BIBLIOGRAPHIC INFORMATION SYSTEM

JOURNAL FULL TITLE: Journal of Biomedical Research & Environmental Sciences
ABBREVIATION (NLM): J Biomed Res Environ Sci **ISSN:** 2766-2276 **WEBSITE:** https://www.jelsciences.com

SCOPE & COVERAGE

- **Sections Covered:** 34 specialized sections spanning 143 topics across Medicine, Biology, Environmental Sciences, and General Science
- Ensures broad interdisciplinary visibility for high-impact research.

PUBLICATION FEATURES

- **Review Process:** Double-blind peer review ensuring transparency and quality
- **Time to Publication:** Rapid 21-day review-to-publication cycle
- **Frequency:** Published monthly
- **Plagiarism Screening:** All submissions checked with iThenticate

INDEXING & RECOGNITION

- **Indexed in:** Google Scholar, IndexCopernicus (**ICV 2022: 88.03**)
- **DOI:** Registered with CrossRef (**10.37871**) for long-term discoverability
- **Visibility:** Articles accessible worldwide across universities, research institutions, and libraries

OPEN ACCESS POLICY

- Fully Open Access journal under Creative Commons Attribution 4.0 License (CC BY 4.0)
- Free, unrestricted access to all articles globally

GLOBAL ENGAGEMENT

- **Research Reach:** Welcomes contributions worldwide
- **Managing Entity:** SciRes Literature LLC, USA
- **Language of Publication:** English

SUBMISSION DETAILS

- Manuscripts in Word (.doc/.docx) format accepted

SUBMISSION OPTIONS

- **Online:** https://www.jelsciences.com/submit-your-paper.php
- **Email:** support@jelsciences.com, support@jbresonline.com

HOME ABOUT ARCHIVE SUBMIT MANUSCRIPT APC

Vision: The Journal of Biomedical Research & Environmental Sciences (JBRES) is dedicated to advancing science and technology by providing a global platform for innovation, knowledge exchange, and collaboration. Our vision is to empower researchers and scientists worldwide, offering equal opportunities to share ideas, expand careers, and contribute to discoveries that shape a healthier, sustainable future for humanity.

JOURNAL OF BIOMEDICAL RESEARCH & ENVIRONMENTAL SCIENCES
issn: 2766-2276
SUBJECT AREA(S): BIOINFORMATICS | BIOLOGY
DOI: 10.37871

SHORT COMMENTARY

Synthetic Data Generation in Biomedical Research: Opportunities, Methods and Applications of Generative Adversarial Networks

Marco Parrillo*

Faculty of Engineering, University Telematics International UNINETTUNO, Rome, Italy

Abstract

The exponential growth of biomedical data, combined with increasingly stringent privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR), has created a significant bottleneck in the development of Artificial Intelligence (AI) and machine learning (ML) models for clinical and translational research. Synthetic data generation has emerged as a promising solution, enabling researchers to produce statistically realistic datasets that preserve the distributional properties of real patient data without exposing sensitive information. This commentary argues that GAN-based methods, and CTGAN in particular, represent a practical and scalable pathway for privacy-preserving biomedical AI: they outperform classical anonymisation techniques in downstream ML utility, handle mixed tabular data types that are pervasive in clinical records, and generalise across oncology, genomics, clinical trial simulation, and electronic health record synthesis. This paper reviews the theoretical foundations and practical applications of synthetic data generation methods in biomedical contexts, with a particular focus on Generative Adversarial Networks (GANs) and their tabular variant, the Conditional Tabular GAN (CTGAN). We further discuss emerging approaches including diffusion-based generative models and federated synthetic data generation. We examine key use cases, outline methodological considerations for validating the fidelity and utility of generated datasets, and address critical limitations including privacy leakage risks, model bias, and unresolved ethical and regulatory questions. Our analysis demonstrates that GAN-based approaches can produce synthetic biomedical records that support downstream ML tasks with accuracy comparable to models trained on real data, opening a viable pathway toward privacy-preserving, data-rich biomedical research.

Introduction

Access to large, well-annotated biomedical datasets is a prerequisite for training robust AI and ML models capable of supporting clinical decision support, disease prediction, genomic

***Corresponding author(s)**

Marco Parrillo, Faculty of Engineering, University Telematics International UNINETTUNO, Rome, Italy
Email: marcoparrillo@gmail.com
DOI: 10.37871/jbres2304

Submitted: 23 May 2026

Accepted: 02 June 2026

Published: 03 June 2026

Keywords

- Synthetic data
- Generative adversarial networks
- CTGAN
- Diffusion models
- Biomedical

How to cite this article: Parrillo M. Synthetic Data Generation in Biomedical Research: Opportunities, Methods and Applications of Generative Adversarial Networks. J Biomed Res Environ Sci. 2026 June 03; 7(6): 7. Doi: 10.37872/jbres2304

analysis, and drug discovery. Yet, the collection and sharing of patient-level health data remains constrained by ethical obligations, informed consent requirements, and regulatory frameworks. Regulations such as HIPAA in the United States and GDPR in Europe impose stringent restrictions on secondary use of personal health information, substantially limiting the availability of data for research purposes [1,2].

Traditional anonymisation methods, including de-identification and k-anonymity, reduce re-identification risk but often degrade the statistical utility of the resulting datasets, making them unsuitable for training complex models. Synthetic data generation offers an alternative paradigm: rather than releasing modified versions of real records, a generative model learns the underlying joint distribution of a dataset and then samples entirely new records from that distribution. If the generative process is sufficiently accurate, synthetic records preserve the statistical relationships present in the original data without being traceable to any real individual [3-5].

The central argument of this commentary is that GAN-based synthetic data generation, and CTGAN in particular, constitutes a mature and deployable solution to this access bottleneck. Compared with classical anonymisation methods, GAN-based approaches preserve downstream ML utility more faithfully while avoiding the need to modify or suppress data fields. Compared with oversampling methods such as SMOTE, they learn the global joint distribution of all variables rather than interpolating locally within individual feature dimensions. The scope of this review encompasses theoretical foundations, benchmark evaluation frameworks, and cross-domain biomedical applications, with particular attention to limitations and open challenges that practitioners must navigate before deploying synthetic data in clinical research workflows.

Among synthetic data methods, deep learning approaches, and Generative Adversarial Networks (GANs) in particular, have demonstrated superior performance in capturing complex, high-dimensional distributions. This paper examines GAN-based synthetic data generation within the biomedical domain, reviews existing applications, discusses evaluation frameworks for assessing the fidelity and downstream utility of generated datasets, and identifies critical limitations including privacy leakage, model bias, and unresolved ethical and regulatory questions.

Methods for Synthetic Biomedical Data Generation

Synthetic data can be generated through a range of approaches of increasing complexity. Classical statistical methods, such as parametric sampling from multivariate normal distributions, Gaussian mixture models (GMMs), and copula-based methods, offer tractable solutions when the data structure is well understood. These approaches model pairwise correlations and marginal distributions, enabling the generation of synthetic observations through Monte Carlo sampling [3].

Data augmentation techniques such as SMOTE (Synthetic Minority Oversampling Technique) and its extensions, including Borderline-SMOTE, ADASYN, and Safe-Level-SMOTE, address class imbalance by interpolating synthetic minority-class examples in feature space. While widely used in biomedical classification tasks, these methods do not learn the global joint distribution of the data and are thus limited to augmenting existing samples rather than generating novel, independent synthetic datasets [1].

Generative Adversarial Networks, introduced by Goodfellow, et al. [6] in 2014, represent a paradigm shift in generative modelling. A GAN consists of two neural networks trained in opposition: a generator G, which maps

random noise z drawn from a standard normal distribution to synthetic samples, and a discriminator D, which attempts to distinguish real from synthetic records. Through iterative adversarial training, G progressively learns to produce samples indistinguishable from real data, while D improves its ability to detect forgeries. At convergence, D assigns equal probability to real and synthetic samples, indicating that G has effectively approximated the true data distribution.

Tabular biomedical data, however, poses specific challenges for standard GAN architectures. Unlike image pixels, tabular features encompass a mixture of continuous and categorical variables, often exhibiting multimodal distributions, severe class imbalance, and complex inter-variable dependencies. The Conditional Tabular GAN (CTGAN), proposed by Xu, et al. [4] directly addresses these challenges through two key innovations. First, mode-specific normalisation fits a Variational Gaussian Mixture model (VGM) to each continuous column, estimating the number of distributional modes and normalising values within each mode; this prevents the gradient vanishing problem that arises from applying standard min-max normalisation to non-Gaussian biomedical variables such as laboratory values, vital signs, or genomic expression levels. Second, a conditional generator combined with a training-by-sampling procedure ensures that rare categorical values, for instance uncommon diagnostic codes or infrequent drug combinations, are visited uniformly during training, preventing mode collapse on dominant categories.

A significant recent development in generative modelling is the emergence of diffusion-based models, which operate by progressively corrupting training data with Gaussian noise and then learning the reverse denoising process. Denoising Diffusion Probabilistic Models (DDPMs), introduced by Ho, et al. [7] demonstrated state-of-the-art performance in image and audio generation. Their adaptation to tabular biomedical data, notably through TabDDPM proposed by Kotelnikov, et al. [8], addresses several persistent limitations of GANs including training instability, mode collapse, and the absence of a stable likelihood objective. TabDDPM formulates denoising steps separately for continuous and categorical feature types, and benchmark evaluations across multiple tabular datasets have shown it to match or exceed CTGAN in distributional fidelity and ML utility metrics.

Notwithstanding these advances, diffusion models require substantially greater computational resources and longer inference times than GAN-based approaches, which may limit their use in resource-constrained clinical settings.

Federated synthetic data generation represents another active frontier, combining the strengths of federated learning and synthetic data. In federated learning, model parameters are trained across distributed hospital sites without centralising patient records, thereby preserving institutional data governance [9]. When federated training is applied to a generative model, the resulting synthesiser can capture population-level distributions spanning multiple sites, producing synthetic datasets that reflect broader demographic and clinical diversity than any single institution could provide. Recent work has demonstrated the feasibility of federated GAN training on EHR data, with a case study in acute myeloid leukaemia demonstrating competitive data fidelity under horizontal federation across simulated institutional nodes [10].

Biomedical Applications

Table 1 provides a structured summary of the principal biomedical application domains reviewed in this section, including

Table 1: Summary of GAN-based synthetic data applications in biomedical research, with key advantages and domain-specific limitations.

Application Area	GAN Approach	Key Advantages	Limitations
Oncology and Cancer Research	CTGAN, conditional GAN	Enables cross-institutional data sharing; supports survival analysis with comparable statistics	Small cohort sizes limit mode diversity; rare tumour subtypes may be underrepresented
Genomics and Personalised Medicine	Attention-based GAN, graph-prior GAN	Preserves population-level linkage disequilibrium; reduces re-identification risk from genetic profiles	High dimensionality increases training instability; validation of genomic fidelity remains challenging
Clinical Trial Simulation	CTGAN, conditional GAN	Accelerates feasibility studies; optimises randomisation strategies without patient exposure	Synthetic cohorts may not capture rare adverse events; regulatory acceptance still evolving
Electronic Health Record (EHR) Synthesis	Recurrent GAN, CTGAN, TabDDPM [8]	Generates longitudinal patient trajectories; enables NLP tool development where real data are scarce	Temporal dependencies are difficult to model faithfully; long sequences increase computational

the GAN approach typically employed, the key advantages, and the domain-specific limitations (Table 1).

Oncology and Cancer Research

In oncology research, sharing patient-level cancer registries across institutions is impeded by data governance agreements and the small sample sizes that characterize rare malignancies. Public Health England has released synthetic cancer datasets to enable researchers to develop and test predictive models, validate study designs, and estimate statistical power before undertaking costly data-access applications. Synthetic records generated from real tumour registries have been shown to support survival analysis and treatment-response modelling with statistical properties comparable to the original cohorts [1,5].

Genomics and Personalized Medicine

In genomics and personalised medicine, the high dimensionality of genomic data, including single-nucleotide polymorphism arrays and RNA-seq expression matrices, exacerbates privacy risks because individuals can be re-identified from their genetic profiles even after de-identification. GAN architectures adapted for high-dimensional sparse data, including variants that incorporate attention mechanisms and graph-based priors, have been applied to generate synthetic genomic datasets for association studies and polygenic risk score development, reducing the exposure of real patient genomes while maintaining population-level linkage disequilibrium patterns. Recent reviews have further highlighted the specific promise of conditional generative models for enabling diverse and equitable representation in precision medicine cohorts [11].

Clinical Trial Simulation

In clinical trial simulation, synthetic cohorts derived from historical trial data allow sponsors to perform in-silico feasibility studies, test inclusion and exclusion criteria, and optimise randomisation strategies before patient recruitment, thereby reducing trial duration and cost. Regulatory agencies, including the European Medicines Agency, have begun to encourage pharmaceutical companies to make trial data more widely accessible; synthetic datasets represent a mechanism for satisfying this requirement without compromising patient confidentiality [1,2].

Electronic Health Record Synthesis

Electronic Health Record (EHR) synthesis is among the most impactful biomedical applications. GAN-based models trained on EHR data can generate realistic longitudinal patient trajectories, including sequences of diagnoses, medications, and laboratory results, enabling the development of clinical NLP tools, early warning systems, and readmission prediction models in institutions that lack sufficient real-world data volumes. Tools such as Synthea, a rule-based synthetic patient generator, complement GAN-based approaches by producing clinically plausible record structures, and hybrid pipelines combining rule-based priors with deep generative models represent a promising avenue for future work [12].

Validation Framework

Validating synthetic biomedical data requires a structured evaluation framework addressing two orthogonal dimensions: fidelity, meaning statistical similarity to real data, and utility, meaning downstream ML performance [3,4].

Fidelity is assessed through univariate distributional comparisons using the Kullback-Leibler (KL) divergence and the Hellinger distance. For a pair of probability distributions P and Q, the Hellinger distance H(P,Q) is bounded in [0, 1], where 0 indicates identical distributions and 1 indicates complete divergence; its bounded nature makes it more interpretable than unbounded divergence measures. The Kolmogorov-Smirnov test provides a non-parametric comparison of empirical cumulative distributions for continuous variables. Multivariate fidelity can be assessed through correlation matrix comparison and propensity score modelling, in which a discriminator classifier is trained to separate real from synthetic records; accuracy near 0.5 indicates indistinguishable datasets [1].

Utility is evaluated through the Train on Synthetic, Test on Real (TSTR) protocol: a supervised classifier is trained exclusively on synthetic data and evaluated on a held-out real test set; its performance is then compared against a classifier trained on real data under the Train on Real, Test on Real (TRTR) benchmark. A small performance gap between TSTR and TRTR indicates that synthetic data preserves the discriminative structure of the original dataset, confirming its suitability for ML model development. Experimental results with CTGAN applied to tabular biomedical datasets have demonstrated TSTR accuracy within 2 to 5 percentage points of TRTR benchmarks across multiple classification tasks [4].

Limitations and Risks of Synthetic Data Generation

Despite its promise, synthetic biomedical data generation carries inherent limitations and risks that practitioners must carefully consider before deployment.

Privacy leakage is perhaps the most critical risk. Although generative models do not store individual records, they can inadvertently memorise rare or outlying training examples, particularly when the training dataset is small. Membership inference attacks have demonstrated that adversaries can determine with non-trivial accuracy whether a specific record was used to train a given generative model, potentially re-identifying individuals from the synthetic output [13]. Differential privacy (DP) mechanisms, such as DP-SGD, can provide formal privacy guarantees by injecting calibrated noise into the training gradient, but do so at the cost of reduced model capacity and lower fidelity synthetic data, creating a fundamental privacy-utility trade-off [14].

Model bias represents a second major concern. Generative models learn the statistical regularities of their training data, including embedded biases related to demographic disparities in healthcare access, historical

underrepresentation of minority populations in clinical studies, and systematic measurement differences across clinical sites. Synthetic data generated from biased sources will propagate and may amplify those biases in downstream models, with potential consequences for clinical equity if the resulting AI tools are deployed differentially across patient populations [15,16].

Additional limitations include the challenge of validating rare clinical events, which by definition have low frequency in training data and may be poorly captured by generative models; the difficulty of faithfully reproducing complex longitudinal temporal dependencies in EHR sequences; and the computational expense of training large GAN or diffusion models on high-dimensional biomedical data.

Ethical Considerations and Regulatory Acceptance

The deployment of synthetic biomedical data in research and clinical AI development raises ethical questions that extend beyond technical privacy guarantees. Patients who have consented to the use of their records for specific research purposes may not have anticipated that their data would be used to train a generative model, or that the resulting synthetic data might be distributed to third parties without further consent. Transparent communication with patients and research ethics boards about the intended use of generative modelling is therefore an important governance consideration, even when strict legal obligations under GDPR or HIPAA may not formally require it [5].

Regulatory frameworks governing the acceptance of synthetic data in drug approval and medical device submissions are still evolving. The European Medicines Agency has encouraged the use of synthetic data for sharing clinical trial results while acknowledging that the conditions under which synthetic data may substitute for real patient data in regulatory submissions remain to be fully defined. The United States Food and Drug Administration has similarly issued draft guidance on AI and ML-based software as a medical device, but has not yet established explicit standards for synthetic training data validation. Researchers and developers are therefore advised to engage proactively with regulatory bodies and to document synthetic data generation and validation protocols transparently in study registrations and publications [1,4].

Conclusions and Future Directions

Synthetic data generation, particularly through GAN-based approaches such as CTGAN, represents a compelling solution to one of the most pressing bottlenecks in biomedical AI: the scarcity of accessible, high-quality, privacy-compliant training data. By learning and replicating the joint statistical structure of real patient records, synthetic datasets enable institutions to share data across organizational boundaries, develop and validate predictive models, and simulate clinical scenarios without exposing individuals to re-identification risk. The comparative analysis of CTGAN and emerging diffusion-based methods such as TabDDPM [8] suggests that no single generative paradigm is universally optimal; the choice of method should be guided by the dimensionality of the target data, the severity of class imbalance, available computational resources, and the required balance between fidelity and privacy. Future research directions include the integration of differential privacy guarantees into GAN and diffusion model training pipelines to provide formal, auditable privacy bounds; the development of federated synthetic data generation protocols that synthesise data collaboratively across multiple hospital sites without centralising records [10]; and the extension of CTGAN-style architectures to longitudinal, time-series EHR data incorporating temporal dependencies [11]. Standardised benchmarking suites for biomedical synthetic data, encompassing

both fidelity metrics and bias audits, would substantially accelerate progress in the field [16]. Regulatory guidance on the acceptance of synthetic data in drug approval and clinical research submissions will also be essential for broader clinical translation, and early and transparent engagement with regulatory bodies is strongly encouraged [17,18].

References

1. Emam K, Mosquera L, Hoptroff R. Practical synthetic data generation. Sebastopol (CA): O'Reilly Media; 2020.
2. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health. 2023;2(1):e0000082. doi:10.1371/journal.pdig.0000082.
3. Figueira A, Vaz B. Survey on synthetic data generation, evaluation methods and GANs. Mathematics. 2022;10(15):2733. doi:10.3390/math10152733.
4. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems. 2019;32.
5. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med. 2023;6:186. doi:10.1038/s41746-023-00927-3.
6. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. In: Advances in Neural Information Processing Systems. 2014;27.
7. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. 2020;33:6840-51.
8. Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. TabDDPM: modelling tabular data with diffusion models. In: Proceedings of the 40th International Conference on Machine Learning (ICML). 2023;202:17564-79.
9. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). 2017;54:1273-82.
10. Isasa I, Catalina M, Epelde G, Aginako N, Beristain A. Synthetic tabular data generation under horizontal federated learning environments in acute myeloid leukemia: case-based simulation study. JMIR Med Inform. 2025;13:e74116.
11. Liu K, Altman RB. Conditional generative models for synthetic tabular data: applications for precision medicine and diverse representations. Annu Rev Biomed Data Sci. 2025;8:21-49. doi:10.1146/annurev-biodatasci-103123-094844.
12. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230-8.
13. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, et al. Synthetic data: what, why and how? arXiv [Preprint]. 2022. Available from: arXiv:2205.03257.
14. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS). New York (NY): ACM; 2016. p. 308-18.
15. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123-44.
16. Shahul Hameed MA, Qureshi AM, Kaushik A. Bias mitigation via synthetic data generation: a review. Electronics. 2024;13:3909. doi:10.3390/electronics13193909.
17. Esteban C, Hyland SL, Ratsch G. Real-valued (medical) time series generation with recurrent conditional GANs. arXiv [Preprint]. 2017. Available from: arXiv:1706.02633.
18. Nikolenko SI. Synthetic data outside computer vision. In: Nikolenko SI. Synthetic data for deep learning. Cham: Springer; 2021. p. 217-26.