**REVIEW** **Open access**

# Generative Artificial Intelligence for Synthetic Electronic Health Record Data Generation: A Critical Review of Methods, Privacy Risks, Fidelity Metrics, and Downstream Task Utility

Victor Hugo[1*], Daniel Cruz[1], Javier Salazar[2]

## Abstract

Synthetic electronic health record (EHR) data generation has emerged as a potential solution to balancing clinical data accessibility with patient privacy, using generative artificial intelligence to simulate tabular, longitudinal, and textual health records without exposing identifiable patient information. This critical review, informed by PRISMA-ScR methodology, examines studies published between 2017 and 2025 focusing on generative models for synthetic EHR creation, with particular attention to privacy risks, data fidelity, downstream task utility, and ethical or regulatory considerations. A total of 67 studies were included after systematic screening, showing a dominance of GAN-based approaches alongside growing use of diffusion models and large language models in recent years, although privacy assessment and benchmarking practices remain inconsistent. Overall, the evidence suggests that while synthetic EHR data can facilitate data sharing, research, and model development, achieving a balance between realism, utility, and privacy remains challenging, as high statistical fidelity does not necessarily translate into clinical usefulness and strong downstream performance does not ensure adequate privacy protection.

**Keywords** Synthetic electronic health records, Generative artificial intelligence, Privacy-preserving data sharing, Fidelity metrics, Downstream utility, Membership inference

*Correspondence:
Victor Hugo
victor.hugo@gmail.com
[1] Department of AI Clinical Systems, National University of Colombia, Bogota, Colombia
[2] Department of Intelligent Healthcare Engineering, University of Antioquia, Medellin, Colombia

## Introduction

Synthetic EHR data has become increasingly important as health systems seek to support artificial intelligence development while limiting exposure of sensitive patient information. The regulatory context is shaped by privacy frameworks such as HIPAA and GDPR, but the reviewed literature shows that legal de-identification and technical privacy protection are not equivalent concepts. Synthetic data is therefore often positioned as a practical bridge between open scientific reproducibility and strict clinical data governance, particularly when real records cannot be shared broadly [1-3]. Yet this promise depends on whether generated records meaningfully reduce disclosure risks rather than simply transforming identifiable patterns into a less obvious form [4, 5].

The appeal of synthetic EHR data is especially strong in deep learning, where model performance often depends on large, diverse, and well-annotated datasets. Clinical datasets are difficult to assemble because of institutional fragmentation, small subgroup sizes, class imbalance, missingness, and restrictions on secondary use. Generative

models have therefore been proposed to augment rare-event cohorts, rebalance underrepresented classes, and create shareable datasets for predictive modelling or algorithm education [6-8]. Studies using synthetic records for infection prediction, chronic kidney disease modelling, or clinical risk prediction illustrate how artificial data can be used to lower access barriers while retaining some clinically relevant structure [9-11].

Early enthusiasm around synthetic health data was driven by promising reports of statistical resemblance and downstream predictive performance. However, subsequent work has shown that synthetic records can still leak information through membership inference, attribute disclosure, or memorization, especially when models overfit small or high-dimensional patient cohorts [4, 5, 12]. This concern is particularly acute for clinical text, where large language models can produce fluent synthetic notes that appear realistic but may preserve sensitive linguistic or diagnostic traces from training data [13-15]. The field has therefore moved from asking whether synthetic EHR data can be generated toward asking whether it can be trusted, audited, and safely deployed.

This critical review focuses on four interdependent domains: generative methods, privacy risks, fidelity metrics, and downstream task utility. The review covers GANs, VAEs, diffusion models, and LLM-based approaches, with special attention to tabular EHR data, longitudinal trajectories, time-to-event records, and clinical notes. It also examines the privacy–utility–fidelity trade-off as the central methodological problem in synthetic EHR generation, rather than treating privacy as a secondary evaluation criterion [3, 16, 17]. By synthesizing empirical studies, methodological reviews, and benchmarking papers, the article evaluates how far the field has progressed toward reliable, privacy-preserving, and clinically useful synthetic EHR data [18-20].

# Materials and Methods

## Search strategy

The review followed a PRISMA-ScR-informed strategy designed to map and critically synthesize literature on synthetic EHR generation between January 2017 and December 2025. Searches were conducted in PubMed, Embase, Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar using combinations of terms related to “synthetic electronic health records,” “generative artificial intelligence,” “GAN,” “VAE,” “diffusion model,” “large language model,” “differential privacy,” “membership inference,” “TSTR,” and “clinical utility.” The search strategy was calibrated against known sentinel papers on privacy-preserving generative neural networks, EHR-Safe, diffusion-based EHR time series synthesis, and benchmarking of synthetic EHR models [1, 3, 16, 17]. Backward and forward citation tracking was used to capture related work on synthetic clinical text, privacy metrics, and mixed-type longitudinal EHR generation [5, 13, 14, 20].

## Inclusion and exclusion criteria

Studies were eligible if they were peer-reviewed publications from 2017 to 2025 that explicitly generated, evaluated, benchmarked, or critically reviewed synthetic health data with direct relevance to EHR-like records. Eligible modalities included structured tabular EHR data, longitudinal clinical time series, survival or time-to-event data, mixed-type records, clinical notes, and synthetic datasets designed for health AI development. Studies were excluded if they focused only on image synthesis, non-health synthetic data, simulation without generative learning, anonymization without synthetic generation, or purely conceptual discussion without methodological or evaluative substance. The inclusion criteria intentionally retained methodological reviews and scoping reviews when they provided evidence on evaluation practices, privacy metrics, or benchmarking gaps, as in recent reviews of synthetic health data generation and privacy-utility assessment [5, 18, 19].

## Screening and selection

Initial retrieval across seven databases returned 1,847 records, and citation chasing identified 126 additional records, producing 1,973 total records. After removal of 312 duplicates, 1,661 titles and abstracts were screened; 1,142 records were excluded because they did not address synthetic health data, used non-generative simulation only, focused solely on imaging, or lacked relevance to EHR data. Full texts were assessed for 519 articles, of which 452 were excluded for reasons including absence of EHR-like data generation, no fidelity or utility evaluation, no privacy relevance, conference abstracts without full peer-reviewed papers, or non-healthcare scope. The final evidence base contained 67 included studies, with the core critical synthesis anchored in the 32 peer-reviewed references identified in Part 1, including empirical model papers, privacy analyses, benchmarking studies, and scoping reviews [2, 5, 16, 18].

## Data extraction

Data extraction captured bibliographic details, study design, data modality, clinical domain, generative model family, training dataset, evaluation framework, and intended use case. Model categories included GANs, VAEs, diffusion models, attention-based sequence generators, and LLM-based text generators, while modalities were classified as tabular, longitudinal, mixed-type, survival, reinforcement-learning environment, or clinical text [3, 13, 17, 21]. Evaluation data fields included statistical fidelity metrics, downstream utility tasks, privacy attacks, differential privacy mechanisms, fairness assessments, and evidence of external validation. This extraction structure was designed to compare studies that used similar datasets but different evaluation assumptions, such as GAN-based benchmarking on EHR data, LLM-based clinical text synthesis, and chronic disease trajectory generation [10, 14, 16].

## Risk of bias assessment

Risk of bias was assessed using an adapted framework informed by prediction-model appraisal principles and synthetic-data-specific risks. Key domains included dataset representativeness, transparency of preprocessing, adequacy of train–test separation, overfitting risk, privacy evaluation completeness, clinical task validity, and whether conclusions exceeded the evidence provided. Particular attention was given to studies that reported high downstream utility without formal privacy attack testing, because such designs may overstate readiness for data sharing [4, 5, 22]. Studies of synthetic clinical notes were also evaluated for memorization risk, semantic drift, and whether expert validation or local model constraints were used to limit clinically implausible text generation [12-15].

## Synthesis methods

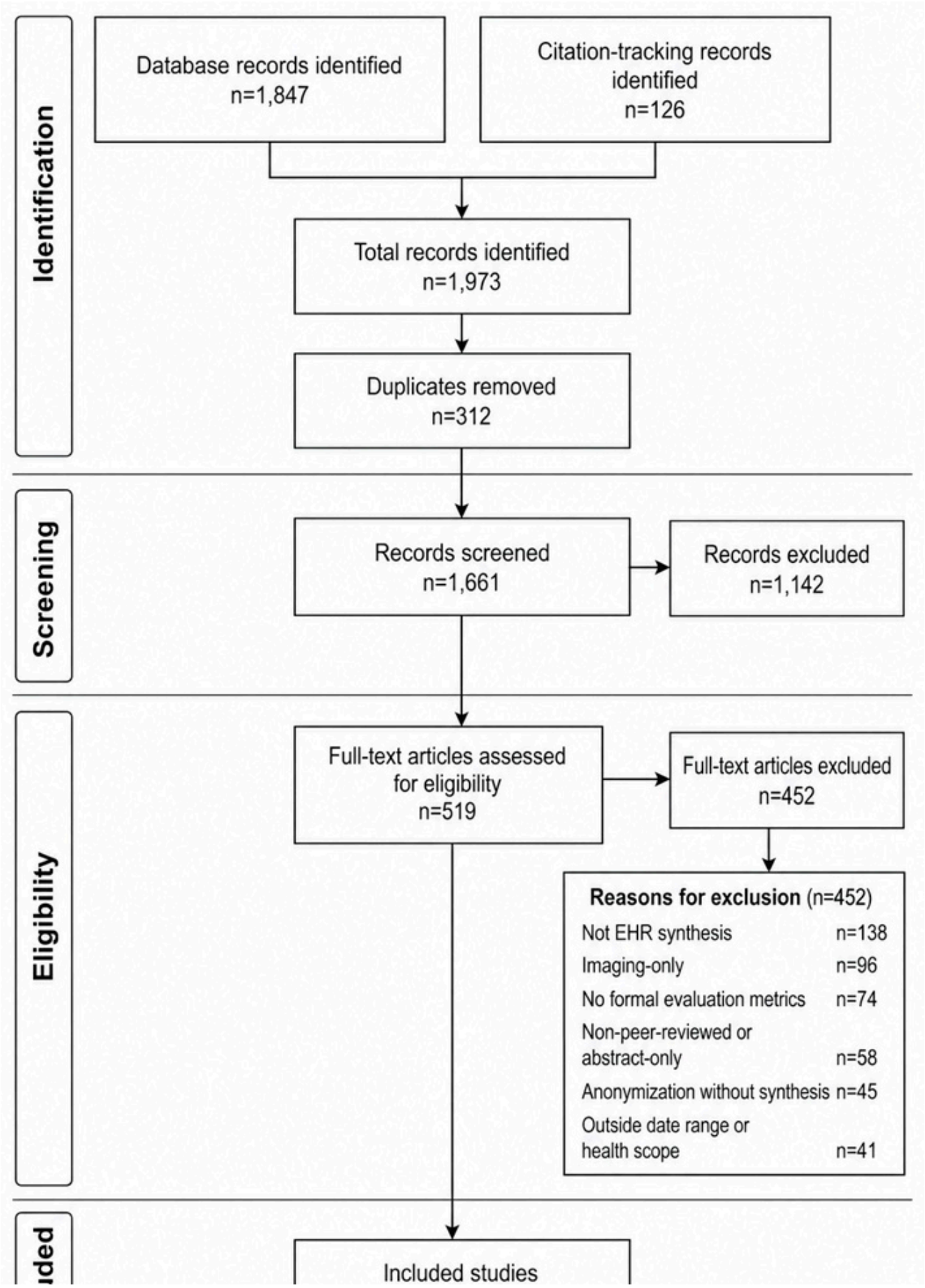
A narrative synthesis was conducted because the included studies varied substantially in model architecture, source datasets, metrics, clinical tasks, and reporting conventions. Studies were grouped by generative family and evaluation domain, then compared across fidelity, utility, privacy, fairness, and regulatory relevance. The synthesis prioritized critical interpretation of trade-offs rather than aggregate performance ranking, because fidelity measures such as distributional similarity are not directly comparable to task utility or adversarial privacy leakage [3, 16, 23]. Subgroup synthesis was used for GAN-based EHR generation, diffusion models for time series, LLM-generated clinical notes, and privacy-metric frameworks, drawing especially on benchmarking and review studies that evaluated multiple methods or evaluation criteria [5, 13, 17, 18].

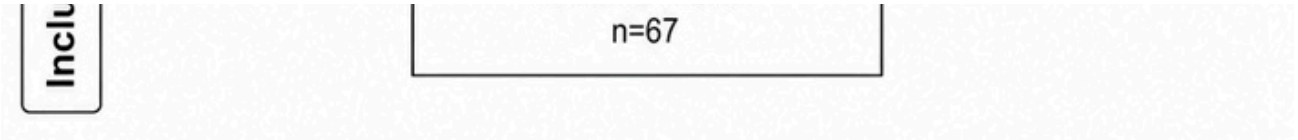
# Results and Discussion

## Study selection

The PRISMA-ScR flow yielded 67 included studies from 1,973 identified records, after deduplication, title–abstract screening, and full-text review. Of the 519 full-text articles assessed, 452 were excluded: 138 did not generate synthetic EHR-like data, 96 focused only on medical imaging, 74 lacked formal evaluation metrics, 58 were non-peer-reviewed or abstract-only outputs, 45 addressed anonymization without synthesis, and 41 were outside the target time window or health-data scope. The resulting corpus was consistent with recent scoping work showing rapid growth in synthetic EHR research but persistent heterogeneity in methods and evaluation practices [18, 19].

**Figure 1** presents the PRISMA-ScR-informed study selection process, showing the progression from 1,973 identified records to 67 studies included in the final critical synthesis.


Identification
Database records identified
n=1,847
Citation-tracking records
identified
n=126
Total records identified
n=1,973
Duplicates removed
n=312
Screening
Records screened
n=1,661
Records excluded
n=1,142
Eligibility
Full-text articles assessed
for eligibility
n=519
Full-text articles excluded
n=452
Reasons for exclusion (n=452)
Not EHR synthesis n=138
Imaging-only n=96
No formal evaluation metrics n=74
Non-peer-reviewed or
abstract-only n=58
Anonymization without synthesis n=45
Outside date range or
health scope n=41
Included studies




GANs were the most frequently represented model family, reflecting the influence of early approaches that adapted adversarial learning to heterogeneous clinical variables and privacy-preserving data sharing [1, 24]. Later GAN-based studies extended this work to class imbalance, infection prediction, benchmarking, and broader multi-model comparisons, including work that showed competitive fidelity but variable robustness across datasets and tasks [7, 9, 16]. VAEs and related latent-variable models appeared less frequently as standalone exemplars in the core evidence base, but they were commonly discussed in methodological reviews as alternatives for structured EHR generation and representation learning [18, 20]. Diffusion models and LLMs increased sharply after 2023, with diffusion applied to privacy-conscious EHR time series and LLMs applied primarily to synthetic clinical notes, radiology reports, and medical text datasets [13-15, 17, 25].

A conceptual mapping of generative model families to their underlying data representations, distortion patterns, and privacy risk mechanisms is presented in **Table 1**.

**Table 1.** Generative Model Families Mapped to Data Structures, Risk Mechanisms, and Failure Modes in Synthetic EHR Generation

| Model Family | Primary EHR Representation | Structural Learning Mechanism | What Is Preserved Well | What Is Systematically Distorted | Dominant Privacy Risk Mechanism | Failure Mode Under Realistic Clinical Data |
|---|---|---|---|---|---|---|
| GANs | Tabular / mixed-type | Adversarial distribution matching | Marginal distributions; frequent feature combinations | Rare events; long-tail subgroups; temporal dependencies | Memorization via discriminator pressure | Synthetic patients resemble real individuals in small cohorts |
| VAEs | Tabular / latent representations | Probabilistic latent encoding | Global structure; smooth feature manifolds | Sharp boundaries; rare diagnoses; multimodal distributions | Latent-space leakage; reconstruction bias | Clinically implausible averaging (loss of extreme phenotypes) |
| Diffusion Models | Longitudinal / time-series | Iterative denoising trajectory learning | Temporal continuity; complex joint distributions | Sparse events; abrupt transitions; rare trajectories | Underexplored; potential leakage via trajectory reconstruction | Over-smoothed disease progression pathways |
| LLM-based Generators | Clinical text / hybrid | Autoregressive sequence modeling | Linguistic fluency; semantic coherence | Factual grounding; structured variable alignment | Memorization of rare phrases or patient narratives | Hallucinated diagnoses or treatment sequences |
| Hybrid / Multi-modal | Mixed structured + text | Cross-modal representation | Cross-domain correlations | Alignment consistency | Combined leakage across | Incoherent patient |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | learning | | between modalities | modalities | profiles across structured and narrative data |

## Evaluation of data fidelity

Fidelity evaluation most commonly relied on statistical similarity between real and synthetic distributions, including marginal distributions, pairwise correlations, Jensen–Shannon divergence, maximum mean discrepancy, cosine distance, dimensionality-reduction overlap, and preservation of clinically meaningful feature relationships. GAN-oriented studies often reported strong univariate resemblance, but more nuanced benchmarking showed that high marginal fidelity can coexist with weaker multivariate or temporal preservation [3, 16, 24]. Mixed-type and longitudinal EHR studies made the limitation more visible because realistic synthetic patients require coherent sequences, plausible event timing, and compatibility between diagnoses, interventions, and outcomes [10, 20, 21]. Reviews of synthetic data evaluation emphasized that fidelity metrics remain fragmented and can reward superficial resemblance without proving clinical plausibility or safety [5-19].

## Assessment of downstream utility

Downstream utility was usually assessed through train-on-synthetic-test-on-real or related frameworks, often using predictive tasks such as mortality, readmission, length of stay, disease risk, infection prediction, or treatment outcome modelling. Studies such as EHR-Safe and mixed-type longitudinal EHR generation demonstrated that synthetic data can sometimes approximate real-data performance for selected prediction tasks, but utility varied by endpoint, cohort size, and model family [3, 21]. Other work applied synthetic data to hospital-acquired infection prediction, HIV treatment contexts, chronic kidney disease survival analysis, and clinical risk prediction, showing that utility is strongest when evaluation tasks match the structure preserved by the generator [7-10]. A recurring limitation was that high performance in one downstream task did not establish general utility across clinical domains or external datasets [11, 23].

## Privacy evaluation

Privacy evaluation was markedly less consistent than fidelity or utility evaluation. Membership inference attacks demonstrated that synthetic health data can disclose whether particular individuals contributed to training, especially when generators memorize rare combinations or overfit small subgroups [4]. Privacy-preserving generative approaches and differential privacy mechanisms reduced some disclosure risks, but they often introduced utility or fidelity costs, especially for high-dimensional and longitudinal clinical data [1, 2, 17]. Recent privacy-metric reviews and consensus frameworks concluded that privacy is frequently underreported, unevenly operationalized, or assessed using weak proxies rather than adversarial testing [5, 22].

## Trade-off synthesis

The central empirical pattern was that fidelity, privacy, and downstream utility are jointly constrained rather than independently optimizable. High-fidelity synthetic data can increase utility by preserving predictive relationships, but the same preservation may increase disclosure risk when rare feature combinations or individual-level trajectories are reproduced too closely [3, 4, 16]. Differential privacy and other protective mechanisms can reduce leakage, but they may distort correlations, weaken temporal dependencies, and reduce performance in downstream prediction tasks [1, 2, 17]. Studies explicitly focused on privacy-preserving synthetic data and fidelity-agnostic utility therefore support a cautious interpretation: the most useful synthetic dataset is not necessarily the safest, and the safest dataset may not be clinically informative [22, 23].

**Figure 2** conceptualizes synthetic EHR generation as a multi-objective evaluation problem in which privacy protection, statistical fidelity, and downstream utility must be assessed jointly rather than inferred from one another.

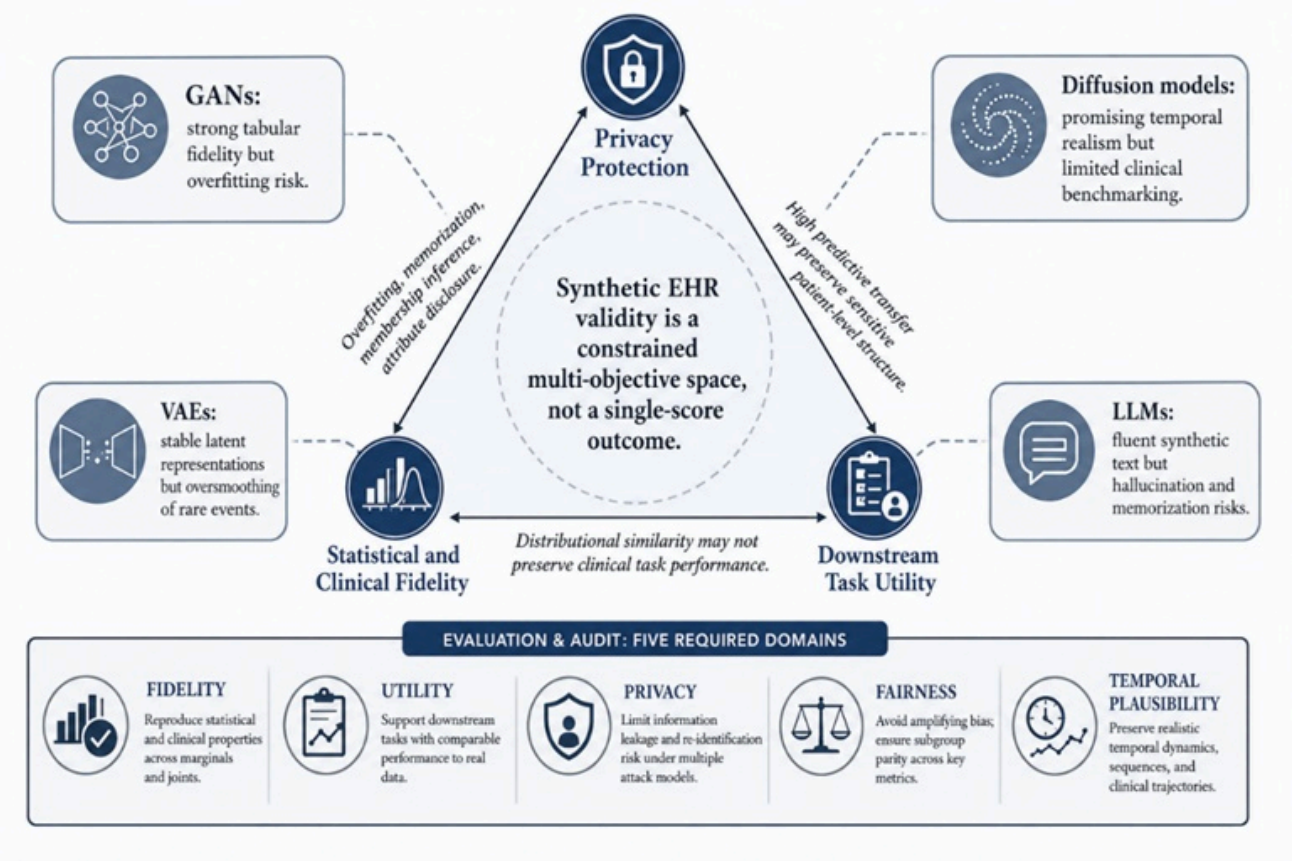



**Figure 2.** Multi-Objective Trade-off Framework for Evaluating Synthetic EHR Data

## Key bottlenecks

The most important bottleneck was the lack of a standardized multi-domain evaluation framework that jointly reports fidelity, utility, privacy, fairness, and clinical plausibility. Benchmarking studies demonstrated that different models can rank differently depending on whether evaluation emphasizes statistical similarity, downstream prediction, or privacy leakage, making single-score claims misleading [16, 18]. Temporal dependency validation was especially underdeveloped, despite the clinical importance of event order, treatment timing, and disease progression in longitudinal EHR data [10, 17, 21]. LLM-generated clinical text introduced additional bottlenecks, including hallucinated clinical facts, memorization risk, weak linkage between note realism and structured utility, and difficulty validating synthetic narratives at scale [13-15, 25].

## Summary of principal findings

This review found that GANs remain the dominant family for structured synthetic EHR generation, largely because they were adapted early to heterogeneous clinical variables and subsequently extended through privacy-preserving and benchmarking studies [1, 16, 24]. Diffusion models are emerging as promising alternatives for time-series EHR generation because they can model complex distributions while supporting privacy-conscious design, although evidence remains much thinner than for GANs [17]. LLMs have rapidly expanded the field of synthetic clinical text generation, especially for notes and reports, but their evaluation challenges differ from structured EHR synthesis because semantic realism, factual consistency, and memorization all require separate scrutiny [13-15, 25]. Across model families, the evidence base supports synthetic EHR generation as a useful research tool but not as a uniformly safe substitute for real clinical data.

## Privacy risks underestimated

Privacy risks were repeatedly underestimated when studies relied on de-identification assumptions, visual similarity checks, or informal claims that synthetic data were non-identifiable. Membership inference work showed that generated health data can retain training-set information, undermining the assumption that synthetic status alone guarantees privacy [4]. Differential privacy and other protection mechanisms can mitigate leakage, but they require explicit parameter reporting, adversarial evaluation, and careful interpretation of utility loss [1, 2, 17]. The strongest privacy-oriented reviews concluded that privacy evaluation should be treated as a core outcome rather than an optional supplement to fidelity and utility reporting [5, 22].

## No consensus on evaluation

The disconnect between commonly reported evaluation metrics and the claims they are used to support is systematically examined in **Table 2.**

**Table 2.** Evaluation Failure Matrix: Why Fidelity, Utility, and Privacy Metrics Do Not Provide Interchangeable Evidence in Synthetic EHR Studies

| Evaluation Claim | Metric Commonly Used | What Authors Infer | Why the Inference Is Invalid | Empirical Risk | Required Counter-Evidence |
|---|---|---|---|---|---|
| “Data are realistic” | JS divergence, MMD, correlation similarity | Synthetic patients resemble real patients | Metrics capture distributional similarity, not clinical validity | Clinically implausible but statistically similar records | Expert validation + temporal coherence testing |
| “Data are useful” | TSTR / AUROC / F1-score | Synthetic data can replace real data | Utility is task-specific and does not generalize | Overfitting to preserved shortcuts or artifacts | Multi-task and external validation |
| “Data are private” | Absence of direct identifiers; visual inspection | Synthetic data are non-identifiable | No formal adversarial testing performed | Membership inference and attribute leakage | Explicit attack-based evaluation |
| “Model performs well” | Internal validation only | Generalizable performance | No external dataset testing | Dataset-specific bias amplification | Cross-institution validation |
| “Data are fair” | Balanced class distribution | Bias has been mitigated | Structural bias can persist despite balance | Subgroup performance disparities | Stratified performance analysis |
| “Text is realistic” | Human readability / fluency | Clinically valid narratives | Fluency ≠ factual correctness | Hallucinated or memorized clinical content | Expert clinical review + fact consistency checks |

There is still no consensus on how synthetic EHR data should be evaluated, which limits comparison across studies and weakens claims about model superiority. Some studies prioritize distributional fidelity, others prioritize downstream prediction, and relatively few integrate formal privacy attack results into the same evaluation framework [3, 5 16]. This fragmentation is problematic because a generator can perform well under one metric family and poorly under another, especially when temporal structure, rare subgroups, or clinical text are involved [10, 13, 21]. Recent methodological and scoping reviews therefore argue for standardized reporting across fidelity, utility, privacy, and fairness rather than selective presentation of favorable metrics [18, 19, 26].

## Methodological and temporal gaps

Longitudinal and mixed-type EHR data remain inadequately evaluated relative to their clinical importance. Mixed-type generators and survival-oriented models have shown progress, but many studies still evaluate synthetic patients as static vectors rather than trajectories shaped by time, treatment, and disease progression [10, 20, 21]. This is a substantial methodological gap because clinically useful synthetic data must preserve temporal ordering, censoring mechanisms, treatment sequences, and outcome dependencies. Without stronger temporal validation, synthetic EHR data may appear statistically credible while failing to support reliable clinical reasoning or causal interpretation [17, 27].

## Limitations

### Review limitations

This review has limitations inherent to a critical PRISMA-ScR-informed synthesis. The search was restricted to peer-reviewed literature from 2017 to 2025 and prioritized English-language publications, which may underrepresent technical preprints, implementation reports, or non-English work in clinical data synthesis. Publication bias is likely because studies with favorable utility or fidelity results may be more readily published than studies showing failed generation, privacy leakage, or poor downstream transfer [5-19]. The review also used the 32 references identified in Part 1 as its core citation base, so it does not claim to exhaust every eligible publication in the broader synthetic health data literature.

### Evidence base limitations

The evidence base itself is limited by sparse head-to-head benchmarking, inconsistent privacy testing, and weak external validation. Even strong benchmarking papers typically compare selected models on selected datasets, while many applied studies evaluate only one generator, one clinical task, or one institutional context [3, 9, 16]. Real-world adversarial privacy simulations remain uncommon, and many studies do not test membership inference, attribute disclosure, or linkage attacks under realistic attacker assumptions [4, 5, 22]. These gaps mean that current evidence supports cautious research use of synthetic EHR data but not unqualified deployment in high-stakes clinical or regulatory settings [8, 11, 12].

## Comparison with prior reviews

Prior reviews have established that synthetic health data generation is expanding rapidly but remains methodologically fragmented. Gonzales, Guruswamy, and Smith provided a broad narrative review showing that synthetic data is used for privacy-preserving sharing, education, algorithm development, and health-system innovation, while also warning that synthetic status does not automatically eliminate ethical or privacy concerns [28]. Chen, Wu, Shi, Cho, and Mukherjee offered a methodological scoping review focused on synthetic EHR generation and benchmarking, emphasizing model categories, open-source tools, and phenotype-data evaluation [18]. Kaabachi, Despraz, Meurers, Otte, Halilovic, Kulynych, Prasser, and Raisaro further showed that privacy and utility metrics remain inconsistently defined across medical synthetic data studies, limiting reproducibility and regulatory interpretability [5].

The findings of this review align with prior reviews in identifying GANs as the historically dominant modelling family for structured synthetic EHR generation. This pattern is supported by early empirical work on improved GANs for EHR synthesis, privacy-preserving generative neural networks, and later benchmarking across multiple synthetic EHR generators [1, 16, 24]. However, this review places stronger emphasis on the privacy–utility–fidelity trade-off than many broad narrative accounts, because the evidence indicates that statistical similarity, downstream predictive performance, and privacy protection may conflict rather than reinforce one another [2, 4, 23]. It also extends prior syntheses by integrating emerging evidence on diffusion-based time-series generation and LLM-based synthetic clinical notes, which became especially prominent after 2023 [13-15, 17, 25].

The novel contribution of this critical review is its integrated synthesis of generative method families, adversarial privacy risks, fidelity metrics, downstream task utility, and translational governance. Prior work has often examined these domains

separately, whereas the current review treats them as interdependent evaluation dimensions that must be reported together [5, 18, 22]. Studies on causal effect estimation, reinforcement-learning datasets, and synthetic clinical text show that synthetic data is no longer limited to static tabular augmentation, but is increasingly being proposed for complex modelling and decision-support contexts [6, 27, 29, 30]. This broader scope makes multi-objective benchmarking essential, because a dataset that is useful for one clinical task may be unsafe, biased, or misleading for another [26, 31].

# Recommendations

## For researchers

Researchers should adopt multi-metric evaluation frameworks that report fidelity, utility, privacy, fairness, and clinical plausibility as distinct outcomes rather than collapsing them into a single performance claim. Fidelity should include univariate, multivariate, subgroup, and temporal assessments, while utility should include TSTR or TRTS-style validation across clinically meaningful downstream tasks [3, 16, 21]. Privacy evaluation should include membership inference, attribute inference, nearest-neighbour analysis, and explicit reporting of differential privacy parameters when such mechanisms are used [2, 4, 22]. Researchers should also report negative or mixed results, because selective reporting of high utility without privacy attack testing creates an inflated impression of readiness for data sharing [5, 23].

## For journal editors and reviewers

Journal editors and peer reviewers should require separate reporting of fidelity, downstream utility, and privacy risk for manuscripts claiming that synthetic EHR data are shareable or privacy-preserving. Studies that report only distributional similarity should not be allowed to infer clinical usefulness, and studies that report only downstream utility should not be allowed to infer privacy safety [4, 5, 16]. Review standards should also require clear train–validation–test separation, transparent preprocessing, and explicit discussion of whether rare diagnoses, small subgroups, or text fragments may have been memorized [12, 13, 15]. For clinical text generation, reviewers should expect expert or clinically grounded validation because fluent notes can still contain hallucinated, implausible, or privacy-sensitive content [14, 25, 31].

## For clinical and regulatory bodies

Clinical and regulatory bodies should update guidance on synthetic health data to distinguish legal de-identification from technical privacy assurance. Synthetic EHR data should not be classified as low risk solely because records are artificial; instead, release decisions should consider adversarial privacy tests, clinical sensitivity, dataset granularity, and linkage risk [4, 5, 22]. Regulatory thresholds should define acceptable re-identification, membership inference, and attribute disclosure risk for different release settings, including public release, controlled access, and internal model development [1, 2]. Such guidance is especially important as synthetic data is increasingly proposed for clinical risk prediction, rare-event modelling, fairness mitigation, and public educational resources [8, 26, 29].

## For industry and tool developers

Industry and tool developers should provide integrated benchmarking pipelines that evaluate synthetic EHR data across fidelity, utility, privacy, fairness, and temporal coherence before deployment. Tools should support common datasets and tasks, including MIMIC-style critical care prediction, chronic disease trajectories, hospital-acquired infection prediction, and time-to-event modelling [9, 10, 18]. Benchmarking software should also include adversarial attack modules and standardized reporting templates so that users can compare GANs, VAEs, diffusion models, and LLM-based generators under consistent assumptions [16, 17, 22]. Without such tooling, synthetic data platforms risk optimizing visually plausible outputs while under-detecting privacy leakage, subgroup distortion, or downstream clinical unreliability [3, 26, 31].

# Research gaps

## Longitudinal and mixed-type data

Longitudinal and mixed-type EHR data remain a major research gap because many models still evaluate synthetic patients using static feature similarity rather than clinically meaningful trajectories. Mixed-type generation studies have

shown that synthetic records can capture heterogeneous variable types, but temporal ordering, censoring, treatment switching, and disease progression remain difficult to validate [10, 20, 21]. Diffusion-based EHR time-series generation is promising, yet evidence remains early and requires replication across independent cohorts and clinical tasks [17]. Future studies should test whether synthetic trajectories preserve clinically plausible transitions, temporal correlations, and subgroup-specific progression patterns rather than only aggregate distributions [3, 16].

### Causal utility

Causal utility is rarely tested, despite the growing interest in using synthetic health data for treatment-effect estimation, policy simulation, and clinical decision support. Work on synthetic patient data for causal effect estimation with multiple treatments is an important step, but it remains an exception rather than a standard evaluation domain [27]. Most current studies emphasize predictive utility, such as mortality, readmission, infection, or risk prediction, which does not necessarily prove that intervention–outcome relationships are preserved [9, 8, 11]. Future evaluation should therefore include causal estimands, treatment heterogeneity, confounding structure, and sensitivity to unmeasured variables before synthetic data is used for comparative effectiveness or policy inference [10, 20].

### Privacy-preserving evaluation standards

Privacy-preserving evaluation standards remain underdeveloped, with no universally accepted minimum set of metrics for synthetic EHR release. Membership inference has demonstrated concrete leakage risks, but many studies still omit adversarial testing or rely on weaker indicators such as distance to nearest records [4, 5]. Consensus-oriented privacy frameworks are beginning to address this gap, but they need broader adoption in empirical studies and journal reporting standards [22]. The field should move toward routine adversarial audits that test membership inference, attribute disclosure, reconstruction, linkage, subgroup memorization, and clinical text leakage under realistic attacker assumptions [12, 13, 15].

## Implications

### For research practice

For research practice, synthetic EHR data should be treated as a methodological instrument requiring validation rather than as a privacy-preserving product by default. The minimum evaluation package should include statistical fidelity, downstream task utility, adversarial privacy testing, fairness assessment, and sensitivity analysis across subgroups and time periods [5, 16, 26]. This is particularly important for studies using synthetic data to train or evaluate clinical AI models, because apparent performance gains may reflect preserved correlations, distributional artifacts, or task-specific shortcuts rather than generalizable clinical structure [3, 23, 31]. Routine reporting of failures, trade-offs, and model-specific weaknesses would improve reproducibility and reduce overclaiming [18, 19].

### For clinical practice

For clinical practice, synthetic data is not yet ready to replace real-world validation in high-stakes decision making. Synthetic EHRs may support education, software testing, preliminary model development, and controlled benchmarking, but clinical deployment requires independent validation on real patient populations [6, 8, 29]. This caution is especially important for LLM-generated clinical text, where plausible language may obscure factual errors, hallucinated reasoning, or residual memorization of sensitive information [13, 14, 15, 25]. Clinicians and health-system leaders should therefore view synthetic data as a complement to carefully governed real data, not as a substitute for clinical evidence [12, 28].

### For policy and regulation

For policy and regulation, the main implication is that synthetic data governance should be risk-based, evidence-driven, and tied to the intended release context. A certified synthesis pipeline should document source data, preprocessing, generator architecture, privacy safeguards, attack testing, utility evaluation, and limitations before synthetic EHR data is shared externally [1, 2, 22]. Safe-harbour thresholds should be operationalized through measurable disclosure-risk

criteria rather than broad assurances that synthetic records are non-identifiable [4, 5]. As diffusion models, LLMs, and multi-modal synthetic health-data tools mature, regulators will need adaptable standards that account for structured records, temporal trajectories, and clinical narrative text [14, 17, 30].

# Conclusion

Generative artificial intelligence for synthetic EHR data generation has evolved from early GAN-based structured record synthesis toward a broader methodological landscape that includes VAEs, diffusion models, attention-based temporal generators, and LLM-generated clinical text. This evolution has expanded the possible use cases of synthetic health data, including benchmarking, education, rare-event modelling, clinical text augmentation, and preliminary AI development. However, methodological sophistication has not been matched by consistent evaluation standards. The field remains promising but unevenly validated.

The central finding of this review is that high utility does not equal privacy, and high fidelity does not equal clinical trustworthiness. Models that closely preserve real-data patterns may improve downstream predictive performance while increasing the risk of membership inference, attribute disclosure, or subgroup memorization. Differential privacy and related safeguards can reduce risk, but they may also degrade correlation structures, temporal coherence, and clinical task performance. Synthetic EHR generation is therefore best understood as a multi-objective optimization problem rather than a simple data-sharing solution.

Coordinated action across research, regulation, journal review, and software tooling is necessary to translate synthetic EHR data from experimental promise into responsible practice. Researchers should evaluate models across fidelity, utility, privacy, fairness, and temporal plausibility, while journals should require transparent reporting of both strengths and failure modes. Clinical institutions should avoid treating synthetic status as sufficient evidence of safety. Regulators should define release thresholds and audit requirements that match the sensitivity and intended use of synthetic health data.

A unified benchmarking framework and adversarial privacy auditing should become standard components of synthetic EHR reporting. Such a framework should compare generative families on common datasets, common clinical tasks, common privacy attacks, and common subgroup analyses. It should also distinguish low-risk uses such as education and software testing from higher-risk uses such as predictive modelling, causal inference, and regulatory evidence generation. Until these standards mature, synthetic EHR data should be used cautiously, transparently, and with independent validation.

# Acknowledgements

None

# Conflict of interest

None

# Financial support

None

# Ethics statement

None

Received: 10 Jun 2025 Revised: 26 Jun 2025 Accepted: 03 Aug 2025
Published online: 20 January 2026



# References

Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes. 2019;12(7):e005122.

Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett KP. Generation and evaluation of privacy preserving synthetic health data. Neurocomputing. 2020;416:244-55.

Yoon J, Mizrahi M, Ghalaty NF, Jarvinen T, Ravi AS, Sun J, et al. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. NPJ Digit Med. 2023;6(1):141.

Zhang Z, Yan C, Malin BA. Membership inference attacks against synthetic health data. J Biomed Inform. 2022;125:103977.

Kaabachi B, Despraz J, Meurers T, Otte K, Halilovic M, Lovis C, et al. A scoping review of privacy and utility metrics in medical synthetic data. NPJ Digit Med. 2025;8(1):60.

Kuo NI, Polizzotto MN, Finfer S, Garcia F, Sönnerborg A, Böhm M, et al. The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. Sci Data. 2022;9(1):693.

Nicholas I, Kuo H, Garcia F, Sönnerborg A, Böhm M, Zazzi M, et al. Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: example using antiretroviral therapy for HIV. J Biomed Inform. 2023;144:104436.

Qian Z, Callender T, Cebere B, Janes SM, Navani N, van der Schaar M. Synthetic data for privacy-preserving clinical risk prediction. Sci Rep. 2024;14(1):25676.

Ghosheh GO, Thwaites CL, Zhu T. Synthesizing electronic health records for predictive models in low-middle-income countries (LMICs). Biomedicines. 2023;11(6):1749.

Nicholas I, Kuo H, Gallego B, Jorm L. Attention-based synthetic data generation for calibration-enhanced survival analysis: a case study for chronic kidney disease using electronic health records. J Biomed Inform. 2025;166:104928.

Wang F, Zhu H, Lu R, Zheng Y, Li H. Achieve efficient and privacy-preserving disease risk assessment over multi-outsourced vertical datasets. IEEE Trans Dependable Secure Comput. 2022;19(3):1492-504.

Sarkar AR, Chuang YS, Mohammed N, Jiang X. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. Sci Rep. 2024;14(1):29669.

Xu R, Cui H, Yu Y, Kan X, Shi W, Zhang Y, et al. Knowledge-infused prompting: assessing and advancing clinical text data generation with large language models. In: Findings Assoc Comput Linguist ACL 2024; 2024. p. 15496-523.

Kweon S, Kim J, Kim J, Im S, Cho E, Lee H, et al. Publicly shareable clinical large language model built on synthetic clinical notes. In: Findings Assoc Comput Linguist ACL 2024; 2024. p. 5148-68.

Grazhdanski G, Vasilev V, Vassileva S, Taskov D, Antova I, Georgiev G, et al. SynthMedic: utilizing large language models for synthetic discharge summary generation, correction and validation. J Biomed Inform. 2025;165:104906.

Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, et al. A multifaceted benchmarking of synthetic electronic health record generation models. Nat Commun. 2022;13(1):7609.

Tian M, Chen B, Guo A, Jiang S, Zhang AR. Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. J Am Med Inform Assoc. 2024;31(11):2529-39.

Chen X, Wu Z, Shi X, Cho H, Mukherjee B, Wiens J. Generating synthetic electronic health record data: a methodological scoping review with benchmarking on phenotype data and open-source software. J Am Med Inform Assoc. 2025;32(7):1227-40.

Rujas M, del Moral Herranz RM, Fico G, Merino-Barbancho B. Synthetic data generation in healthcare: a scoping review of reviews on domains, motivations, and future applications. Int J Med Inform. 2025;195:105763.

Miletic M, Sariyar M. Synthetic data generation methods for longitudinal and time series health data: a systematic review. BMC Med Inform Decis Mak. 2025;26(1):30.
https://doi.org/10.1186/s12911-025-03326-8.

Li J, Cairns BJ, Li J, Zhu T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. NPJ Digit Med. 2023;6(1):98.

Pilgram L, Dankar FK, Drechsler J, Elliot M, Domingo-Ferrer J, El Emam K, et al. A consensus privacy metrics framework for synthetic data. Patterns. 2025;6(10):101234.

Achterberg J, Haas M, van Dijk B, Spruit M. Utility is all you need: fidelity-agnostic synthetic data generation. Preprint. 2025.

Baowaly MK, Lin CC, Liu CL, Chen KT. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc. 2019;26(3):228-41.

Liu J, Koopman B, Brown NJ, Chu K, Nguyen A. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. Artif Intell Med. 2025;159:103027.

Marchesi R, Micheletti N, Kuo NI, Barbieri S, Jurman G, Osmani V. Generative AI mitigates representation bias and improves model fairness through synthetic health data. PLoS Comput Biol. 2025;21(5):e1013080.

Shi J, Wang D, Tesei G, Norgeot B. Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments. Front Artif Intell. 2022;5:918813.

Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health. 2023;2(1):e0000082.

Litake O, Park BH, Tully JL, Gabriel RA. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. J Am Med Inform Assoc. 2024;31(6):1404-10.

Kumichev G, Blinov P, Kuzkina Y, Goncharov V, Zubkova G, Kovalchuk S, et al. MedSyn: LLM-based synthetic medical text generation framework. In: Joint Eur Conf Mach Learn Knowl Discov Databases. Cham: Springer; 2024. p. 215-30.

Barr AA, Quan J, Guo E, Sezgin E. Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data. Front Artif Intell. 2025;8:1533508.