. Open collaboration, multimodal synthesis, and ethical governance are common aspirations.	Future visions include cross-domain benchmarking and causal SDG (Chen et al., 2025b; Rao et al., 2025), domain-specific open repositories (Ibrahim et al., 2025), and privacy-compliant LLM workflows (Alismail & Lanquillon, 2025; Liu et al., 2024). Integration of ethical-by-design standards remains a key research imperative.

Publication and language bias. The evidence base is skewed towards English-language publications and towards journals that are well indexed. This can under-represent conference-first disciplines and non-English communities, and can inflate the apparent consensus around practices prevalent in those outlets.

Quality appraisal subjectivity. We used the DARE-4 checklist to appraise candidate secondary studies and applied a decision threshold for inclusion. Although the rubric is simple, item-level judgments can introduce subjectivity, especially where reporting is terse. Different thresholds or tie-breaking rules could shift the final set and the associated descriptive statistics.

Transparency signals are under-resolved. Repository links, extraction sheets, and preregistration are inconsistently reported. Absence of evidence is not evidence of absence, since artefacts may exist but be difficult to locate or may sit behind institutional access. Our tallies should therefore be read as lower bounds.

Temporal interpretation. Observations about trends or shifts in SDG practices are based on cross-sectional comparisons of surveys published in different years, rather than on longitudinal measurement of the same practices over time. As such, these observations indicate relative emphasis rather than documented evolution.

External validity. Our synthesis emphasises what surveys report, not what primary studies necessarily achieve in deployment. Recommendations captured here should be viewed as reflective of secondary literature consensus, which may diverge from best practice in specific industrial or regulatory contexts.

Ethical and legal nuance. Privacy, fairness, and safety claims are often context-dependent. Many surveys provide high-level guidance rather than jurisdiction-specific analyses. Our consolidation inherits this limitation and should not be treated as legal or compliance advice.

Reproducibility of our review. Although we standardised templates and normalised terms, some steps inevitably rely on author judgment. Minor differences in screening decisions, inclusion thresholds, or normalisation rules could yield slightly different counts and classifications.

Taken together, these limitations mean that our quantitative summaries should be read as directional signals rather than definitive measurements. We mitigate over-interpretation by foregrounding qualitative patterns and by recommending a compact, reusable evaluation baseline, but further updates and replications are encouraged as the literature grows.

7. Conclusion and recommendations

7.1. Conclusion

This tertiary study consolidates a fragmented body of secondary research on synthetic data generation, providing a unified and critically appraised overview of how surveys have evolved in scope, methodology, and transparency. By systematically identifying and evaluating 17 methodologically sound secondary studies through the DARE-4 framework, we delineated the contours of the current SDG landscape.

The evidence reveals a clear concentration of survey activity within healthcare and biomedical domains, reflecting the pressing need for privacy-preserving and data-accessible solutions in sensitive environments. At the same time, non-health domains remain comparatively underexplored, and few reviews integrate multi-modal or cross-domain analyses. This asymmetry underscores both the maturity achieved in medical applications and the opportunity for methodological diversification.

From an evaluative standpoint, most surveys converge on a small set of recurrent metrics–chiefly those assessing fidelity and downstream utility–while privacy, fairness, and diversity are less consistently examined. Despite this imbalance, an encouraging trend toward standardisation and benchmarking is emerging, supported by open-source initiatives and structured evaluation frameworks. Transparency, however, remains limited: only a minority of studies release artefacts or protocols, restricting reproducibility and meta-analytic depth.

Overall, the field of SDG survey research stands at a point of consolidation rather than closure. The transition from GAN-based to foundation-model and agentic architectures marks a methodological shift that demands renewed reflection on ethics, sustainability, and governance. By synthesising current practices and exposing structural gaps, this tertiary review offers an evidence-based foundation for advancing SDG toward more reproducible, equitable, and domain-general methodologies.

7.2. Practical implications

This subsection highlights the practical implications derived from our synthesis and clarifies how our findings advance beyond existing SDG surveys. For teams adopting SDG, three take-aways emerge. First, evaluation should be reported as a small, comparable battery rather than a bespoke assortment. At minimum, combine univariate tests for marginals, multivariate distances for joint structure, dependency diagnostics, and task utility under matched training and validation splits. Second, incorporate at least one privacy audit and, where applicable, report a differential privacy budget alongside attack success rates. Third, publish the exact evaluation script and configuration to allow others to reproduce the claimed trade-offs on their own data. These practices align with what the strongest surveys already advocate and would immediately improve cross-study comparability.

7.3. Theoretical implications and directions for future research

Beyond aggregating prior findings, this section articulates theoretical propositions that follow from recurring patterns observed across the analysed SDG surveys.

Building upon these insights, the reviewed literature collectively outlines a roadmap for advancing SDG research and practice.

First, developing **cross-domain benchmarking frameworks** that integrate fidelity, utility, privacy, diversity, and fairness metrics is critical. While diversity and privacy are currently underrepresented in many SDG surveys, our analysis highlights that they should be elevated to first-class outcomes alongside the more frequently operationalised dimensions.

Second, the emergence of **agentic and multi-modal SDG systems** represents a promising avenue for creating adaptable, domain-aware generative pipelines capable of handling complex data distributions.

Third, fostering **open-source ecosystems and shared repositories**, as initiated by Alismail and Lanquillon (2025), Chen et al. (2025b) and Ibrahim et al. (2025), is essential to enhance transparency, collaboration, and community-driven validation.

From a theoretical standpoint, the synthesis supports a set of explicit propositions that extend beyond integrative summarisation. First, evaluation in synthetic data generation cannot be reduced to downstream utility alone; fidelity, privacy, and diversity represent

orthogonal quality dimensions that must be jointly considered to characterise generative validity. Second, the heterogeneity observed across survey taxonomies indicates that SDG is best theorised as an ecosystem of domain- and modality-conditioned practices rather than a single, unified methodological paradigm. Third, transparency and reproducibility emerge not as secondary reporting concerns, but as structural preconditions for meaningful evaluation and cross-study comparability.

Finally, future work should adopt an **ethical-by-design** approach to SDG, embedding privacy, fairness, and interpretability as core design principles rather than auxiliary features. Integrating interdisciplinary perspectives from law, ethics, and data governance will be essential to ensure responsible adoption of SDG technologies across scientific and industrial contexts.

Beyond these recommendations, the synthesis also supports a small set of conceptual propositions at the meta-level of SDG survey research.

Proposition 1 (Evaluation imbalance). Across the analysed surveys, evaluation practices show a systematic imbalance favouring utility and fidelity, while privacy and diversity are less consistently operationalised and reported. This pattern recurs across domains and survey types, indicating a structural bias in how SDG quality is conceptualised.

Proposition 2 (Procedural robustness). The robustness of SDG surveys is driven less by the breadth of models or metrics reviewed than by the explicit specification of evaluation protocols, transparency artefacts, and reproducibility assumptions. Surveys lacking procedural clarity limit cross-study comparability regardless of technical scope.

7.4. Recommendations for future surveys and meta-research

1. **Broaden domain and modality coverage.** Commission surveys that target under-represented verticals and structured modalities such as graphs, geospatial streams, heterogeneous time series beyond health, and non-medical multi-modal pipelines. Cross-domain syntheses should make transferability explicit rather than incidental.
2. **Adopt protocolised, artefact-first reporting.** Release a machine-readable extraction sheet, the screening log, and code for figures and tables under a persistent identifier. Where possible, preregister the survey protocol and include versioning of living artefacts.
3. **Standardise evaluation summaries.** Report the evaluation battery in a compact table that lists metrics, criteria, protocols, and decision thresholds, with a short justification for each. Encourage re-use of an identical template across surveys to build a comparable corpus for future tertiary synthesis.
4. **Elevate privacy and diversity to first-class outcomes.** Privacy auditing and coverage diagnostics should be treated as mandatory components when summarising SDG methods, not as optional appendices. Whenever possible, empirical attacks should be complemented with formal guarantees, and limitations and failure modes should be made explicit rather than implicit.
5. **Map evidence to deployment decisions.** Add practitioner-oriented guidance that connects model families and safeguards to concrete deployment choices, including governance workflows and risk registers for sensitive contexts.

CRediT authorship contribution statement

Navid Nobani: Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Conceptualization; **Giovanni Officioso:** Writing – review & editing, Writing – original draft, Investigation, Data curation; **Filippo Pallucchini:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation; **Giancarlo Sperlì:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Conceptualization; **Fabio Mercorio:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Conceptualization.

Data availability

No data was used for the research described in the article.

Appendix A. Evaluation metrics taxonomy

Evaluation metrics play a pivotal role in assessing the quality, reliability, and real-world applicability of synthetic data. Across recent surveys, four principal perspectives–Utility, Fidelity, Diversity, and Privacy–have emerged as the core pillars of SDG evaluation. These dimensions collectively capture how well synthetic data can (i) support downstream analytical or predictive tasks, (ii) replicate the statistical and structural properties of the original data, (iii) maintain sufficient variability and coverage to avoid mode collapse, and (iv) preserve confidentiality and fairness for individuals or groups. To consolidate the heterogeneous evaluation practices reported in the literature, we systematise existing metrics into a unified taxonomy that reflects their conceptual focus and methodological scope. This taxonomy aims to offer a clear, reproducible reference for both researchers and practitioners, facilitating consistent comparison and transparent reporting across future SDG studies. Fig. A.4 presents a truncated version of the taxonomy.

Utility. The novelty of this taxonomy does not lie in introducing new evaluation metrics, but in providing a normalised cross-survey mapping of evaluation practices, exposing frequency and coverage signals across surveys, and deriving a compact baseline evaluation battery grounded in observed practice.

- **Utility and task performance.** Accuracy, Precision, Recall, F1-score, F2-score, AUC, ROC, AUCPR, Balanced accuracy, Classifier utility, ML parameter comparison, Nearest-neighbour adversarial accuracy, GAN-train, GAN-test, Calibration error, Brier score, Predictive validity, Domain gap measure.

- **Explainability and robustness.** Feature importance stability, XAI rule similarity, Faithfulness, Monotonicity, Incompleteness, Reproducibility tests, Ablation study metrics, Sensitivity analysis, Robustness to perturbations.

Fidelity.

- **Statistical fidelity.** Kolmogorov-Smirnov (KS) test, Chi-squared test, Pearson correlation, Mutual information difference, Jensen-Shannon divergence, Kullback-Leibler (KL) divergence, Wasserstein distance, Maximum Mean Discrepancy (MMD), Hellinger distance, Variation distance, RMSE, MSE, SRMSE, Cosine similarity, Euclidean distance, Mahalanobis distance.
- **Textual and linguistic quality.** BLEU, ROUGE (n-gram), METEOR, BERTScore, CLIP score, Perplexity, BM25, Recall@k, Factual consistency, Faithfulness, Coherence, Readability, Factual correctness.
- **Temporal consistency.** Dynamic Time Warping (DTW), Time Warp Edit Distance (TWED), Autocorrelation function, Cross-correlation, Mean-over-time RMSE, Temporal trend preservation, Latent temporal statistics, Cycle preservation, Next-step prediction accuracy.
- **Visual and structural quality.** Inception Score (IS), Fréchet Inception Distance (FID), Kernel Inception Distance (KID), Structural Similarity Index (SSIM), Multi-scale SSIM (MS-SSIM), Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS), Feature Similarity Index (FSIM), Universal Quality Index (UQI), Contrast Noise Ratio (CNR), Visual inspection, Expert scoring.

Diversity.

- **Diversity and coverage.** Coverage ratio, Diversity score (DS), Category coverage, Mode collapse ratio, Sample diversity index, Feature distribution similarity (FDS), Pairwise correlation difference, Attribute entropy, Correlation matrix distance (CMD).
- **Domain-specific validity.** Clinical correctness, Kaplan-Meier divergence, Concordance index (C-index), Genomic Synthetic Fidelity (GSF), Clinical Synthetic Fidelity (CSF), Expert evaluation, Country or sector indicators, Domain-specific resemblance.

Privacy.

- **Privacy and disclosure risk.** Differential privacy ϵ, k-anonymity, ℓ-diversity, t-closeness, Membership inference risk, Attribute inference risk, Identity disclosure risk, Distance to closest record (DCR), Over-fitting gap, Privacy leakage rate, Re-identification risk, Privacy-utility trade-off.
- **Fairness and bias.** Statistical Parity Difference (SPD), Disparate Impact, Equal Opportunity, Equalised Odds, Fairness prediction score, Demographic parity, True positive rate parity, Fairness violation index.

Please note that coverage ratio is assigned to the diversity dimension because it is predominantly used in SDG surveys to quantify how broadly the synthetic data spans the real data space. In contrast, recall is assigned to utility, as it is typically reported as part of downstream predictive performance evaluation.

What is reported most often ? Table 6 lists the top metrics by total mentions in the aggregate. The most frequent items are general-purpose predictive metrics (Accuracy, F1, Recall, AUROC) and classical distributional distances/tests (Wasserstein/EMD, JS/KL divergence, RMSE, MSE, KS, Chi-square), with image-style proxies (FID, Inception Score) also prominent. In this appendix, *mentions* denotes the number of secondary studies that explicitly report or discuss a given metric, rather than the frequency of use in the underlying primary studies. This reflects the tertiary scope of our analysis, which focuses on reporting practices in surveys.

How are these metrics used in practice ? **Accuracy, F1, Precision, Recall, AUROC (Utility).** In surveys, these are almost always applied to *downstream predictive tasks* to test whether models trained with synthetic data behave comparably to models trained with real data. Two common protocols are: (i) *Train on Synthetic, Test on Real* (TSTR, akin to GAN-train) and *Train on Real, Test on Synthetic* (TRTS/TSTS, akin to GAN-test) to gauge transfer and over/under-fitting (Figueira & Vaz, 2022; Perkonoja et al., 2023). AUROC, Accuracy, F1, Precision and Recall are then compared to their real-data baselines to quantify utility gaps. In classification-heavy domains (e.g., EHR risk prediction or vision benchmarks), AUROC and F1 are preferred to balance class imbalance and discrimination (Delussu et al., 2024; Kaabachi et al., 2025; Paulin & Ivasic-Kos, 2023; Perkonoja et al., 2023).

Wasserstein/EMD, MMD, JS/KL, KS, Chi-square, RMSE/MSE (Fidelity). These quantify *statistical resemblance* between real and synthetic data: univariate alignment (KS for continuous, Chi-square for categorical), marginal moment errors (RMSE/MSE), and multivariate shifts (Wasserstein/EMD, MMD, JS/KL). Surveys recommend using a *portfolio* spanning univariate, bivariate (correlation/covariance errors), and multivariate distances to avoid false confidence from any single statistic (Kaabachi et al., 2025; Perkonoja et al., 2023). Wasserstein/EMD and MMD are frequently highlighted for capturing higher-order structure beyond simple marginals.

FID, Inception score (Diversity/fidelity proxies). Originally from image synthesis, surveys report these as *embedding-space* proxies for realism and coverage; they correlate with visual quality and mode coverage (FID lower is better; IS higher is better) and are commonly reported in vision or image-like pipelines (Delussu et al., 2024; Paulin & Ivasic-Kos, 2023). Some surveys caution that FID/IS should be complemented with task-specific or domain statistics to avoid over-reliance on a single proxy (Kaabachi et al., 2025).

Cosine similarity, correlation errors, PCA/t-SNE overlays (Fidelity diagnostics). These are used to *diagnose representation similarity* (e.g., pairwise feature relations, or overlap in low-dimensional projections), especially in tabular and text-like data. They help validate that dependencies are preserved, not just marginals (Kaabachi et al., 2025; Perkonoja et al., 2023).

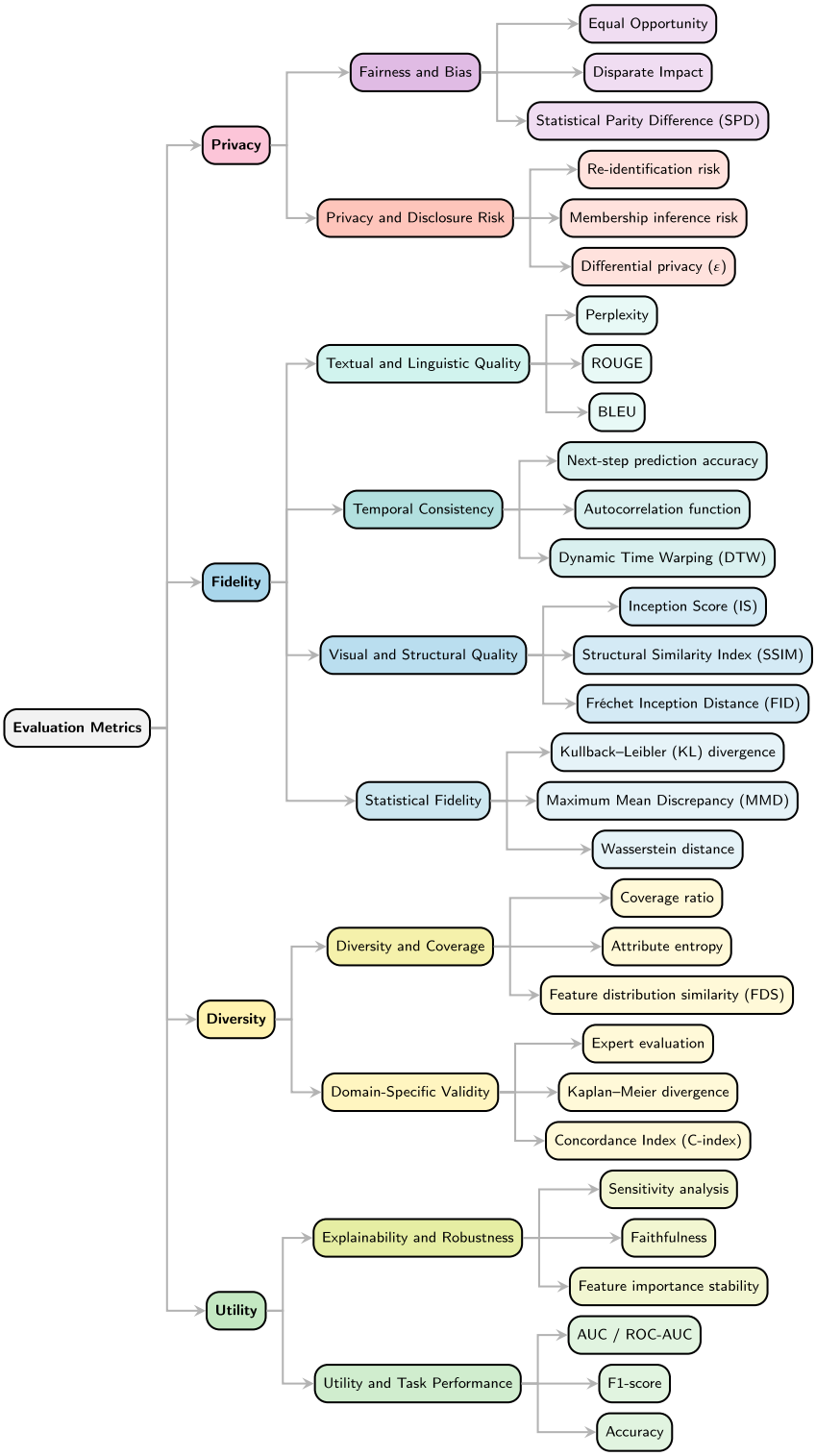

Fig. A.4. Taxonomy of evaluation metrics for synthetic data generation. For visual clarity, only the top 3 most frequently used metrics across the analysed studies are displayed.

Coverage/diversity metrics. Where reported, these include coverage ratios, collapse diagnostics, or image-oriented measures like MS-SSIM/LPIPS alongside FID/IS to ensure modes are not missing. A smaller subset of surveys also mentions α-Precision/β-Recall (fidelity/diversity) and "authenticity" to detect over-fitting (Figueira & Vaz, 2022; Miró-Nicolau et al., 2025; Paulin & Ivasic-Kos, 2023).

Privacy metrics. Though least frequent overall, surveys increasingly report: *membership inference* (attack precision/recall or success rate), *attribute inference*, *linkability/closest-record distance*, and classical *re-identification risk* estimators. When available, *differential privacy* is summarised with the privacy budget ϵ (and sometimes δ), but many surveys emphasise *empirical* risk auditing over formal guarantees (Hyrup et al., 2025; Kaabachi et al., 2025). Recommended practice is to pair at least one *formal* (DP ϵ if applicable) with one or more *empirical* attack-based audits (e.g., membership/attribute inference) (Hyrup et al., 2025).

References

Abowd, J. M. (2018). The US census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2867).

Akpinar, M. H., Sengur, A., Salvi, M., Seoni, S., Faust, O., Mir, H., Molinari, F., & Acharya, U. R. (2024). Synthetic data generation via generative adversarial networks in healthcare: A systematic review of image-and signal-based studies. *IEEE Open Journal of Engineering in Medicine and Biology*, *6*, 1–17.

Alaa, A., Van Breugel, B., Saveliev, E. S., & Van Der Schaar, M. (2022). How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning* (pp. 290–306). PMLR.

Alismail, A., & Lanquillon, C. (2025). A survey of LLM-based methods for synthetic data generation and the rise of agentic workflows. In *International conference on human-computer interaction* (pp. 119–135). Springer.

Andreini, P., Tanfoni, M., Bonechi, S., & Bianchini, M. (2026). Leveraging synthetic data for zero-shot and few-shot circle detection in real-world domains. *Pattern Recognition*, *172*, 112407. https://doi.org/10.1016/j.patcog.2025.112407

Baowaly, M. K., Lin, C.-C., Liu, C.-L., & Chen, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, *26*(3), 228–241.

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(1), 4.

Bouguettaya, S., Pupo, F., Chen, M., & Fortino, G. (2025). A meta-survey of generative AI in education: Trends, challenges, and research directions. *Big Data and Cognitive Computing*, *9*(9), 237.

van Breugel, B., Liu, T., Oglic, D., & van der Schaar, M. (2024). Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, *2*(12), 991–1004.

Cai, X., Sun, Y., Lin, Z., Li, R., & Cai, T. (2025). Differentially private synthetic data generation for robust information fusion. *Information Fusion*, *124*, 103373. https://doi.org/10.1016/j.inffus.2025.103373

Camino, R. D., Hammerschmidt, C. et al. (2018). Generating multi-categorical samples with generative adversarial networks. In *Icml 2018 workshop on theoretical foundations and applications of deep generative models* (pp. 1–10).

Chatterjee, S., Hazra, D., & Byun, Y.-C. (2025). Gan-based synthetic time-series data generation for improving prediction of demand for electric vehicles. *Expert Systems with Applications*, *264*, 125838.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T.-H., Chou, H.-H., Zou, S., Chen, Y.-H., & Hsieh, S.-Y. (2025a). System-level integration of deep learning and computer vision for contact ring seal defect detection in semiconductor manufacturing. *Expert Systems with Applications*, (p. 129551).

Chen, X., Wu, Z., Shi, X., Cho, H., & Mukherjee, B. (2025b). Generating synthetic electronic health record data: A methodological scoping review with benchmarking on phenotype data and open-source software. *Journal of the American Medical Informatics Association*, *32*(7), 1227–1240.

Chen, Z., Qiao, F., Chen, H., Zhang, W., Ren, P., Lin, S., Sun, J., & Li, Y. (2026). Custom training data: Aligning model distribution via prefix-guided preference data synthesis. *Information Processing & Management*, *63*(1), 104337. https://doi.org/10.1016/j.ipm.2025.104337

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference* (pp. 286–305). PMLR.

Delussu, R., Putzu, L., & Fumera, G. (2024). Synthetic data for video surveillance applications of computer vision: A review. *International Journal of Computer Vision*, *132*(10), 4473–4509.

Dhinakaran, D., Kumar, N. J., Ponnuviji, N. P. et al. (2025). Safeguarding confidentiality and privacy in cloud-enabled healthcare systems with spectrasafe encryption and dynamic k-anonymity algorithm. *Expert Systems with Applications*, *279*, 127584.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). Carla: An open urban driving simulator. In *Conference on robot learning* (pp. 1–16). PMLR.

Egger, J., Pepe, A., Gsaxner, C., Jin, Y., Li, J., & Kern, R. (2021). Deep learning–a first meta-survey of selected reviews across scientific disciplines, their commonalities, challenges and research impact. *PeerJ Computer Science*, *7*, e773.

Feng, Y., Li, L., Qin, X., & Zhang, B. (2025). Improving event representation learning via generating and utilizing synthetic data. *Information Processing & Management*, *62*(4), 104083. https://doi.org/10.1016/j.ipm.2025.104083

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733.

Gayathri, R. G., Sajjanhar, A., & Xiang, Y. (2024). Hybrid deep learning model using SPCAGAN augmentation for insider threat analysis. *Expert Systems with Applications*, *249*, 123533.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 139-144.

Goyal, M., & Mahmoud, Q. H. (2024a). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, *13*(17). https://doi.org/10.3390/electronics13173509

Goyal, M., & Mahmoud, Q. H. (2024b). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, *13*(17), 3509.

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). Prisma2020: An r package and shiny app for producing prisma 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews*, *18*(2), e1230.

Hellwig, N. C., Fehle, J., & Wolff, C. (2025). Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Systems with Applications*, *261*, 125514.

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

Hernández-Ferrándiz, D., Pantrigo, J. J., & Cabido, R. (2025). A parametric synthetic data generator for training learning-based sperm analysis systems. *Expert Systems with Applications*, *271*, 126614.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Huang, Y., Wu, S., Gao, C., Chen, D., Zhang, Q., Wan, Y., Zhou, T., Xiao, C., Gao, J., Sun, L. et al. (2024). Datagen: Unified synthetic dataset generation via large language models. In *The twelfth international conference on learning representations (ICLR 2024)* p. . Conference paper https://openreview.net/forum?id=F5R0lG74Tu.

Hyrup, T., Lautrup, A. D., Zimek, A., & Schneider-Kamp, P. (2025). A systematic review of privacy-preserving techniques for synthetic tabular health data. *Discover Data*, *3*(1), 1–32.

Ibrahim, M., Al Khalil, Y., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., Elen, B., Ertaylan, G., & Dumontier, M. (2025). Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*, *189*, 109834.
Jordon, J., Yoon, J., & Van Der Schaar, M. (2018). Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.
Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. *NPJ Digital Medicine*, *8*(1), 60.
Kalibatiene, D., & Jolanta, M. (2025). From manual to automated systematic review: Key attributes influencing the duration of systematic reviews in software engineering. *Computer Standards & Interfaces*, *96* (p. 104073).
Kotti, Z., Galanopoulou, R., & Spinellis, D. (2023). Machine learning for software engineering: A tertiary study. *ACM Computing Surveys*, *55*(12). https://doi.org/10.1145/3572905
Kurdi, G. (2025). Supporting systematic review workflows in computing: a functional evaluation of electronic databases. *Journal of Documentation*, *81* (pp. 1–25).
Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Kamp, P. (2024). Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data. *ACM Computing Surveys*, *57*(4), 1–38.
Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review*, *54*(7), 4997–5053.
Little, C., Elliot, M., & Allmendinger, R. (2023). Federated learning for generating synthetic data: A scoping review. *International Journal of Population Data Science*, *8*(1), 2158.
Liu, Y., Acharya, U. R., & Tan, J. H. (2024). Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation. *Computer Methods and Programs in Biomedicine*, *260*, (p. 108571).
Loni, M., Poursalim, F., Asadi, M., & Gharehbaghi, A. (2025). A review on generative AI models for synthetic medical text, time series, and longitudinal data. *npj Digital Medicine*, *8*(1), 281.
Miró-Nicolau, M., Jaume-i Capó, A., & Moyà-Alcover, G. (2025). A comprehensive study on fidelity metrics for XAI. *Information Processing & Management*, *62*(1), 103900. https://doi.org/10.1016/j.ipm.2024.103900
Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Bmj*, *339*.
van Mourik, F., Jutte, A., Berendse, S. E., Bukhsh, F. A., & Ahmed, F. (2024). Tertiary review on explainable artificial intelligence: Where do we stand? *Machine Learning and Knowledge Extraction*, *6*(3), 1997–2017.
Nadăş, M., Dioşan, L., & Tomescu, A. (2025). Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, *13*.
Nikolenko, S. I. et al. (2021). Synthetic data for deep learning (vol. 174). Springer.
Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, *74*, 1–26.
Paulin, G., & Ivasic-Kos, M. (2023). Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial Intelligence Review*, *56*(9), 9221–9265.
Perkonoja, K., Auranen, K., & Virta, J. (2023). Methods for generating and evaluating synthetic longitudinal patient data: a systematic review. arXiv preprint arXiv:2309.12380.
Ping, H., Stoyanovich, J., & Howe, B. (2017). Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th international conference on scientific and statistical database management* (pp. 1–5).
Rao, H., Liu, W., Wang, H., Huang, I., He, Z., Huang, X. et al. (2025). A scoping review of synthetic data generation for biomedical research and applications. arXiv preprint arXiv:2506.16594.
Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441.
RezvaniNejad, M., & Yameqani, A. S. (2025). Wdae-gan: A hybrid dual autoencoder and generative adversarial framework with wavelet denoising for credit card fraud detection. *Expert Systems with Applications*, (p. 130078). https://doi.org/10.1016/j.eswa.2024.130078
Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.
Ruiz-Gándara, A., Casales-García, V., & González-Abril, L. (2025). Artificial generation of survey data on the expected bitterness of beer. *Expert Systems with Applications*, *275*, 126950. https://doi.org/10.1016/j.eswa.2025.126950
Rujas, M., del Moral Herranz, R. M. G., Fico, G., & Merino-Barbancho, B. (2025). Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications. *International Journal of Medical Informatics*, *195*, 105763.
Schieber, H., Demir, K. C., Kleinbeck, C., Yang, S. H., & Roth, D. (2024). Indoor synthetic data generation: A systematic review. *Computer Vision and Image Understanding*, *240*, 103907.
Shahul Hameed, M. A., Qureshi, A. M., & Kaushik, A. (2024). Bias mitigation via synthetic data generation: A review. *Electronics*, *13*(19), 3909.
Shi, L., Giunchiglia, F., Wang, H., Cheng, Y., Song, R., Shi, D., Diao, X., & Xu, H. (2026). From text mining to intelligent debate: Task frameworks and technological evolution in computational argumentation. *Information Processing & Management*, *63*(2, Part B), 104465. https://doi.org/10.1016/j.ipm.2025.104465
Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on security and privacy (SP)* (pp. 3–18). IEEE.
Singh, A. K., Rao, A., Chattopadhyay, P., Maurya, R., & Singh, L. (2024). Effective plant disease diagnosis using vision transformer trained with leafy-generative adversarial network-generated images. *Expert Systems with Applications*, *254*, 124387.
Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data–anonymisation groundhog day. In *31st USENIX security symposium (USENIX security 22)* (pp. 1451–1468).
Steinbacher, M., Raddant, M., Karimi, F., Camacho Cuena, E., Alfarano, S., Iori, G., & Lux, T. (2021). Advances in the agent-based modeling of economic and social behavior. *SN Business & Economics*, *1*(7), 99.
Stenger, M., Leppich, R., Foster, I., Kounev, S., & Bauer, A. (2024). Evaluation is key: A survey on evaluation measures for synthetic time series. *Journal of Big Data*, *11*(1), 66.
Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, *33*, 18583–18599.
Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238.
Wang, T., Chen, X., Liu, Z., & Yao, S. (2026a). Diffushield: Flexible privacy-preserving synthetic face generation via generative diffusion model. *Information Fusion*, *125*, 103451. https://doi.org/10.1016/j.inffus.2025.103451
Wang, X., Chen, C., Yang, F., Gong, X., & Zhao, S. (2026b). Srmer: Synthetic-to-real multimodal emotion recognition. *Information Fusion*, *127*, 103869. https://doi.org/10.1016/j.inffus.2025.103869
Wang, Z., & Sun, J. (2022). Promptehr: Conditional electronic healthcare records generation with prompt learning. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 2873–2885).
Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, *32*, 7335–7345
Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, 244–255.
Zhang, H., Guo, H., Bai, H., Zhang, C., & Li, L. (2026). Maml-based temporal supervised information maximizing gan for few-shot time series data generation. *Expert Systems with Applications*, *297*, 129342. https://doi.org/10.1016/j.eswa.2025.129342
Zhang, H., Jing, Y., Zhang, F., Li, Z., Wang, X. S., Chen, Z., & Lv, C. (2025). TabTransGAN: A hybrid approach integrating GAN and transformer architectures for tabular data synthesis. *Information Processing & Management*, *62*(5), 104220. https://doi.org/10.1016/j.ipm.2025.104220
Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., & Xiao, X. (2017). Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, *42*(4), 1–41.