NAR Genomics and Bioinformatics, 2026, **8**, lqag012
https://doi.org/10.1093/nargab/lqag012
AI Methods and Foundation Models for Biology

OXFORD

An ELIXIR scoping review on domain-specific evaluation metrics for synthetic data in life sciences

Styliani-Christina Fragkouli[1,2], Somya Iqbal[3], Lisa Crossman[4,5], Barbara Gravel[6,7,8], Nagat Masued[9], Mark Onders[10], Devesh Haseja[11], Alex Stikkelman[12], Alfonso Valencia[9,13], Tom Lenaerts[6,7,8], Fotis Psomopoulos[2], Pilib Ó Broin[11], Núria Queralt-Rosinach[12], Davide Cirillo[9,*]

[1]Department of Biology, National and Kapodistrian University of Athens, Athens 15772, Greece
[2]Institute of Applied Biosciences, Centre for Research & Technology Hellas, Thessaloniki 57001, Greece
[3]University of Edinburgh, Usher Institute, Nine BioQuarter, Edinburgh EH16 4UX, United Kingdom
[4]SequenceAnalysis.co.uk, Norwich Research Park, Norwich NR4 7UG, United Kingdom
[5]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom
[6]Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles-Vrije Universiteit Brussel, Brussels 1050, Belgium
[7]Machine Learning Group, Université Libre de Bruxelles, Brussel 1050, Belgium
[8]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussel 1050, Belgium
[9]Barcelona Supercomputing Center (BSC), Plaça Eusebi Güell, 1-3, Barcelona 08034, Spain
[10]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, United States
[11]School of Mathematical and Statistical Sciences, College of Science and Engineering, University of Galway, Galway, H91 TK33, Ireland
[12]Department of Human Genetics, Leiden University Medical Center, Leiden 2333, The Netherlands
[13]ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

*To whom correspondence should be addressed. Email: davide.cirillo@bsc.es

Abstract

Synthetic data (SD) has become an increasingly important asset in the life sciences, helping address data scarcity, privacy concerns, and barriers to data access. Creating artificial datasets that mirror the characteristics of real data allows researchers to develop and validate computational methods in controlled environments. Despite its promise, the adoption of SD in life sciences hinges on rigorous evaluation metrics designed to assess their fidelity and reliability. To explore the current landscape of SD evaluation metrics in distinct life sciences domains, the ELIXIR Machine Learning Focus Group performed a systematic review of the scientific literature following the PRISMA guidelines. Six critical domains were examined to identify current practices for assessing SD. Findings reveal that, while generation methods are rapidly evolving, systematic evaluation is often overlooked, limiting researchers' ability to compare, validate, and trust synthetic datasets across different domains. This systematic review underscores the urgent need for robust, standardized evaluation approaches that not only bolster confidence in SD but also guide its effective and responsible implementation. By laying the groundwork for establishing domain-specific yet interoperable standards, this scoping review paves the way for future initiatives aimed at enhancing the role of SD in scientific discovery, clinical practice and beyond.

Introduction

Synthetic data (SD) is gaining traction across fields to address the scarcity of real data (RD), privacy concerns, and access limitations, while enabling robust machine learning (ML) and artificial intelligence (AI) applications and benchmarking studies [1]. By generating artificial datasets, SD allows researchers to develop and validate computational methods in controlled environments by facilitating the evaluation of statistical models and predictive algorithms. However, as technologies continue to advance, the volume and complexity of biological and clinical data have grown significantly. While this expansion presents new opportunities for discovery, it also introduces challenges in ensuring data availability, quality and privacy protection. A promising solution lies in SD, which enables the creation of high-fidelity datasets that bypass privacy restrictions while preserving essential statistical properties [2].

Beyond practical considerations, SD is particularly valuable in contexts where strict data governance regulations limit access to RD. Legal frameworks like the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) impose stringent requirements on sharing and using sensitive data in healthcare research [3]. SD offers a potentially viable solution by providing synthetic versions of patient records that maintain statistical properties of RD while reducing privacy risk. This facilitates inter-institutional collaborations by loweing bureaucratic barriers and ensuring compliance with ethical and legal standards [4].

The effective integration of SD into scientific and clinical workflows depends on its ability to accurately mirror the structure of RD. Evaluating SD quality remains a major challenge, as it requires assessing both structural and inferential fidelity in addition to its practical utility for downstream anal-

Received: September 27, 2025. **Revised:** January 7, 2026. **Accepted:** January 19, 2026

yses. Evaluation needs vary by application: some focus on improving ML/AI model performance, while others aim to uncover novel patterns for hypothesis generation. Consequently, there is no one-size-fits-all evaluation method. Assessment strategies must be tailored to the goals and context of each use case [5].

To address these challenges, robust evaluation recommendations and best practices are essential for assessing SD across diverse applications. Much like ongoing initiatives in areas such as cross-domain FAIR data assessment and the Critical Assessment of Structure Prediction (CASP), the development of guidelines for SD could progressively lead to a more structured and coherent framework for evaluating prediction methods across domains. Such guidelines will enhance the trustworthiness of SD, ensuring its reliability in biomedical research and clinical decision-making [6]. Furthermore, as SD continues to mature, its successful integration into healthcare will require more rigorous evaluation recommendations and clear regulatory guidelines as well as ongoing adaptation to uphold its reliability and ethical implementation in medical research and practice.

The ELIXIR ML Focus Group is part of ELIXIR, the European infrastructure for life sciences, which brings together over 250 research organizations across 24 countries. In this work, a collaborative effort was led by the ELIXIR ML focus group, specifically the Task Force on SD, in order to establish an evaluation set of guidelines for SD in life sciences within Europe. A scoping review based on the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) standards was carried out by experts in the Task Force across different domains adjacent to life sciences in order to identify the current evaluation metrics that are utilized to assess SD (individual flow diagrams for each domain are available in Supplementary Figs S1–S6). The key domains selected for the review are genomics, transcriptomics, proteomics, phenomics, imaging, and electronic health records (EHRs), ensuring a thorough and comprehensive investigation of the topic. Figure 1 displays a word cloud visualization of the metrics identified across the six domains, illustrating the diversity of the metrics in a visual format (with all metrics visualized collectively in Supplementary Fig. S7).

This scoping review maps current methodologies and highlights gaps in evaluation strategies to support the systematic assessment of SD quality, reliability, and utility in life sciences research. While existing studies provide in-depth surveys of SD quality metrics in biomedicine, they tend to focus on specific aspects, like privacy and utility [7], benchmarking general-purpose implementations [8], or particular healthcare domains collectively [9]. To our knowledge, this is the first comprehensive review that systematically examines evaluation metrics for SD across multiple application domains in the life lciences, addressing each domain in detail.

Materials and methods

Inclusion criteria

Following the PRISMA [10] guidelines, we devise a systematic approach to identifying relevant literature on SD across various domains, ensuring the inclusion of high-quality, peer-reviewed, and openly accessible research. To identify relevant studies, we conducted a comprehensive literature search using inclusion criteria as shown in the PRISMA diagram for each domain (Supplementary Figs S1–S6 and Supplementary Material).

The selection of studies for this scoping review was guided by predefined inclusion criteria to ensure a comprehensive yet focused assessment of evaluation metrics for SD. Literature searches were conducted across multiple databases, including PubMed, SCOPUS, and Google Scholar, while Web of Science (WoS) and DOAJ.org as optional sources as their specialized focus might not align with all research requirements. To maintain relevance to current research trends only studies published within the last ten years were included.

All selected publications were in English and accessible as Open Access articles to facilitate transparency and reproducibility. Furthermore, only peer-reviewed journal publications were considered to ensure the scientific rigor and credibility of the included studies. Given the large volume of unspecific results typically obtained by Google Scholar using queries, those search results were ranked by relevance [11], with the top 20 articles reviewed for potential inclusion. While this approach efficiently captures commonly referenced literature, we acknowledge that relevance rankings are algorithmically determined and not publicly documented. To mitigate this limitation, all retrieved records were manually inspected to ensure representativeness and to confirm that key, highly cited papers in each domain were included. This approach ensures that the review captures high-quality, methodologically sound literature that contributes meaningfully to the evaluation of SD (see Table 1).

Search queries

A diverse set of queries were formulated, tailored to two key aspects: first, to accurately reflect the specific domain of the evaluation, and second, to align with the search engine used in each case. All queries incorporated the terms “synthetic” and “evaluation”, followed by domain-specific keywords to ensure relevance for each domain. The primary domains deliberately chosen for exploration include genomics, transcriptomics, proteomics, phenomics, imaging and EHRs in order to ensure a comprehensive investigation of the topic. For further details please see the Supplementary Material file.

Identification of relevant publications

The PRISMA 2020 flow diagram, depicted for each domain search in Supplementary Figs S1–S6, provides a structured representation of the study selection process in systematic reviews that involve searches from databases and registers. It consists of three main stages: identification, screening and inclusion. During the identification stage, records are collected from multiple sources including databases and registers. Before the screening process begins, duplicate records are removed and additional exclusions may occur (as described later in the limitations section) due to automation tools or other filtering criteria.

During the screening stage, the remaining records undergo an initial review to determine their relevance. At this step, reports are sought for retrieval and any that cannot be accessed are documented. The retrieved reports are then assessed for eligibility and exclusions are made based on specific reasons, which are clearly recorded (i.e. not relevant for evaluating SD).

Finally, in the inclusion stage, the studies that meet all criteria are incorporated into the systematic review and the total number of included reports is documented. This method

Figure 1. Word cloud visualization of the identified evaluation metrics in six application domains in the life sciences. Legend: Word cloud visualization of the metrics identified across genomics, transcriptomics, proteomics, phenomics, imaging, and EHRs. The figure illustrates the diversity of terms used as evaluation metrics in different domains of SD research. Alt text: A word cloud showing various metric names of different sizes, representing how frequently each metric appears across six SD domains including genomics, transcriptomics, proteomics, phenomics, imaging, and EHRs.

Table 1. The inclusion criteria used for our comprehensive literature search in order to identify relevant studies

Criteria	Details
Databases	PubMed, SCOPUS, Google Scholar, Web of Science (WoS) (optional), DOAJ.org (optional)
Timeframe	Studies published in the last 10 years
Language	English
Accessibility	Open Access
Peer-reviewed	Articles from peer-reviewed journals
Ranking	In the case of Google Scholar, top 20 results ranked by relevance

ensures transparency and reproducibility by visually summarizing the number of records processed at each stage and detailing exclusions, thus enhancing clarity. All domain-related PRISMA flowcharts can be found in the Supplementary Material (Supplementary Figs S1–S6).

The complete list of all extracted evaluation metrics, their assigned domains, and the corresponding source publications are provided in the supplementary dataset (Supplementary Files 1 and 2), enabling full transparency and reproducibility of the results.

Limitations

During the search process, several observations and challenges were identified. The term "synthetic" often refers to synthetic biology rather than SD generation, leading to unrelated findings. Additionally, many studies mention "synthetic data" in abstracts without detailing evaluation metrics or generation tools. The number of relevant papers varies significantly by domain, with fewer studies found in highly specialized fields such as synthetic phenomics. Furthermore, additional care was necessary to distinguish studies focused on synthetic peptides and data derived from them, as opposed to computationally generated SD in proteomics. Although the majority of the literature search and curation were conducted in the first half of 2024, we continuously reviewed additional papers in each domain and found no additional relevant metrics beyond those that we had already surveyed. This strengthens the authors' confidence that this review accurately reflects the current state of the art.

Results

Synthetic genomics

Domain overview

Synthetic genomics involves using computational tools to design, analyze and simulate synthetic genomes. It aims to create accurate representations of genetic sequences by leveraging bioinformatics and computational biology along with wet-lab expertise. Some key concepts include designing genomes computationally, modelling the functionality of the synthesized genomes, utilizing this type of data for the development and the benchmarking of bioinformatics tools. Designing optimized genomic sequences for specific functions is driving advancements in vaccine development [12] as well as synthetic and evolutionary biology [13]. However, challenges such as data limitations, biosecurity risks, and model interpretability persist, highlighting the need for solutions to ensure safety and ethical compliance.

Review outcomes

After the examination of 769 articles, 7 publications were identified appropriate for our effort. These selected articles fall under two main categories, reviewing genomics simulating tools and suggesting guidelines for the evaluation [14]. Due to the broad use of the term 'synthetic,' many irrelevant publications (either outside of the inclusion criteria or due to technical aspects that were unsuitable for this review) were initially identified, requiring us to refine the queries with more appropriate search terms. Furthermore, the term "synthetic DNA" typically refers to data that is generated in wet labs rather than in-silico.

Evaluation metrics

The reported metrics can be divided into two groups: those that can be found and utilized in a standard bioinformatics pipeline and analysis, and those tailored to specific use cases.

- **Bioinformatics related metrics** [15–17]: Proportion of reads on each chromosome, coverage, type of reads, paired-end fragment length, alignment quality, rates of sequencing error, substitution, insertion, and deletion rates, quality scores for reads, GC-coverage bias, computational cost, mapping sensitivity of bases in reads, read mapping precision, Phred score.
- **Use case related metrics** [18]: Principal component analysis (PCA) to evaluate qualitatively by visualization and supervised discriminators for quantitative evaluation.
- **Statistics based metrics** [19]: Mean of significant variables in case/control, dispersion parameter of the negative binomial distribution, and signal to noise ratio, sample size, number of variables, mean of insignificant variables.

Synthetic transcriptomics

Domain overview

Synthetic transcriptomics is a key research area within the field of SD generation for biological studies, enabling the creation and analysis of artificial RNA-Seq data that mirrors the characteristics of RD [20]. Such SD can aid in standard downstream transcriptomics analyses like the identification of differentially expressed genes and isoforms and pathway enrichment analysis [6, 21], as well as more advanced applications, such as the identification of clinically relevant biomarkers. Synthetic transcriptomic data is often used in the context of oversampling approaches to correct imbalance between different sample groups within a dataset. This can help to compensate for bias due to underrepresentation of smaller sample groups, lead to increased statistical power, and reduce the effect of outliers [22]. A key factor contributing to the increasing use of synthetic transcriptomic data relates to the fact that the process of RNA isolation, library preparation, and sequencing to obtain RD is relatively time, labour, and cost intensive, making the alternative approach of generating realistic SD attractive in terms of both efficiency and conservation of resources [23]. This is particularly true for the adoption of synthetic transcriptomic data for the training of ML algorithms where adequate sample size, and restricted access to RD can be a concern.

Review outcomes

The search criteria for 'synthetic transcriptomic' or 'synthetic RNA' resulted in 424 articles (139 in Scopus, 23 in Pubmed, 16 in Web of Science, 246 in Google Scholar). After removing duplicate records ($n = 14$), additional filtering was carried out to exclude articles not meeting other criteria (articles in other languages, preprints or conference proceedings ($n = 120$), articles found to be >10 years old ($n = 3$), articles belonging to a different domain ($n = 33$). Ultimately, 84 articles remained for curation and the resulting evaluation metrics from those that report them are summarized below.

A key aspect of SD evaluation in this domain is the distinction between intrinsic and extrinsic metrics. Extrinsic evaluation refers to performance-based metrics that assess how well

a model behaves when trained on SD compared to RD. Common extrinsic metrics include precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), and other standard measures used in ML/AI. These metrics are particularly relevant when assessing the usability of SD in real-world applications. On the other hand, intrinsic evaluation focuses on the inherent quality of SD itself, independent of any downstream task. These metrics assess characteristics such as distribution similarity, perplexity (e.g. in the case of SD for language models), and various statistical measures that capture how closely SD resembles RD. One approach to intrinsic evaluation involves confusion matrices, which can provide insight into how different SD generation methods relate to one another, helping quantify similarities or systematic biases. By combining both intrinsic and extrinsic evaluation methods, a more comprehensive understanding of SD quality can be achieved.

Evaluation metrics

Many different evaluation parameters were identified during the review. Here, we have attempted to segregate them into different categories according to their underlying properties.

- **Quantitative metrics:** These metrics are used for comparison of quantitative parameters between the SD and RD. These can be subclassified as follows:
  - Performance metrics: These are the standard statistical metrics that are generally used to test the performance of the data generation models. Examples include Sensitivity, F1-score, Precision, AUC-ROC score, AUC-PRC score, Confusion matrix.
  - Classification metrics: These metrics are used to benchmark the classification accuracy of the model. They consist of parameters like Matthews Correlation Coefficient and Cohen's Kappa.
  - Clustering metrics: The clustering metrics assess the quality of clusters formed from SD. Examples are Adjusted Rand Index and Normalized Mutual Information.
  - Ranking metrics: They are used to compare the rank of the data generated by the model in order of relevance. Examples are Spearman's rho and Kendall's tau.
  - Regression metrics: These metrics evaluate the predictive ability of a given model. Examples are Huber's Loss and Error rate.
- **Biologically informed/use case metrics:** These metrics assess the difference between the biological properties of the SD and RD. Examples include: read coverage, read count, read depth, and related read metrics like reads per kilobase million (RPKM) and fragments per kilobase million (FPKM), overlap in annotation/classification of single cells, fold change in gene/isoform expression, and clustering accuracy.
- **Qualitative metrics:** These metrics focus on the comparison of the aspects of data generating algorithms. Examples include: Speed, Complexity of datasets, Privacy, Diversity.

Synthetic proteomics

Domain overview

Proteomics falls under the biological study of proteins and is embedded in the wider Omics field. The breakdown of proteomics can be split into: quantitative proteomics, functional proteomics, structural proteomics, protein–protein interactions, qualitative methods with mass spectrometry (MS) imaging, proximity extension assays, and aptamer based platforms amongst others [24]. The instrumentation and technical methodologies also span multiple formats, with emerging single cell proteomics making a foray [25]. Comprehensive overview of the listed methodologies and their nuanced analytical approaches are described in [26, 27] for further reading. Data generated from Quantitative Mass Spectrometry (qMS) proteomics, includes spectral peaks, abundance values, intensity signals in MS, amounts of protein in samples, and many other measurement outputs depending on sub-field and instrumentation, which are of relevance when discussing SD in this domain.

Review outcomes

We identified 129 initial articles and 68 after duplicate removals, which went through an abstract screening phase, before the full text review ($n = 57$), and then data extraction ($n = 15$). The initial searches used combined queries per database (see Supplementary Material), Pubmed $n = 52$, Scopus $n = 35$, Google Scholar $n = 24$ (top 20 by relevance plus second query), Web of Science ($n = 16$), DOAJ.org ($n = 6$), total 129 papers. The reviewed papers did not have SD generation as the core focus, but rather SD was used as part of wider research questions, and evaluation metrics were extracted from multiple sections to infer any measures taken to evaluate the data.

Evaluation metrics

The reported metrics span the broad themes of quantitative, biological and qualitative sub categories. An example of sub category types would be statistical measures for quantitative metrics, such as false discovery rate (FDR) during identification and quantification stages of MS spectra. For qualitative metrics, this would include expert inspections, overlays, benchmarking, whilst biological modetrics would include similarity measures against biological benchmarks, closeness to true outputs from protein/peptide measures, and/or synthetic peptide standards.

The papers included reported evaluations of the SD produced, and further details around how they assessed these data in relation to overall study goals.

Types of SD generated or used in the included papers:

1. Synthetic mass spectra, spectral data, spectra, isobaric peptides [28–34]
2. Synthetic 2-dimensional gel electrophoresis (2DGE) spots, images [35, 36]
3. Synthetic expression data, synthetic features of proteomic data for designated models (high dimensional formats, signal intensities, perturbations, cut-offs) [37–40]
4. RPPA, P100 assays, and global chromatin profiling (GCP) [38]
5. Human Protein Atlas (HPA), large-scale immunohistochemistry (IHC) [41]
6. Networks (biological with proteins) [41, 42]

The main metrics and most represented area of proteomics from the papers was from MS based proteomics research. A detailed list of quality metrics identified for proteomics can be found in the Supplementary Material.

- **Quantitative metrics (statistical and performance):** Signal-to-noise ratio (SNR), spot detection efficiency, sensitivity and FDR, low-abundance protein detection, peak separation (e.g. normalized separation parameter), perturbation modelling (Gaussian, Poisson), threshold-based matching (cosine similarity, m/z tolerance), AUC and ROC curves, F1 score, MSPE and mean absolute error (MAE), feature matrix noise modelling, binomial and Markov chain Monte Carlo (MCMC) - based distributions, dataset balancing, and replicate consistency.
- **Qualitative metrics:** Perceptual image evaluation, visual distribution comparisons (e.g. RD versus SD), manual peptide inspection, replicate similarity, network connectivity patterns, and use of labeled ground truth for validation.
- **Biological and domain-specific metrics:** Protein abundance ranges, peptide length distributions, isotope pattern fidelity, expression heterogeneity (SNR), reduction of isobaric interference, representation of disease-associated proteins (e.g. known subsets, module overlaps), and molecular ion similarity.
- **Comparison metrics:** Prediction overlap percentage, number of recovered oscillating proteins, performance comparisons between RD and SD models, accuracy of identifications (e.g. peptide-spectrum match (PSM) counts), and graphical analysis under varying noise and perturbation scenarios.

Metrics like ROC, AUC, MSPE, and F1 were the most suitable metrics for assessing SD for model based validations, whilst noise and perturbations were used to mimic real time variations and heterogeneity, the LogFC thresholds thereby maintained relevancy to downstream analytics and classification. Ground truth recovery was identified as a way to ensure robustness in noisy network based paradigms, with grid searches allowing for many simulations improving reproducibility, whereas those employing manual inspection or subjective evaluation provide expert interaction with quantitative metrics. Metrics like pooling and identifying isobaric peptide signals as well as peptide length are considered contextually relevant items related to MS instrumentation.

Of the metric types recorded, in the context of proteomic data the comparison metrics would be deemed bespoke in that the model applications, which were designed to be validated themselves were used to show comparative metrics between RD and SD, with the main goal of highlighting the efficacy of the model as opposed to the SD generation. However, those which carefully considered instrumentation parameters and key features from RD could still provide a richer view of the SD quality metrics. The metrics used for algorithmic performance should ideally be comparable between SD and RD.

Some papers focused on SD tailored to specific model evaluations such as linear or correlation models without aiming to replicate RD. These SD were often designed for targeted validation tasks rather than broad mimicry of RD. However, reporting on how well the SD captured the intended characteristics was often insufficient. In cases where SD generation methods were adopted from prior work, descriptions of the process and evaluation were frequently minimal. Although the use of existing methods is common, clearly reporting the conditions and assumptions remains important. Furthermore, the review did not apply snowballing techniques, making it difficult to retrieve details when SD generation was only referenced indirectly.

Research papers using synthetic peptides for quantification required nuanced reviewer skills, the generation of synthetic peptides has existing and established quality control metrics, however papers which used synthetic spectra from synthetic peptide quantification for model evaluation purposes and provided evaluations either on how they then used MS to quantify alongside the RD or how they assessed these data in their own right were included for those metrics.

Many of the excluded papers had limited or unclear reporting of evaluation metrics which required further inference. However there was a critical evaluation paper on the use of artificial spectra data which is already successfully used in MS peptide identifications and considering these SD for training algorithms in *de novo* peptide identification [43]. Further expert commentary on the ML development in proteomics with relevance to SD and SD generation in the field can be found elsewhere [44, 45].

The range of papers included in the review represent a wide scope of the types of data and contexts available to proteomics and the extracted metrics for SD quality can provide an insightful index for future standardization.

Synthetic phenomics

Domain overview

Phenomics is a relatively new discipline of Biology that has been applied in several fields, mainly in crop sciences [46, 47]. Phenomics is the systematic study of the complete set of expressed phenotypes and the structure of such a set in relation to the genetic factors and environmental perturbations that caused their expression [48]. It combines concepts and methods from various disciplines to understand the complex interactions between genotype, phenotype, and environmental factors that influence organismal development, function, and evolution [49, 50]. While traditional approaches to phenomics rely heavily on experimental measurements [51], synthetic phenomics offers a complementary strategy that harnesses the power of computation to create simulated phenotypes.

Synthetic phenomics is a subfield within the broader scope of phenomics, focusing on the creation and utilization of SD to investigate complex traits. By leveraging state-of-the-art computational models and simulation techniques, we can now generate vast quantities of SD that closely resemble phenotypes from RD. This innovative approach has the potential to significantly enhance our understanding of complex traits, such as those involved in cancer, neurological disorders, and metabolic diseases [52]. Overall, phenomics has the potential to transform our understanding of complex diseases and traits, paving the way for the development of personalized therapies and targeted interventions.

Review outcomes

From the initial large number of publications [i.e. a total of 6920 identified from the queries on PubMed ($n = 2$), Web of Science ($n = 21$), Google Scholar ($n = 6890$), and ELIXIR curation registry ($n = 7$)] only 22 were finally selected for the review after the screening process. Before screening, we applied an extra filter to reduce the untreatable number of records obtained from Google Scholar to the top 20 ranked per relevance. This resulted in 50 records to be screened. We

excluded five synthetic phenomics datasets for not providing a peer-reviewed supporting publication. From the 45 retrieved reports, we excluded 3 publications for not being open access, and 20 for being false positives, i.e. they did neither generate nor use SD. We observed that 8 out of these 20 false positives were due to the 'synthetic biology' topic of the publication. Noteworthy, of the final 22 publications we reviewed, a significant number, the 91% (20/22), fall into the Plant Sciences domain, and the majority of publications, the 32% (7/22), were issued in 2021.

Evaluation metrics

Only three publications from the total reviewed publications, i.e. 14%, mentioned the use of evaluation metrics to assess the quality of the SD used or/and generated. The metrics identified were both quantitative and qualitative and overall dependent on the modality of the generated dataset:

1. **Quantitative metrics** [53]: widths and heights in pixels of real versus synthetic maize tassels images; structural similarity index computation (SSIM) to indicate the similarity between two images, real and synthetic; usability of SD by accuracy score of a classifier model.
2. **Perceptual qualitative metrics** [53]: human perception annotation about real or synthetic image; human rating on the similarity of a pair real-synthetic image.
3. **Count distributions similarity** [54]: distributions of the count table elements.
4. **Distributions of information measures** [54]: joint entropies and any information theoretic measure which is a function of the entropies such as multi-information Ω.
5. **Statistical testing of distribution equivalence** [54]: quantile-quantile plot of *P*-values from Epps–Singleton tests comparing synthetic versus expected distributions.
6. **Realism of the simulated dataset** [55]: Euclidean distance.
7. **Image spatial resolution** [55]: spatial resolution per pixel.

Synthetic imaging

Domain overview

Synthetic imaging is a rapidly evolving field in medical imaging, leveraging advanced computational techniques to generate realistic medical images. These synthetic images are pivotal in various applications, including training ML/AI models, enhancing diagnostic accuracy, and facilitating the development of new imaging technologies. Synthetic imaging presents numerous advantages and diverse applications in the medical field [56]. One key benefit is data augmentation, where synthetic images enrich existing datasets by providing a wider variety of samples, effectively addressing the limitations of RD. This approach also enhances privacy protection, as SD eliminates the need to use sensitive patient information, while simultaneously reducing the costs and effort associated with annotating real medical images. Furthermore, synthetic imaging serves as a valuable tool for training and educating healthcare professionals, offering a risk-free environment that safeguards patient confidentiality. Beyond education, it opens up new avenues for research, including modality translation and the development of AI-driven diagnostic tools.

Review outcomes

A comprehensive review of synthetic imaging literature from the past decade was conducted, which identified and screened a large number of publications. The initial search outcomes were: PubMed identified 13 articles, with 10 fitting the criteria after applying the last 10 years filter; SCOPUS identified 21 articles, with 18 fitting the criteria after filtering; and Google Scholar identified 307 articles, with 296 fitting the criteria after applying the last 10 years filter.

After examining 341 articles, 45 articles were finally identified as an appropriate fit for our effort. These selected articles fall under two main categories: those reviewing tools for simulating synthetic images and those suggesting guidelines for evaluation. The initial search yielded a large number of publications with only a few that were relevant, mostly due to the broad usage of the term "synthetic", prompting the need for more precise search terms.

Evaluation metrics

Evaluating the quality and performance of synthetic images involves both quantitative and qualitative metrics.

Quantitative metrics as identified in the curated publications [57–93] include:

- **Peak signal-to-noise ratio (PSNR)**: Measures the ratio between the maximum possible power of a signal and the power of corrupting noise, indicating image quality.
- **Structural similarity index (SSIM)**: Assesses the similarity between two images, focusing on luminance, contrast, and structure.
- **Mean squared error (MSE)**: Calculates the average squared differences between the generated and reference images.
- **Mean absolute error (MAE)**: Measures the average absolute differences between the pixel values of the synthetic and real images.
- **Fréchet inception distance (FID)**: Evaluates the distance between the distributions of real and synthetic images in a feature space, providing an overall measure of image quality.

Use case related metrics: These metrics are tailored for specific applications:

- **Visual Inspection:** Expert radiologists visually assess image quality, anatomical detail, and clinical utility.
- **Radiologist evaluation:** Radiologists use rating scales to assess image sharpness, contrast, and artifact presence.

Statistics based metrics: These metrics are derived from statistical analyses:

- **Wilcoxon signed-rank test:** A nonparametric test to compare paired samples, like synthetic and real images.
- **Friedman test:** A nonparametric test to detect differences in treatments across multiple test attempts.
- **Principal Component Analysis (PCA):** Used for qualitative data visualization.
- **Supervised discriminators:** Employed for quantitative evaluation, measuring how well synthetic images can be distinguished from real ones.

Synthetic EHRs

Domain overview

EHRs serve as an extensive, lifelong repository of patient health information, encompassing diverse data types like medical history, medications, allergies, lab results and more. These records, often dispersed across multiple platforms and locations, provide a detailed, chronological overview of a patient's medical journey and facilitate data sharing across healthcare settings. However, the fragmented nature of EHR systems complicates access to patient records, creating challenges for clinical practitioners, researchers, and developers [94].

Review outcomes

After curating 83 articles, 18 of them were finally selected. Regarding the generation methods the main sources mainly involve types of GANs, probabilistic models, classification models, weighted bayesian association rule mining, Gaussian multivariate, bayesian networks and sequential tree synthesis. As indicated by [94] and further supported by other works [3, 95–97], different evaluation metrics are often utilized, usually based on the type of generation method that was initially implemented to produce SD, which makes the direct comparison of the results quite challenging. On the other hand, the work of [98] distinguishes the metrics used based on data type focusing on tabular data and time series.

Evaluation metrics

A plethora of metrics are utilized for the evaluation of EHR SD. It is important to clarify that this literature is primarily targeting the structured part of the EHRs, and therefore the textual form is not taken into consideration (nor any AI-drive aspects). As such, the SD in this case are tabular data and sometimes longitudinal data. The majority of the publications that were selected indicated three main categories of evaluation metrics regarding the fidelity of the SD, the utility and finally the preservation of privacy. Furthermore, many studies note the distinction of the latter category into further information disclosure related subcategories. The first subcategory includes those regarding the identity disclosure which occurs when an attacker can accurately determine that a specific individual is part of the training dataset. This risk arises when one or more SD points closely resemble a RD sample used in generating the synthetic dataset, potentially revealing its inclusion in the original confidential data. The second subcategory includes those that deal with the attribute disclosure which refers to the risk that an attacker can accurately predict the original values of sensitive attributes for an individual in the confidential dataset. This occurs when the attacker uses known variables from the RD, combined with patterns in the SD, to infer the sensitive values. The likelihood of such disclosure depends on factors like the number of known variables, the size of the synthetic dataset, and the configuration of the attack model.

The metrics that were identified after carefully curating the papers that were included in the manuscript are categorized and presented below:

- **Fidelity evaluation**: variable distributions, frequency of data features, dimensional probability or probability distributions, Consultation with clinical experts (qualitatively), Data similarity, Synthetic to Synthetic (STS), Real to Synthetic (RTS) and Real to Real (RTR), Pairwise Pearson correlation coefficient, most common values, Statistical similarity (mean, std, miss rate), Visualization techniques (PCA, histograms, and correlation matrices), Nearest neighbor adversarial accuracy (AA) and fidelity loss, average trends, Train an ML classifier to label data as real or synthetic, distribution of attributes, Kolmogorov–Smirnov (K–S), Welsch *t*-tests, Student *t*-test, Chi Square tests, Dimension-wise distribution similarity [Dimension-wise Probability, Dimension-wise Average, K–S test, Support Coverage, Kullback–Leibler divergence (KLD)] [99–102].
- **Utility evaluation**: Augment data for ML model training, Use SD in ML models, KLD, pairwise correlation difference, log-cluster, support coverage, cross-classification, Multivariate Hellinger distance, Wasserstein Distance, Cluster analysis measure, Propensity and prediction MSE, Dimension-wise distribution, Column-wise correlation, Latent cluster analysis, Clinical knowledge violation, Medical concept abundance, Feature selection [103, 104].
- **Privacy evaluation**: Identity disclosure and attribute disclosure, Distance to the closest record, Membership attack, Max-RTS similarity, Formulation of DP, Privacy loss, Maximum mean discrepancy, JS divergence and Wasserstein distance, Differential privacy cost, Distance to the optimal point, Attribute inference risk, Membership inference risk, Meaningful identity disclosure risk, Nearest neighbor adversarial accuracy risk [94, 101, 103–105].
- **Extra metrics**: ML performance metrics (Sensitivity, Specificity, Positive Predictive Value/Precision, Negative Predictive Value, AUC score, Precision–Recall AUC score), Rule similarity (for specific data type generation) [105–107].

Discussion

Our review of evaluation metrics for SD across six distinct life sciences domains (genomics, transcriptomics, proteomics, phenomics, imaging, and EHRs) highlights both the diversity and domain-specific nature of the metrics commonly used in the field. While SD is playing an increasingly important role in scientific research, evaluating its quality and applicability remains a significant challenge, particularly when comparing across these diverse domains. Although SD is tailored to specific applications within each domain, the lack of standards makes it difficult to assess and compare SD quality even within the same application domain. This challenge becomes even more pronounced with the growing interest in multimodal applications of SD, where multiple data types from different domains are artificially generated and merged. This trend highlights the opportunity to develop new evaluation metrics tailored to these complex use cases, which will require a separate review as applications begin to develop further.

In total, after the systematic screening of over 8000 initial records retrieved from multiple databases 188 publications were finally curated across all SD domains. The refinement process involved multiple layers of filtering, including duplicate removal, exclusion of preprints and nonpeer-reviewed material, non-English texts, articles older than 10 years and records that misused or ambiguously applied the term "synthetic". A total of 156 metrics were identified (Fig. 2), of which 142 are unique to a single domain and 14 are shared across multiple domains. Notably, domains with a larger num-

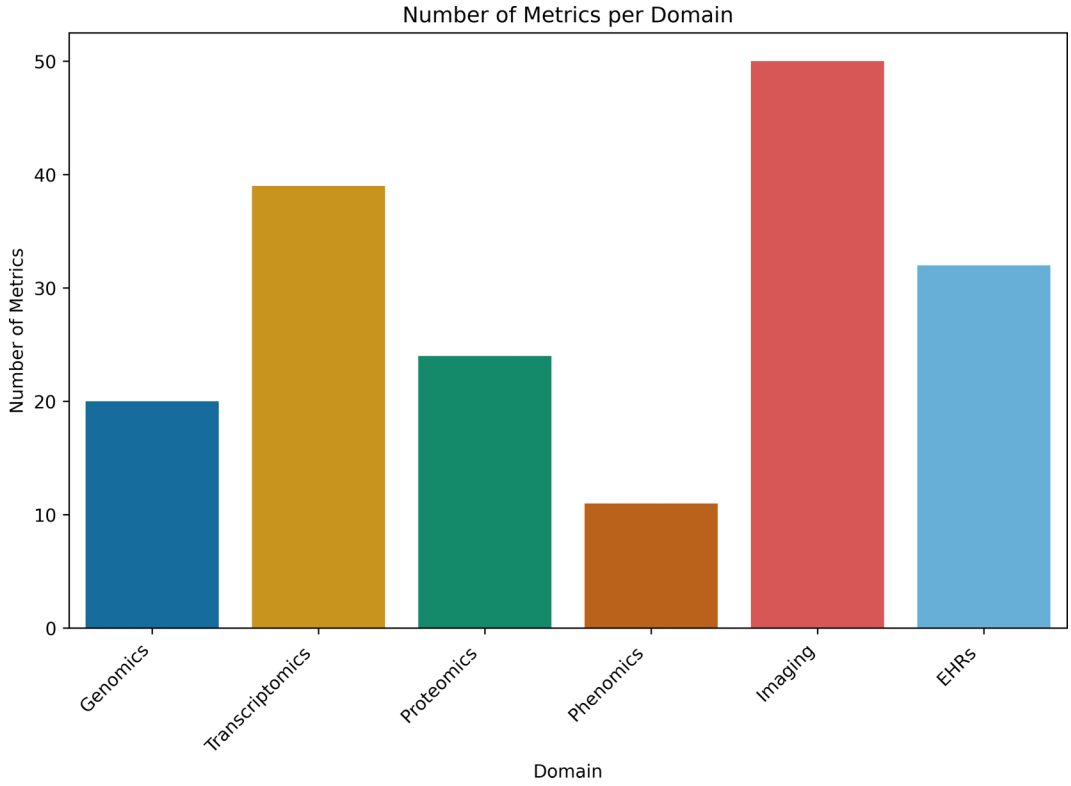

Figure 2. Barplot of number of SD evaluation metrics per domain. Legend: Barplot showing the number of SD evaluation metrics that were retrieved for each domain in the systematic review. The figure highlights the differences in how many metrics have been reported across genomics, transcriptomics, proteomics, phenomics, imaging, and EHRs. Alt text: A bar chart with six bars, each representing a life sciences domain, showing how many SD evaluation metrics were found in each domain.

ber of reported metrics (such as EHRs) do not necessarily reflect greater methodological maturity; instead, they may indicate fragmentation and the absence of widely adopted evaluation standards. This is an interesting observation that emerged from our curation effort that needs further investigation in future works. These shared metrics are illustrated in Fig. 3 and explored in greater detail in Supplementary Fig. S8.

In the domain of **genomics** the review found that the term "synthetic" has often a different meaning, with many studies referring to wet-lab synthesized DNA rather than computationally generated sequences. This issue underscores the need for clearer terminology and domain-specific criteria for evaluating SD in genomics. Regarding the curation effort, the outlined metrics can be put into three categories: bioinformatics-related metrics, use case-related metrics and finally statistics-based metrics. Similarly, in **transcriptomics**, SD plays a crucial role in gene expression analysis and ML/AI applications. The primary focus for the application of SD in this domain include data augmentation and balancing underrepresented classes to increase statistical power for the downstream analyses. Extrinsic metrics assess model performance when trained on SD versus RD, using measures like precision, recall, F1-score and AUC-ROC, which are crucial for real-world applicability. Intrinsic metrics, on the other hand, evaluate the inherent quality of SD, focusing on distribution similarity, perplexity and statistical comparisons. While extrinsic metrics ensure that SD is suitable for practical applications, intrinsic metrics provide fundamental insights into data fidelity, consistency and variability. By combining both approaches, a more comprehensive and robust SD assessment can be achieved, contributing to more reliable benchmarking frameworks.

In **proteomics** research, SD is frequently used, yet its evaluation often lacks standardized metrics. While many reviewed studies describe SD generation methods, few detail how the data is assessed. However, emerging metrics, such as ROC, AUC, peak accuracy, sensitivity, specificity, and image fidelity, suggest the potential for developing best practices across subfields like 2DGE and MS. These metrics are highly context-dependent, varying with the methodology employed. Clear ontologies and evaluation frameworks are needed to guide SD use across proteomics workflows, which involve multiple stages from sample preparation to data interpretation. For quantitative MS, this includes considering different acquisition methods (e.g. DDA, DIA), labelling strategies, and spectral features like noise and fragmentation. Notably, this review excluded synthetic peptide production, which follows separate standards. Ultimately, SD in proteomics aims to support, not replace, RD, enabling more efficient hypothesis testing and model development ahead of experimental validation.

In **phenomics**, the reviewing process found that while SD offers a promising approach for studying genotype-phenotype interactions, the primary focus remains on data generation rather than systematic assessment, indicating a gap that needs to be addressed to ensure the reliability of SD for this domain. Furthermore, the dominance of plant-based studies suggests a potential bias in SD research, requiring further exploration in human-related phenomics applications. The metrics here can be divided into various categories, including qualitative, per-

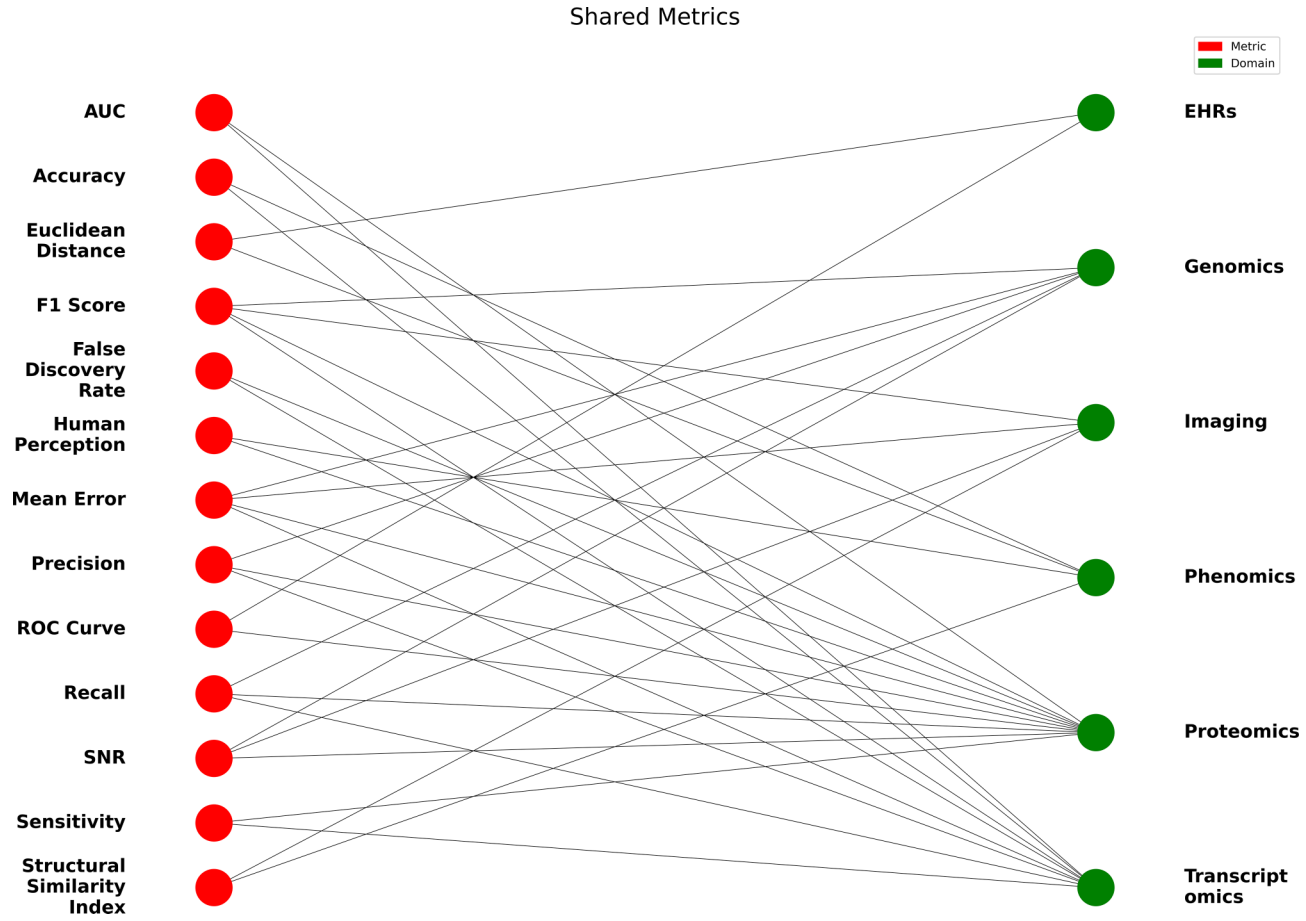

Figure 3. Bipartite graph of shared evaluation metrics across domains. Legend: Bipartite graph displaying evaluation metrics (nodes on the left) that are shared across multiple domains (nodes on the right). This visualization shows the connections between metrics and the domains where they are applied, revealing overlap among life sciences areas in SD evaluation practices. Note: Recall and sensitivity are equivalent metrics and are both mentioned due to their common usage in the literature. AUC denotes the area under the curve, SNR the signal-to-noise ratio, EHRs electronic health records and ROC the receiver operating characteristic curve. Alt text: A bipartite network diagram showing which evaluation metrics are shared between different life sciences domains.

ceptual qualitative, distribution similarity, information measure, statistical testing of distribution equivalence, realism of SD and image spatial resolution.

Medical **Imaging** is one of the more mature fields in terms of SD applications, as synthetic medical images are widely used for training AI models and augmenting RD. However, despite the significant number of studies reviewed, most publications focused on the development of synthetic imaging tools rather than comprehensive evaluations of their reliability. Metrics like structural similarity and image fidelity were commonly reported, but there was a lack of consistency in benchmarking methodologies. The metrics that were identified as relevant in this domain can be split into quantitative, use case-related and statistics-based metrics.

Lastly, in the domain of **EHRs**, SD is used to generate privacy-preserving patient records for research and AI model training. The review found a wide variety of generation methods, including GANs, probabilistic models and Bayesian networks, each employing different evaluation metrics. The lack of uniformity in assessment strategies makes it difficult to establish clear guidelines for evaluating synthetic EHR data, further emphasizing the need for a standardized approach. At this point it should be noted that both Medical Imaging and EHR SD applications are particularly relevant in a clinical context, and the respective need for more rigorous and standardized evaluation ways to assess the clinical utility of the SD in these fields. Importantly, clinical utility is intrinsically use-case dependent. What constitutes "useful" SD varies according to whether the downstream task involves risk prediction, association testing, subgroup comparison or model development. Evaluating utility therefore requires examining whether analyses performed on SD lead to the same clinical conclusions as those produced by RD. Regarding synthetic EHRs, the work of [108] provides a clear example of this by comparing confidence-interval overlap in estimates across real and synthetic datasets, demonstrating how closely clinical inference aligns when predicting diabetes onset. Similarly, the work of [109] extends utility assessment beyond statistics by incorporating clinical judgment. Clinicians evaluated whether learned relationships were medically plausible and a blinded experiment asked them to distinguish real from synthetic patient records. Their performance offers a practical benchmark of clinical credibility and consistency of analytical outcomes, highlighting the importance of evolving domain experts in these evaluations.

A key challenge across domains is the inconsistency in evaluation criteria, which often depends on the SD generation method. Some studies emphasize statistical similarity

to RD, while others focus on downstream utility, like improving ML/AI performance. Additional approaches assess data integrity, reproducibility, or real-world applicability. This diversity highlights the need for domain-specific yet practical evaluation guidelines that ensure consistent, rigorous, and meaningful assessments tailored to each domain's unique needs.

Furthermore, as members of ELIXIR Europe [110], we advocate for interoperability in SD evaluation metrics that align with existing standards and related initiatives. This approach is consistent with the software best practices defined under the ELIXIR-STEERS project [111] and the software quality indicators developed within the European Open Science Cloud (EOSC) ecosystem [112]. For example, the EVERSE project [113] aims to establish a framework for research software excellence by integrating community curation, quality assessment, and best practices, which will be encapsulated in the Research Software Quality toolkit (RSQkit) as a comprehensive knowledge base for high-quality software development across disciplines. The interoperability of SD evaluation metrics could be achieved by developing dedicated ontologies and semantic strategies to represent the evaluation metrics to allow their integration in SD generation benchmarking platforms, such those currently developed in European projects like SYNTHIA [114] and SYNTHEMA [115].

Despite advancements in medical AI research, the clinical adoption of emerging AI solutions remains challenging due to limited trust and ethical concerns. Several recent initiatives are contributing to the development of robust best practices and guidelines for the evaluation of AI, such as FUTURE-AI consortium [116] and the AHEAD project [117], addressing ethical concerns, privacy implications, fairness, transparency, and the broader societal impact.

We believe that by integrating both technical and societal evaluation approaches, more robust and comprehensive recommendations for assessing SD can be established. Such an integrated approach would not only promote responsible AI innovation but also ensure that the deployment of SD across various domains is both reliable and aligned with broader societal values.

Finally, this foundational mapping also lays the groundwork for implementing the metrics collected in this work in a FAIR and actionable evaluation framework for the ELIXIR community and beyond, accompanied by practical guidelines for assessing SD across biomedical domains.

Conclusions

This ELIXIR scoping review highlights the urgent need for standardized evaluation metrics to assess the quality, reliability, and applicability of SD across scientific domains. The review found that most current research emphasizes SD generation rather than its evaluation, creating an imbalance that may hinder the credibility and adoption of SD in scientific and clinical settings. To address this, future work should focus on developing robust, shared evaluation frameworks to ensure consistency, comparability, and trust in SD use. Establishing clear assessment guidelines will support better validation of SD methods and promote responsible integration into research and healthcare. The review's findings lay the groundwork for initiatives aimed at improving SD evaluation practices, ensuring synthetic datasets are effectively and ethically applied across disciplines.

Acknowledgements

The authors would like to acknowledge The ELIXIR Machine Learning Focus Group for helpful discussions. Also, the authors would like to express their gratitude to Magnus Palmblad, Rahuman Sheriff, Sara Morsy, Ruben Branco, Tim Beck, Salvador Capella-Gutiérrez, and Sucheta Ghosh for their valuable support and insightful advice.

Author contributions: D.C. and N.Q.R. conceived and designed the review. S.C.F., S.I., B.G., N.R., D.H., A.T., L.C., and N.Q.R. conducted the literature search, analyzed and interpreted the findings, and contributed to manuscript writing. A.V., T.L., F.P., and P.O.B. supervised the work. All authors reviewed and approved the final manuscript.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by ELIXIR, the research infrastructure for life-science data. The Centre for Research & Technology Hellas (CERTH), Barcelona Supercomputing Center (BSC), and Leiden University Medical Center (LUMC) receive support from SYNTHIA. SYNTHIA (Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications) is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement no. 101172872. The JU receives support from the European Union's Horizon Europe research and innovation programme, COCIR, EFPIA, Europa Bío, MedTech Europe, Vaccines Europe, and DNV. The UK consortium partner, The National Institute for Health and Care Excellence (NICE) is supported by UKRI Grant 10132181. Funded by the European Union, the private members, and those contributing partners of the IHI JU. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

Data availability

No new data were generated or analysed in support of this research.

References

1. El Emam K, Mosquera L, Fang X *et al.* An evaluation of the replicability of analyses using synthetic health data. *Sci Rep* 2024;14:6978. https://doi.org/10.1038/s41598-024-57207-7
2. Pantanowitz J, Manko CD, Pantanowitz L *et al.* Synthetic data and its utility in pathology and laboratory medicine. *Lab Invest* 2024;104:102095. https://doi.org/10.1016/j.labinv.2024.102095
3. Pezoulas VC, Zaridis DI, Mylona E *et al.* Synthetic data generation methods in healthcare: a review on open-source tools and methods. *Comput Struct Biotechnol J* 2024;23:2892–910. https://doi.org/10.1016/j.csbj.2024.07.005

4. Susser D, Schiff DS, Gerke S *et al.* Synthetic Health Data: real ethical promise and peril. *Hastings Cent Rep* 2024;54:8–13. https://doi.org/10.1002/hast.4911
5. Figueira A, Vaz B. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics* 2022;10:2733. https://doi.org/10.3390/math10152733
6. Queralt-Rosinach N, Alshaikhdeeb B, Bolzani L *et al.* Infrastructure for synthetic health data. 2023. https://doi.org/10.37044/osf.io/q4zgx
7. Kaabachi B, Despraz J, Meurers T *et al.* A scoping review of privacy and utility metrics in medical synthetic data. *NPJ Digit Med* 2025;8:60. https://doi.org/10.1038/s41746-024-01359-3
8. Santangelo G, Nicora G, Bellazzi R *et al.* How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data. *BMC Med Inform Decis Mak* 2025;25:89. https://doi.org/10.1186/s12911-024-02731-9
9. Rujas M, Martín Gómez Del Moral Herranz R, Fico G *et al.* Synthetic data generation in healthcare: a scoping review of reviews on domains, motivations, and future applications. *Int J Med Informatics* 2025;195:105763. https://doi.org/10.1016/j.ijmedinf.2024.105763
10. Page MJ, McKenzie JE, Bossuyt PM *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. https://doi.org/10.1136/bmj.n71
11. Beel J, Gipp B. Google Scholar's ranking algorithm: the impact of citation counts (an empirical study). In: *Third International Conference on Research Challenges in Information Science, 2009*. Fez, Morocco: IEEE, 2009. https://doi.org/10.1109/rcis.2009.5089308
12. Farahani AF, Kasraei N. Evaluating the impact of artificial intelligence on vaccine development: lessons learned from the COVID-19 pandemic. medRxiv. 2024; https://doi.org/10.1101/2024.10.23.24315991
13. Nguyen E, Poli M, Durrant MG *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* 2024;386:eado9336. https://doi.org/10.1126/science.ado9336
14. Chen H-S, Hutter CM, Mechanic LE *et al.* Genetic simulation tools for post-genome wide association studies of complex diseases. *Genet Epidemiol* 2015;39:11–9. https://doi.org/10.1002/gepi.21870
15. Milhaven M, Pfeifer SP. Performance evaluation of six popular short-read simulators. *Heredity* 2023;130:55–63. https://doi.org/10.1038/s41437-022-00577-3
16. Alosaimi S, Bandiang A, van Biljon N *et al.* A broad survey of DNA sequence data simulation tools. *Brief Funct Genomics* 2020;19:49–59. https://doi.org/10.1093/bfgp/elz033
17. Donato L, Scimone C, Rinaldi C *et al.* New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies. *Neural Comput Applic* 2021;33:15669–92. https://doi.org/10.1007/s00521-021-06188-z
18. Perera M, Montserrat DM, Barrabes M *et al.* Generative moment matching networks for genotype simulation. *Annu Int Conf IEEE Eng Med Biol Soc* 2022;2022:1379–83. https://doi.org/10.1109/EMBC48229.2022.9871045
19. Yang S, Guo L, Shao F *et al.* A systematic evaluation of feature selection and classification algorithms using simulated and real miRNA sequencing data. *Comput Math Methods Med* 2015;2015:1. https://doi.org/10.1155/2015/178572
20. Shakola F, Palejev D, Ivanov I. A framework for comparison and assessment of synthetic RNA-seq data. *Genes* 2022;13:2362. https://doi.org/10.3390/genes13122362
21. Thind AS, Monga I, Thakur PK *et al.* Demystifying emerging bulk RNA-seq applications: the application and utility of bioinformatic methodology. *Brief Bioinform* 2021;22:bbab259. https://doi.org/10.1093/bib/bbab259
22. Bej S, Galow A-M, David R *et al.* Automated annotation of rare-cell types from single-cell RNA-sequencing data through synthetic oversampling. *BMC Bioinformatics* 2021;22:557. https://doi.org/10.1186/s12859-021-04469-x
23. Treppner M, Salas-Bastos A, Hess M *et al.* Synthetic single cell RNA sequencing data from small pilot studies using deep generative models. *Sci Rep* 2021;11:9403. https://doi.org/10.1038/s41598-021-88875-4
24. Cui M, Cheng C, Zhang L. High-throughput proteomics: a methodological mini-review. *Lab Invest* 2022;102:1170–81. https://doi.org/10.1038/s41374-022-00830-7
25. Ctortecka C, Mechtler K. The rise of single-cell proteomics. *Anal Sci Adv* 2021;2:84–94. https://doi.org/10.1002/ansa.202000152
26. Graves PR, Haystead TAJ. Molecular biologist's guide to proteomics. *Microbiol Mol Biol Rev* 2002;66:39–63. https://doi.org/10.1128/MMBR.66.1.39-63.2002
27. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;422:198–207. https://doi.org/10.1038/nature01511
28. Bob K, Teschner D, Kemmer T *et al.* Locality-sensitive hashing enables efficient and scalable signal classification in high-throughput mass spectrometry raw data. *BMC Bioinformatics* 2022;23:287. https://doi.org/10.1186/s12859-022-04833-5
29. Lee C-Y, Wang D, Wilhelm M *et al.* Mining the human tissue proteome for protein citrullination. *Mol Cell Proteomics* 2018;17:1378–91. https://doi.org/10.1074/mcp.RA118.000696
30. Ovando-Vázquez C, Cázarez-García D, Winkler R. Target–decoy MineR for determining the biological relevance of variables in noisy datasets. *Bioinformatics* 2021;37:3595–603. https://doi.org/10.1093/bioinformatics/btab369
31. Ruffieux H, Carayol J, Popescu R *et al.* A fully joint Bayesian quantitative trait locus mapping of human protein abundance in plasma. *PLoS Comput Biol* 2020;16:e1007882. https://doi.org/10.1371/journal.pcbi.1007882
32. Xu J, Wang L, Li J. Biological network module-based model for the analysis of differential expression in shotgun proteomics. *J Proteome Res* 2014;13:5743–50. https://doi.org/10.1021/pr5007203
33. Marczyk M, Polanska J, Polanski A. Improving peak detection by Gaussian mixture modeling of mass spectral signalIn. In: *Third International Conference on Frontiers of Signal Processing (ICFSP), 2017*. Paris, France: IEEE, 2017. https://doi.org/10.1109/icfsp.2017.8097057
34. Cubison MJ, Jimenez JL. Statistical precision of the intensities retrieved from constrained fitting of overlapping peaks in high-resolution mass spectra. *Atmos Meas Tech* 2015;8:2333–45. https://doi.org/10.5194/amt-8-2333-2015
35. Matuzevičius D. Synthetic data generation for the development of 2D gel electrophoresis protein spot models. *Appl Sci* 2022;12:4393. https://doi.org/10.3390/app12094393
36. Goez MM, Torres-Madronero MC, Rothlisberger S *et al.* Joint pre-processing framework for two-dimensional gel electrophoresis images based on nonlinear filtering, background correction and normalization techniques. *BMC Bioinformatics* 2020;21:376. https://doi.org/10.1186/s12859-020-03713-0
37. Du L, Zhang J, Liu F *et al.* Identifying associations among genomic, proteomic and imaging biomarkers via adaptive sparse multi-view canonical correlation analysis. *Med Image Anal* 2021;70:102003. https://doi.org/10.1016/j.media.2021.102003
38. Dhruba SR, Rahman A, Rahman R *et al.* Recursive model for dose-time responses in pharmacological studies. *BMC Bioinformatics* 2019;20:317. https://doi.org/10.1186/s12859-019-2831-4
39. De Los Santos H, Bennett KP, Hurley JM. MOSAIC: a joint modeling methodology for combined circadian and non-circadian analysis of multi-omics data. *Bioinformatics* 2021;37:767–74. https://doi.org/10.1093/bioinformatics/btaa877
40. Mansouri M, Khakabimamaghani S, Chindelevitch L *et al.* Aristotle: stratified causal discovery for omics data. *BMC Bioinformatics* 2022;23:42. https://doi.org/10.1186/s12859-021-04521-w

41. Kavran AJ, Clauset A. Denoising large-scale biological data using network filters. *BMC Bioinformatics* 2021;22:157. https://doi.org/10.1186/s12859-021-04075-x
42. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015;11:e1004120. https://doi.org/10.1371/journal.pcbi.1004120
43. McDonnell K, Howley E, Abram F. Critical evaluation of the use of artificial data for machine learning based peptide identification. *Comput Struct Biotechnol J* 2023;21:2732–43. https://doi.org/10.1016/j.csbj.2023.04.014
44. Neely BA, Dorfer V, Martens L *et al.* Toward an integrated machine learning model of a proteomics experiment. *J Proteome Res* 2023;22:681–96. https://doi.org/10.1021/acs.jproteome.2c00711
45. Leeming MG, Ang C-S, Nie S *et al.* Simulation of mass spectrometry-based proteomics data with Synthedia. *Bioinform Adv* 2023;3: vbac096. https://doi.org/10.1093/bioadv/vbac096
46. Micol JL. Leaf development: time to turn over a new leaf? *Curr Opin Plant Biol* 2009;12:9–16. https://doi.org/10.1016/j.pbi.2008.11.001
47. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. *Nat Rev Genet* 2010;11:855–66. https://doi.org/10.1038/nrg2897
48. Zavafer A, Bates H, Mancilla C *et al.* Phenomics: conceptualization and importance for plant physiology. *Trends Plant Sci* 2023;28:1004–13. https://doi.org/10.1016/j.tplants.2023.03.023
49. Tolani P, Gupta S, Yadav K *et al.* Big data, integrative omics and network biology. *Adv Protein Chem Struct Biol* 2021;127:127–60. https://doi.org/10.1016/bs.apcsb.2021.03.006
50. Singh N, Ujinwal M, Singh A. Advances in agricultural bioinformatics: an outlook of multi "omics" approaches. In: *Bioinformatics in Agriculture*. Elsevier, 2022. p. 3–21. https://doi.org/10.1016/b978-0-323-89778-5.00001-5
51. Yi L. LiDAR: an important tool for next-generation phenotyping technology of high potential for plant phenomics? *Comput Electron Agric* 2015;119:61–73. https://doi.org/10.1016/j.compag.2015.10.011
52. Murtaza H, Ahmed M, Khan NF *et al.* Synthetic data generation: state of the art in health care domain. *Comput Sci Rev* 2023;48:100546. https://doi.org/10.1016/j.cosrev.2023.100546
53. Shete S, Srinivasan S, Gonsalves TA. TasselGAN: an application of the generative adversarial model for creating field-based maize tassel data. *Plant Phenomics* 2020; 2020, 2020:8309605. https://doi.org/10.34133/2020/8309605
54. Kunert-Graf JM, Sakhanenko NA, Galas DJ. Optimized permutation testing for information theoretic measures of multi-gene interactions. *BMC Bioinformatics* 2021;22:180. https://doi.org/10.1186/s12859-021-04107-6
55. Li Y, Zhan X, Liu S *et al.* Self-supervised plant phenotyping by combining domain adaptation with 3D plant model simulations: application to wheat leaf counting at seedling stage. *Plant Phenomics* 2023;5:0041. https://doi.org/10.34133/plantphenomics.0041
56. Koetzier LR, Wu J, Mastrodicasa D *et al.* Generating synthetic data for medical imaging. *Radiology* 2024; 312:, 312:e232471. https://doi.org/10.1148/radiol.232471
57. Liu Y, Niu H, Ren P *et al.* Generation of quantification maps and weighted images from synthetic magnetic resonance imaging using deep learning network. *Phys Med Biol* 2022;67:25002. https://doi.org/10.1088/1361-6560/ac46dd
58. Galbusera F, Bassani T, Casaroli G *et al.* Generative models: an upcoming innovation in musculoskeletal radiology? A preliminary test in spine imaging. *Eur Radiol Exp* 2018;2:29. https://doi.org/10.1186/s41747-018-0060-7
59. Biradar N, Dewal ML, Rohit MK. Edge preserved speckle noise reduction using integrated fuzzy filters. *Int Sch Res Notices* 2014;2014:1. https://doi.org/10.1155/2014/876434
60. Colleoni E, Psychogyios D, Van Amsterdam B *et al.* : Simulation-supervised image synthesis for surgical instrument segmentation. *IEEE Trans Med Imaging* 2022;41:3074–86. https://doi.org/10.1109/TMI.2022.3178549
61. Nykänen O, Nevalainen M, Casula V *et al.* Deep-learning-based contrast synthesis from MRF parameter maps in the knee joint. *Magn Reson Imaging* 2023;58:559–68. https://doi.org/10.1002/jmri.28573
62. Hu S, Lei B, Wang S *et al.* Bidirectional mapping generative adversarial networks for brain MR to PET synthesis. *IEEE Trans Med Imaging* 2022;41:145–57. https://doi.org/10.1109/TMI.2021.3107013
63. Qin Z, Liu Z, Zhu P *et al.* A GAN-based image synthesis method for skin lesion classification. *Comput Methods Programs Biomed* 2020;195:105568. https://doi.org/10.1016/j.cmpb.2020.105568
64. Wang Z, Lin Y, Cheng K-TT *et al.* Semi-supervised mp-MRI data synthesis with StitchLayer and auxiliary distance maximization. *Med Image Anal* 2020;59:101565. https://doi.org/10.1016/j.media.2019.101565
65. Zhao T, Hoffman J, McNitt-Gray M *et al.* Ultra-low-dose CT image denoising using modified BM3D scheme tailored to data statistics. *Med Phys* 2019;46:190–8. https://doi.org/10.1002/mp.13252
66. Gourdeau D, Duchesne S, Archambault L. On the proper use of structural similarity for the robust evaluation of medical image synthesis models. *Med Phys* 2022;49:2462–74. https://doi.org/10.1002/mp.15514
67. Shah D, Suresh A, Admasu A *et al.* A survey on evaluation metrics for synthetic material micro-structure images from generative models. arXiv, https://doi.org/10.48550/ARXIV.2211.09727, 03 November 2022, preprint: not peer reviewed.
68. Mahmood F, Chen R, Durr NJ. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Trans Med Imaging* 2018;37:2572–81. https://doi.org/10.1109/TMI.2018.2842767
69. Korol FJ, Stevens T. Gan quality metrics for evaluating computed tomography image generators. *SSRN J* 2022. https://doi.org/10.2139/ssrn.4049605
70. Korkinof D, Rijken T, O'Neill M *et al.* High-resolution mammogram synthesis using progressive generative adversarial networks. arXiv, https://doi.org/10.48550/ARXIV.1807.03401, 03 November 2019, preprint: not peer reviewed.
71. Aljohani A, Alharbe N. Generating synthetic images for healthcare with novel deep Pix2Pix GAN. *Electronics* 2022;11:3470. https://doi.org/10.3390/electronics11213470
72. Sivanesan U, Braga LH, Sonnadara RR *et al.* Unsupervised medical image segmentation with adversarial networks: from edge diagrams to segmentation maps. arXiv, https://doi.org/10.48550/ARXIV.1911.05140, 12 November 2019, preprint: not peer reviewed.
73. Mahmood F, Chen R, Sudarsky S *et al.* Deep learning with cinematic rendering: fine-tuning deep neural networks using photorealistic medical images. *Phys Med Biol* 2018;63:185012. https://doi.org/10.1088/1361-6560/aada93
74. O'Reilly JA, Asadi F. Identifying obviously artificial medical images produced by a generative adversarial network. *Annu Int Conf IEEE Eng Med Biol Soc* 2022;2022:430–3. https://doi.org/10.1109/EMBC48229.2022.9871217
75. Asadi F, O'Reilly JA. Artificial computed tomography images with progressively growing generative adversarial network. *13th Biomedical Engineering International Conference (BMEiCON), 2021*. Ayutthaya, Thailand: IEEE, 2021. https://doi.org/10.1109/bmeicon53485.2021.9745251
76. Li HB, Prabhakar C, Shit S *et al.* A domain-specific perceptual metric via contrastive self-supervised representation: applications on natural and medical images. arXiv,

https://doi.org/10.48550/ARXIV.2212.01577, 3 December 2022, preprint: not peer reviewed.
77. Korkinof D, Harvey H, Heindl A *et al.* Perceived realism of high-resolution generative adversarial network-derived synthetic mammograms. *Radiol Artif Intell* 2021;3:e190181. https://doi.org/10.1148/ryai.2020190181
78. Havaei M, Mao X, Wang Y *et al.* Conditional generation of medical images via disentangled adversarial inference. *Med Image Anal* 2021;72:102106. https://doi.org/10.1016/j.media.2021.102106
79. Dong J, Liu C, Man P *et al.* Fp-GAN with fused regional features for the synthesis of high-quality paired medical images. *J Healthc Eng* 2021;2021:1. https://doi.org/10.1155/2021/6678031
80. Liang Z, Huang JX, Antani S. Image translation by ad CycleGAN for COVID-19 X-ray images: a new approach for controllable GAN. *Sensors* 2022;22:9628. https://doi.org/10.3390/s22249628
81. Nguyen LX, Le HQ, Park S-B *et al.* A new chapter for medical image generation: the stable diffusion method. In: *International Conference on Information Networking (ICOIN), 2023*, Bangkok, Thailand: IEEE, 2023. https://doi.org/10.1109/icoin56518.2023.10049010
82. Sivanesan U, Braga LH, Sonnadara RR *et al.* : Unsupervised image synthesis and segmentation based on shape priors. arXiv, https://doi.org/10.48550/ARXIV.2102.02690, 4 February 2021, preprint: not peer reviewed.
83. Barrera K, Merino A, Molina A *et al.* Automatic generation of artificial images of leukocytes and leukemic cells using generative adversarial networks (syntheticcellgan). *Comput Methods Programs Biomed* 2023;229:107314. https://doi.org/10.1016/j.cmpb.2022.107314
84. Wahid KA, Xu J, El-Habashy D *et al.* Deep-learning-based generation of synthetic 6-minute MRI from 2-minute MRI for use in head and neck cancer radiotherapy. *Front Oncol* 2022;12:975902. https://doi.org/10.3389/fonc.2022.975902
85. Konar AS, Paudyal R, Shah AD *et al.* Qualitative and quantitative performance of magnetic resonance image compilation (MAGiC) method: an exploratory analysis for head and neck imaging. *Cancers* 2022;14:3624. https://doi.org/10.3390/cancers14153624
86. Simona B, Thibeau-Sutre E, Maire A *et al.* Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation with U-Net derived models. In: Išgum I, Colliot O, *Medical Imaging: Image Processing, 2022*. San Diego, California, United States: SPIE, 2022. https://doi.org/10.1117/12.2608565
87. Fujita S, Hagiwara A, Takei N *et al.* Accelerated isotropic multiparametric imaging by high spatial resolution 3D-QALAS with compressed sensing: a phantom, volunteer, and patient study. *Invest Radiol* 2021;56:292–300. https://doi.org/10.1097/RLI.0000000000000744
88. Moeck P, DeStefano P. Accurate lattice parameters from 2D-periodic images for subsequent Bravais lattice type assignments. *Adv Struct Chem Imag* 2018;4:5. https://doi.org/10.1186/s40679-018-0051-z
89. Wang C, Ren Q, Qin X *et al.* The same modality medical image registration with large deformation and clinical application based on adaptive diffeomorphic multi-resolution demons. *BMC Med Imaging* 2018;18:21. https://doi.org/10.1186/s12880-018-0267-3
90. Ger RB, Yang J, Ding Y *et al.* Accuracy of deformable image registration on magnetic resonance images in digital and physical phantoms. *Med Phys* 2017;44:5153–61. https://doi.org/10.1002/mp.12406
91. Li H, Wu J, Miao A *et al.* Rayleigh-maximum-likelihood bilateral filter for ultrasound image enhancement. *Biomed Eng Online* 2017;16:46. https://doi.org/10.1186/s12938-017-0336-9
92. Miao J, Huang F, Narayan S *et al.* A new perceptual difference model for diagnostically relevant quantitative image quality evaluation: a preliminary study. *Magn Reson Imaging* 2013;31:596–603. https://doi.org/10.1016/j.mri.2012.09.009
93. Mori S, Hirai R, Sakata Y *et al.* Deep neural network-based synthetic image digital fluoroscopy using digitally reconstructed tomography. *Phys Eng Sci Med* 2023;46:1227–37. https://doi.org/10.1007/s13246-023-01290-z
94. Goncalves A, Ray P, Soper B *et al.* Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 2020;20:108. https://doi.org/10.1186/s12874-020-00977-1
95. Achterberg JL, Haas MR, Spruit MR. On the evaluation of synthetic longitudinal electronic health records. *BMC Med Res Methodol* 2024;24:181. https://doi.org/10.1186/s12874-024-02304-4
96. Yan C, Zhang Z, Nyemba S *et al.* Generating synthetic electronic health record data using generative adversarial networks: tutorial. *JMIR AI* 2024;3:e52615. https://doi.org/10.2196/52615
97. Budu E, Etminani K, Soliman A *et al.* Evaluation of synthetic electronic health records: a systematic review and experimental assessment. *Neurocomputing* 2024;603:128253. https://doi.org/10.1016/j.neucom.2024.128253
98. Ghosheh GO, Li J, Zhu T. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Comput Surv* 2024;56:1–34. https://doi.org/10.1145/3636424
99. Hernandez M, Epelde G, Alberdi A *et al.* Synthetic data generation for tabular health records: a systematic review. *Neurocomputing* 2022;493:28–45. https://doi.org/10.1016/j.neucom.2022.04.053
100. Hernadez M, Epelde G, Alberdi A *et al.* Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods Inf Med* 2023;62:e19–38. https://doi.org/10.1055/s-0042-1760247
101. Yoon J, Mizrahi M, Ghalaty NF *et al.* EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digit Med* 2023;6:141. https://doi.org/10.1038/s41746-023-00888-7
102. Bietsch D, Stahlbock R, Voß S. Synthetic data as a proxy for real-world electronic health records in the patient length of stay prediction. *Sustainability* 2023;15:13690. https://doi.org/10.3390/su151813690
103. El Emam K, Mosquera L, Fang X *et al.* Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR Med Inform* 2022;10:e35734. https://doi.org/10.2196/35734
104. Yan C, Yan Y, Wan Z *et al.* A Multifaceted benchmarking of synthetic electronic health record generation models. *Nat Commun* 2022;13:7609. https://doi.org/10.1038/s41467-022-35295-1
105. Sun C, van Soest J, Dumontier M. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *J Biomed Inform* 2023;143:104404. https://doi.org/10.1016/j.jbi.2023.104404
106. Kharya S, Soni S, Swarnkar T. Generation of synthetic datasets using weighted bayesian association rules in clinical world. *Int J Inf Tecnol* 2022;14:3245–51. https://doi.org/10.1007/s41870-022-01081-x
107. Narteni S, Orani V, Ferrari E *et al.* A new XAI-based evaluation of generative adversarial networks for IMU data augmentation. In: *IEEE International Conference on E-health Networking, Application & Services (HealthCom), 2022*. Genoa, Italy: IEEE, 2022. https://doi.org/10.1109/healthcom54947.2022.9982780
108. Ji E, Ohn JH, Jo H *et al.* Evaluating the utility of data integration with synthetic data and statistical matching. *Sci Rep* 2025;15:19627. https://doi.org/10.1038/s41598-025-01514-0
109. Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: preserving data utility and patient privacy. *Comput Intell* 2021;37:819–51. https://doi.org/10.1111/coin.12427

110. Welcome to ELIXIR. https://elixir-europe.org (16 March 2025, date last accessed).
111. ELIXIR-STEERS. https://elixir-europe.org/about-us/how-funded/eu-projects/steers. (16 March 2025, date last accessed).
112. EOSC Association. Review of software quality attributes and characteristics. https://eosc.eu/task-force-deliverab/review-of-software-quality-attributes-and-characteristics. (16 March 2025, date last accessed).
113. European Virtual Institute for Research Software Excellence. The EVERSE project. https://everse.software. (16 March 2025, date last accessed).
114. SYNTHIA. https://www.ihi-synthia.eu. (16 March 2025, date last accessed).
115. SYNTHEMA. https://synthema.eu. (16 March 2025, date last accessed).
116. Lekadir K, Frangi AF, Porras AR *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 2025;388:e081554. https://doi.org/10.1136/bmj-2024-081554
117. European Commission. AI for health: evaluation of applications & datasets. CORDIS | European Commission 2024. https://cordis.europa.eu/project/id/101183031. (16 March 2025, date last accessed).

Received: September 27, 2025. **Revised:** January 7, 2026. **Accepted:** January 19, 2026