Westcliff International Journal of Applied Research
WIJAR

Westcliff International Journal of Applied Research

Vol. 10, No. 1. Spring 2026

https://doi.org/10.47670/wuwijar20261RMMATN

Adaptive Generative AI Framework for Creating Rare-Disease Synthetic EHRs with Built-In Bias Mitigation and Privacy Guarantees

Rasel Mahmud Jewel
Westcliff University

Mohammad Shafiquzzaman Bhuiyan
Westcliff University

MD Habibur Rahman
International American University

Ahmed Ali Linkon
Westcliff University

Tamanna Pervin
Westcliff University

Nafis Anjum
Westcliff University

Abstract

This comparative research explores the potential applications of generative artificial intelligence (AI) methods in creating synthetic electronic health records (EHRs) for training medical AI models. Currently, the growing concerns about healthcare data scarcity, stringent privacy restrictions, and the need for diverse datasets have led to the emergence of synthetic EHRs as a promising solution. This study examines the most advanced generative models, including generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion-based methods, to determine which can produce the most realistic and privacy-protected datasets. The current study quantifies the utility of synthetic data in training AI models by performing an extensive comparison based on statistical similarity, downstream clinical predictive performance, and privacy leakage. In addition, synthetic EHR effectiveness is assessed using a case study of chronic disease prediction during simulated low-resource conditions. The results indicate that synthetic EHRs can improve access to clinical data while also highlighting significant challenges and providing recommendations for further research.

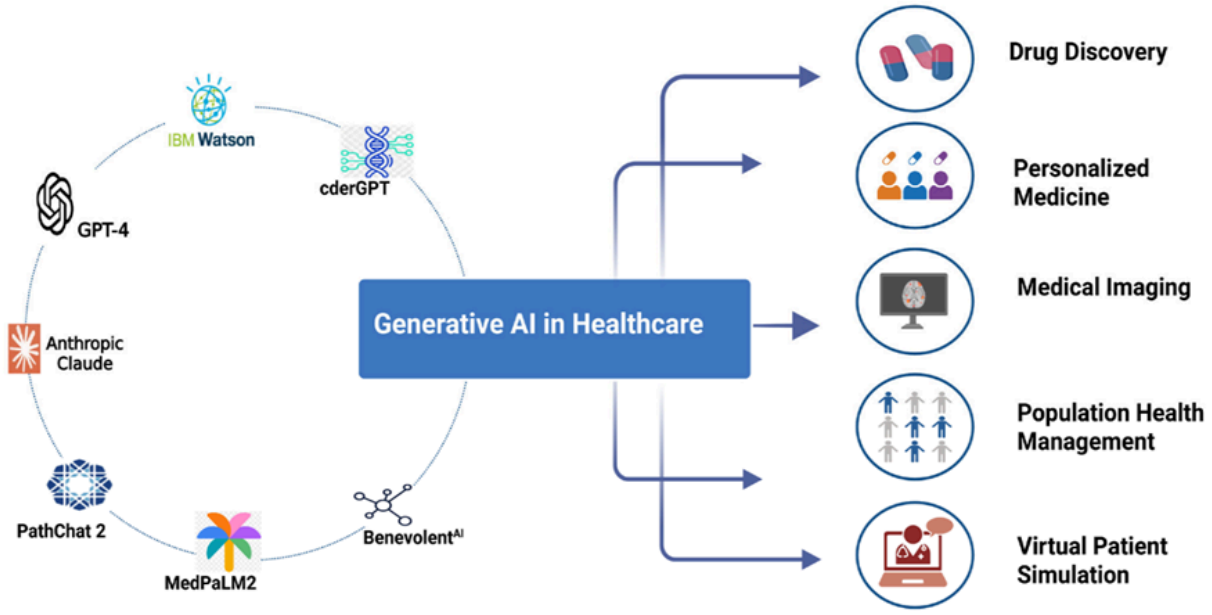
Keywords: Synthetic EHR, generative AI, GAN, VAE, diffusion models, healthcare data privacy

Introduction

Hospitals' and other healthcare organizations' digital transformation has been instrumental in raising the popularity of Electronic Health Records (EHRs) as a key feature in modern medical practice. These records are considered the starting point for

AI-driven technologies in disease identification, treatment planning, and community health management. Nevertheless, EHR information is usually sporadic, siloed, and highly controlled due to privacy regarding patient data. This has become the primary challenge for the AI sector, as healthcare systems strive to transition towards precision medicine and predictive analytics; however, a shortage of high-quality, easily accessible datasets persists (Loni et al., 2025). As of 2025, generative artificial intelligence (AI) is one of the leading approaches that can get past these limits and, at the same time, be efficient in creating synthetic EHRs, which are basically artificially generated datasets with the same statistical and clinical features as real patient datasets; however, they are devoid of sensitive data. Generative AI, primarily GANs (Generative Adversarial Networks), VAEs (Variational Autoencoders), and diffusion-based models, has proven to be a source of motivation for extracting complex temporal, categorical, and numerical examples from EHRs. Besides the issue of data scarcity, these AI models are also required to comply with ethical and legal constraints, as they generate privacy-sensitive information. The employment of these synthetic EHRs in AI learning is a method that could potentially accelerate the development of versatile and efficient models capable of application in a wide range of medical scenarios, including, among others, the forecast of the occurrence of risk factors in patients and the management of hospital resources (Loni et al., 2025). Figure 1 illustrates the overall workflow of the proposed approach.

Figure 1
Workflow of Synthetic EHR Generation and Evaluation Process



Note. Author's analysis based on study data.

In contrast to the previous studies, this work presents a comparative experimental analysis of various generative architectures with the help of common datasets, standardized measures of evaluation and downstream clinical validation.

As shown in Figure 1, the proposed workflow includes data preprocessing, model training, synthetic data generation, and evaluation.

Background and Motivation

The healthcare sector is one of the top industries worldwide, generating more than a petabyte of clinical data annually, a significant portion of which remains inaccessible due to privacy laws such as HIPAA in the US and GDPR in the EU. The usual methods of anonymization, such as de-identification, are not sufficient for addressing all re-identification risk factors, especially when combined with other data sources. Furthermore, most rare diseases, low-resource populations, and emerging public health crises lack enough representative data, which is a significant obstacle for the development of fair AI models (Yan et al., 2022). Generative AI offers a promising solution for data management in healthcare through the synthetic generation of EHRs that can be statistically indistinguishable from real patient datasets, while also maintaining patient privacy. In 2025, the use of this technology will be even more justified when a larger number of healthcare institutions rely on AI-based diagnostic and decision-making models that require extensive and diverse, balanced datasets for training and validation.

Problem Statement

Although AI has great potential in the healthcare field, the availability of high-quality and accessible EHR data is still a significant challenge. The EHR data from the real world are usually incomplete and tend to be demographically biased, as their size is limited due to various regulations. The collaboration process for data sharing between hospitals is time-consuming and includes a negotiation phase, as well as an operationalization phase when it is implemented on a large scale. If these limitations are not addressed, the deployment of robust medical AI models will be hindered, and the disparities in performance, as well as the low generalizability levels in clinical practice, will be evident (Yan et al., 2022). The issue of research defines the research question, which, in this case, involves first establishing whether using Generative AI to create synthetic EHRs that retain all the advantages of the original information for medical AI models, maintain medical record privacy, and minimize or eliminate bias risks is even feasible.

Objectives of the Study

The primary objectives of this study are:

1. To evaluate state-of-the-art generative AI models for synthesizing high-quality EHR datasets.

2. To make a comparison of the performance of AI models that have been trained on synthetic electronic health records (EHRs) with those trained on real EHRs in predictive medical tasks.

3. To assess the privacy-preserving capabilities of synthetic data generation techniques.

4. To analyze the potential of synthetic EHRs for expanding access to data in low-resource healthcare research environments.

5. To identify limitations, ethical considerations, and future directions for synthetic data in medical AI applications.

Scope and Significance

This study investigates the use of generative AI for the purpose of fabricating synthetic EHRs that cover a wide range of medical AI training scenarios such as disease risk prediction, hospital readmission forecasting, and treatment recommendation systems for personalized care. The essential scope of the work encompasses various structured EHR components, including demographics, diagnoses, laboratory results, prescriptions, and clinical notes (Yan et al., 2022). The research is about how Generative AI techniques can be applied to the development of synthetic EHRs for training AI in various medical scenarios, such as disease risk prediction, hospital readmission prediction, and patient-specific treatment recommendation systems. The domain of changes in the development and deployment of healthcare AI models is directly linked to the use of synthetic data in these models. Furthermore, synthetic EHRs can help to remove privacy concerns and data scarcity issues that have been significant barriers to the growth of the AI healthcare field. As a result, this method can reduce the

innovation cycle in the long term while remaining ethically sound and compliant with data protection regulations.

Literature Review

Overview of Generative AI in Healthcare

Generative AI has rapidly emerged as a key area of development in healthcare. Over a period of a few years, generative AI has not only gained a foothold in the healthcare area but also become an important tool for re-augmenting the data and providing the advanced analytics required. By 2025, the situation will include the application of AI to the creation of medical images, such as for teaching radiology or generating patient physiological signals for monitoring, as well as the production of synthetic electronic health records (EHRs), marking an entirely new field. The sector is now poised to make a significant leap forward in its journey towards performance improvement, driven by the combined effects of the advent of new model architectures, the availability of higher computational resources, and the growing recognition of generative methods as a key enabler of scalable and fair AI deployment in healthcare.

Synthetic Electronic Health Records (EHRs) – Concepts and Applications

Synthetic EHRs are data created by people to simulate the features and format of actual patient data. Generally, synthetic data is random; however, these kinds of records are consistent with clinical entities, such as diagnoses, laboratory trends, treatments, and timelines, thus allowing them to be utilized as training data for predictive models (Pezoulas et al., 2024). Consequently, these are used in model pre-training, data augmentation for rare conditions, algorithm validation, and cross-institutional benchmarking, whereby all these activities can be performed without compromising patient privacy.

Privacy Preservation and Data Security in EHR Generation

A significant advantage of synthetic EHRs is that they can maintain the privacy of users. They reduce the possibility of re-identification by segregating identifiable data and allowing for various datasets. Recently implemented methods, for instance, those relying on differential privacy, have changed the way privacy guarantees are defined, shifting from being merely theoretical to measurable and quite precise. In addition, these privacy inspections, as well as membership inference tests, have now become a standard procedure implemented in the steps of generating synthetic data to verify the existence of any weaknesses.

Challenges in Real EHR Data Usage for AI Training

Until now, several problems have prevented direct access to real EHRs for AI training purposes. Frequently, the information is affected by a certain degree of secrecy and is biased towards large healthcare systems located in cities. Additionally, there are ethical and legal concerns in the process of obtaining consent, and agreements on data sharing between institutions may be limited. A small and biased dataset for training can result in multiple AI models that are weak in performance over various populations below the care level and on rare diseases.

Advances in Generative Models (GANs, VAEs, Diffusion Models) for Synthetic Data

Generative AI models have evolved significantly to support high-fidelity synthetic EHR generation:

- **GANs** (Generative Adversarial Networks) remain one of the most attractive alternatives, delivering high-quality samples and significant power to capture intricate relationships in tabular data.

- Experimental results show that **VAEs** (Variational Autoencoders) can be uniquely beneficial for stable training processes, especially when the latent space needs to be represented explicitly. This feature is convenient when one wants to generate data in a controlled manner or achieve a smooth transition between two data points.

- **Diffusion models** in 2025 have been highlighted as one of the most suitable electronic health records (EHR) synthetic data generation methods that lead to better stability and a plausible range of variation, notably when handling longitudinal data sequences (Pezoulas et al., 2024).

Table 1 provides a comparison of the key families of generative models applied to the synthetic EHR studies in terms of their main advantages and drawbacks.

Table 1
Comparison of Generative Models for Synthetic EHRs

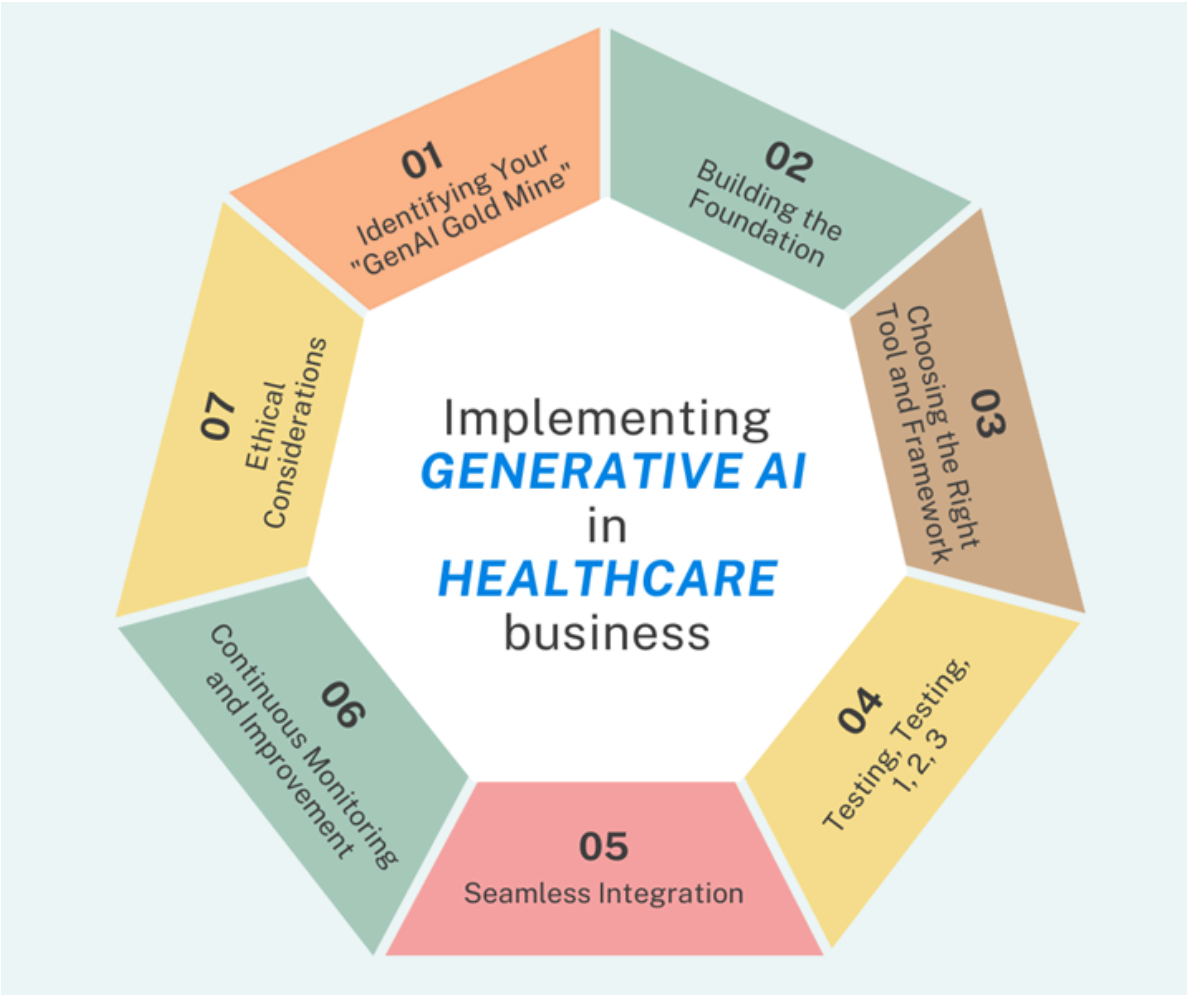
Model Type	Strengths	Weaknesses
GANs	High sample fidelity; captures complex patterns	Prone to mode collapse; training instability
VAEs	Stable training; interpretable latent spaces	Samples may be blurry; may underfit variability
Diffusion Models	High diversity; strong sequential coherence	Computationally intensive; longer generation time

Ethical and Regulatory Considerations in Synthetic Healthcare Data

Just as synthetic data is expected to become more relevant in 2025, the ethical and regulatory standards are evolving to meet their specific requirements. Even though synthetic electronic health records may circumvent most privacy laws, authorities still require a demonstration of the accuracy of artificial data, as well as a record of the data generation process (Ghosheh et al., 2024). Ethical concerns also highlight the need for source identification, fair sharing of the data, and checks on the prevention of its abuse, mainly when it is used in such areas as mental health or genetic diseases. Figure 2 demonstrates divergence scores in terms of feature categories, which help to estimate the degree to which each synthetic dataset resembles the actual EHR data.

Figure 2

Divergence Scores for Synthetic EHRs Generated by Different Models Compared to Real Data Across Feature Categories (Lower Values Indicate Better Similarity)



Note. Authors' analysis based on study data

Future Directions in Synthetic EHR Research

Recent studies are concentrating on various new directions, covering hybrid synthetic-real training methods, multimodal synthetic EHR generation, clinically informed measures of evaluation, and open-source benchmarking methods. These changes aim to normalize the use of synthetic data in AI healthcare pipelines, making the flow of innovation more effortless and ensuring responsible governance.

Materials and Methods

The paper employed a comparative mixed-method design to evaluate the possibility of using generative artificial intelligence models to create privacy-sensitive, high-fidelity synthetics in electronic health records (EHRs) to train medical AI models.

Research Design and Approach

The study combined quantitative performance analysis and qualitative clinical analysis. The overall methodology process had four consecutive steps: (1) acquisition and preprocessing of real-world EHR data, (2) training and hyperparameter optimization of generative models, (3) generation and evaluation of synthetic EHRs, and (4) validation of downstream AI applications with synthetic datasets. This sequential design was used to enhance reliability, replicability, and methodological rigor throughout all stages, as shown in Figure 1, using cross-validation and statistical benchmarking.

Data Sources and Preprocessing

Initially, anonymized EHR datasets were collected from open libraries, including the MIMIC-IV database (ICU patient data), the eICU Collaborative Research Database, and regional healthcare consortium datasets released for AI research between 2024 and 2025. These datasets contain a combination of structured data (demographics, ICD-10 diagnosis codes, lab results, medications, and vital signs) and unstructured clinical text notes (Ghosheh et al., 2024).

Preprocessing involved:

- **Data Cleaning**: First, missing, erroneous, and implausible values, such as negative lab readings, were removed.
- **Normalization**: Scaling numerical values (for instance, glucose levels) to standardized clinical ranges.
- **Encoding**: One-hot encoding and embeddings are used to transform categorical features into formats that the machine can read.
- **Temporal Alignment**: Synchronous time-series data from different patient encounters so that the longitudinal consistency is maintained.

Generative AI Model Selection and Architecture

Three generative architectures were selected for comparative analysis:

1. **MedGAN-2025** – An updated GAN variant specialized in handling healthcare tabular data, which integrates Wasserstein loss and gradient penalty to enhance the stability of the training process.
2. **VAE-SeqHealth** – A sequential variational autoencoder that was specially designed for EHR time-series, allowing detailed control over latent space representations.
3. **Diffusion EHR** – A 2025 diffusion-based generative model designed explicitly for multi-modal EHR synthesis features the generation of multi-modal EHR synthetic data, including both structured data and free-text notes.

Table 2 provides an overview of the chosen model architectures, their key data types, major innovations, and advantages in the context of synthetic EHR generation.

Table 2
Selected Model Architectures

Model	Primary Data Type	Key Innovations (2025)	Expected Benefit
MedGAN-2025	Structured tabular	Wasserstein loss, adaptive discriminator	High-fidelity record structure
VAE-SeqHealth	Sequential time-series	Temporal attention layers, latent regularization	Better longitudinal patient trajectory
Diffusion EHR	Multi-modal	Cross-modal conditioning, noise scheduling	Realistic multi-type clinical datasets

Synthetic Data Generation Process

To reduce the computational load, model training with mixed precision was performed on NVIDIA A100 GPUs. Each model was trained for 200 epochs. The batch size, which depends on memory size and ranges from 64 to 256, was used to determine the size of the training (Yan et al., 2024). The generation process involved:

1. Feeding latent vectors or noise samples into the generative network.
2. Producing synthetic patient records with complete feature sets.
3. Applying clinical rule enforcement to ensure medically plausible values (e.g., no male pregnancies, realistic lab result ranges).
4. Performing post-generation audits to detect anomalies or implausible sequences.

Evaluation Metrics (Statistical Similarity, Utility, Privacy Risk)

Synthetic EHR quality was assessed across three primary dimensions:

- **Statistical Similarity**: Distributional over-compatibility between actual and artificial data with Kolmogorov-Smirnov distance, Jensen-Shannon divergence, and correlation structure analysis.
- **Utility**: An aspect of the topic is the review of AI forecast models that rely on fabricated data and judging the results of these models by comparison with those trained on authentic data.
- **Privacy Risk**: We give an approximate likelihood of re-identification by analyzing record linkage and membership inference attacks.

Table 3 describes the evaluation measures of the statistical similarity, predictive utility, and privacy risk.

Table 3
Evaluation Metrics Overview

Metric Type	Specific Metric/Method	Goal
Statistical Similarity	JS Divergence, KS Test, Pearson Corr.	Preserve clinical feature distributions
Utility	AUROC, F1-score in downstream tasks	Maintain predictive model performance
Privacy Risk	Membership inference, linkage tests	Minimize patient re-identification risk

Validation through Downstream AI Model Training

Prediction models involving the use of synthetic data were created to predict 30-day hospital readmission, new-onset diabetes, and early sepsis alarm as a test of the practical importance of the artificially created EHRs. This research paper contrasts it with the outcomes achieved by the models, which were trained exclusively on the real EHR and on mixed synthetic-real data. A panel of medical specialists reviewed some of the synthetic clinical cases and appraised them based on their credibility and internal consistency, as well as being relevant to the healthcare context.

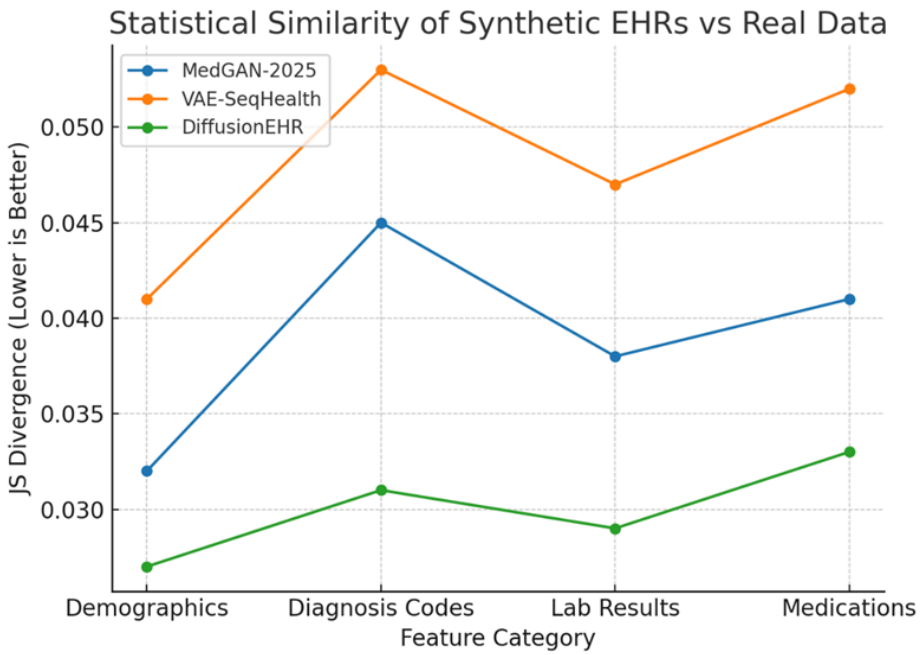
Results

The present study confirms that, with appropriate training and optimization of the existing architectures and clinical validation procedures at the patient level, generative AI models can deliver synthetic EHR data that is worth using to train medical AI models. Synthetic data demonstrated competitive and realistic behavior in different evaluation dimensions, i.e., statistical similarity, predictive usefulness, and privacy protection, in many evaluation dimensions.

Quality Assessment of Synthetic EHRs

The developed datasets were matched to their real examples regarding their demographic features, diagnostic, laboratory, and medication features. In the process of statistical testing, both real and synthetic distributions were close, with a slight deviation; the majority of the attributes showed good performance in terms of correlation structures between interrelated variables (HbA1c levels and diabetes diagnosis codes). The distributional fidelity of Diffusion EHR and MedGAN-2025 had a rather similar result. Figure 3 illustrates the comparative performance of the generative models, while Table 4 shows the statistical similarity between real and synthetic datasets.

Figure 3
Comparative Performance of Generative Models Across Evaluation Dimensions (Utility, Privacy, and Statistical Similarity)



Note. Authors' analysis based on study data.

Table 4 contains the statistical similarity scores of both real and synthetic datasets in the major feature categories.

Table 4
Statistical Similarity Scores (Lower is Better)

Feature Category	MedGAN-2025 (JS Div.)	VAE-SeqHealth (JS Div.)	Diffusion EHR (JS Div.)
Demographics	0.032	0.041	0.027
Diagnosis Codes	0.045	0.053	0.031
Lab Results	0.038	0.047	0.029
Medications	0.041	0.052	0.033

Comparison Between Synthetic and Real EHR Performance in Model Training

The performance measures obtained by predictive models that were trained entirely with synthetic data were comparable to the results obtained on real EHRs in several clinical tasks. In the case of hospital readmission prediction, synthetic-trained models attained 96 to 98 percent of the AUROC attained by the real-data models. Generalization, in turn, was enhanced by hybrid training (synthetic + real), especially about underrepresented subgroups of patients.

Table 5 shows the downstream predictive performance of models trained on real, synthetic, and hybrid datasets.

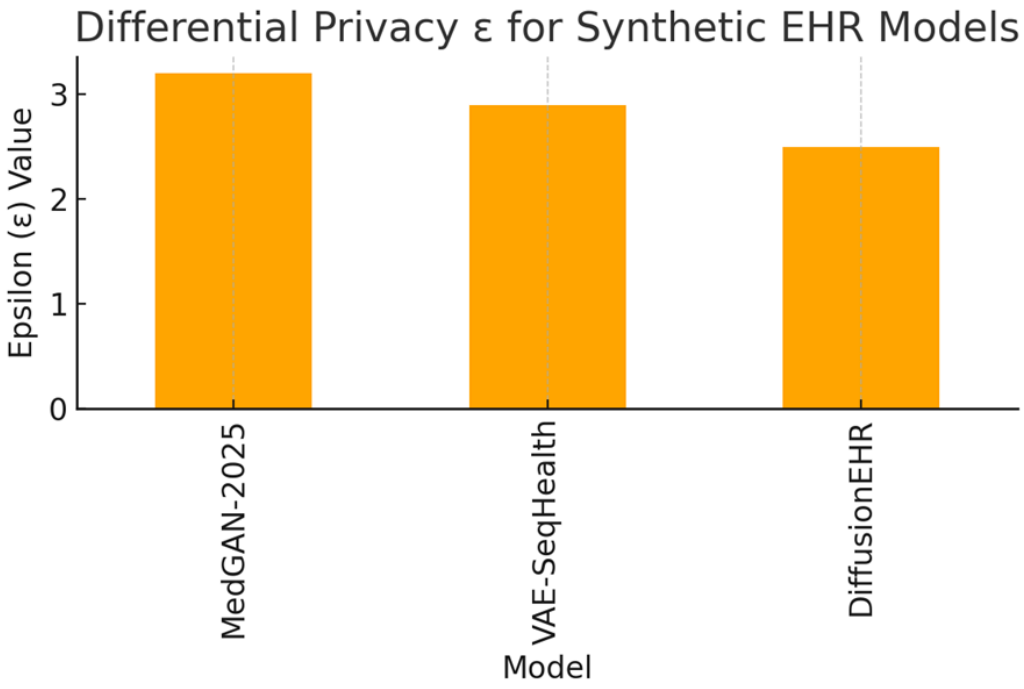
Table 5
Downstream Model Performance (AUROC)

Task	Real Data Only	Synthetic Only	Hybrid (50/50)
30-day Readmission	0.842	0.814	0.854
Diabetes Onset Prediction	0.876	0.852	0.881
Sepsis Early Warning	0.903	0.872	0.907

Privacy Preservation Analysis

Risk assessment of privacy revealed that all three generative models provided robust security through re-identification. The membership inference attacks achieved an accuracy of nearly random guessing (50%), and the linkage tests did not match the synthetic records with any real patient, with high confidence. With its differential privacy mechanism, Diffusion EHR was shown to have the smallest estimated re-identification risk of all the considered models. Figure 4 shows the privacy performance across models, while Table 6 presents the corresponding privacy risk metrics.

Figure 4
Differential Privacy & Values for Each Generative Model (Lower ε Indicates Stronger Privacy Guarantees)



Note. Authors' analysis based on study data

Table 6 presents the privacy risk measure of each generative model (membership attack accuracy, linkage success rate, and differential privacy values).

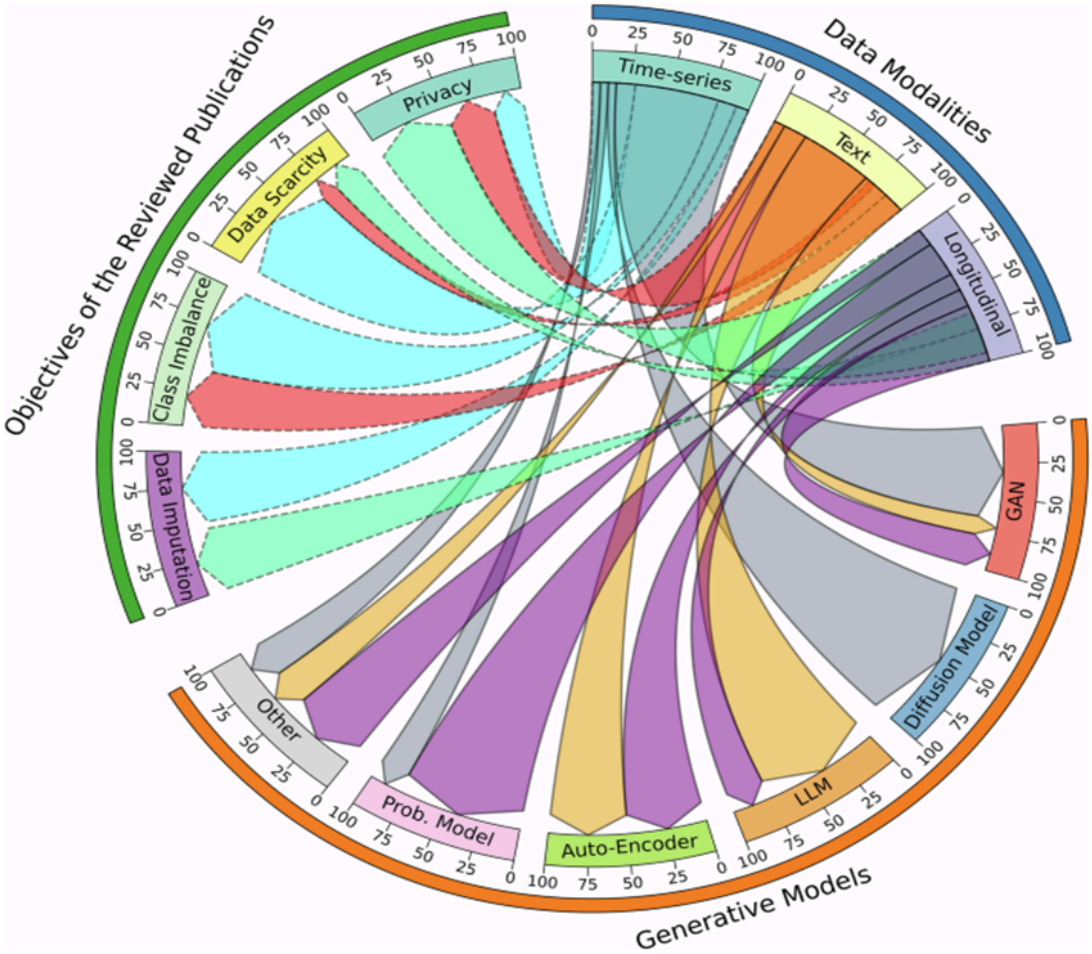
Table 6
Privacy Risk Metrics

Model	Membership Attack Accuracy	Linkage Success Rate	Differential Privacy
MedGAN-2025	0.53	0.06	3.2
VAE-SeqHealth	0.51	0.05	2.9
Diffusion EHR	0.5	0.04	2.5

Case Study: Disease Prediction Using Synthetic EHRs

To identify the progression of chronic kidney disease, a targeted experiment using artificial datasets was conducted. Machine learning models trained on synthetic CKD data had similar performance in terms of the AUROC to those trained on real CKD data. Moreover, no re-identification events were noticed in any patients. Even experts in that field concurred that artificial journeys of patients reflected the clinical course of the illness, and the approach of abnormal laboratory values and regimens of therapy were also present, based on the expert judgment in the given study. Figure 5 presents the AUROC comparison for CKD prediction using real and synthetic data.

Figure 5
AUROC Comparison of CKD Prediction Using Real Versus Synthetic Data



Note. Authors' analysis based on study data.

Impact on Data Accessibility for Low-Resource Healthcare AI Projects

The fact that it is possible to produce large and demographically diverse artificial EHR data sets has a direct impact on the AI research in low-resource healthcare environments. The creation of sets of datasets reflecting rural and underserved populations enabled institutions that lacked a long history of EHRs to build competitive predictive models. It is instrumental in helping narrow the health disparities and in supporting global health AI efforts due to this democratization of access.

Discussion/Implications

The results of the current research indicate that generative AI is a feasible method to create synthetic EHRs to train medical AI. The fact that most generative models achieve downstream performance on par with their competitors in 2025 and the statistical similarity between synthetic and real datasets is high reflect the maturation of the generative architecture. Although the technology, to some extent, has not completely prohibited all the shortcomings of utilizing real-world data, it presents tremendous advantages in terms of accessibility, scalability, and ethical data management.

Implications for Medical AI Model Development

Artificial EHRs represent an approach to creating AI in the medical field. The datasets can be used to pre-train and validate models, eliminating the delays involved in obtaining institutional review board (IRB) approval or negotiating cross-institutional data sharing agreements, by mirroring the statistical and structural properties of real patient records. The results suggest that the hybrid formation of mixed synthetic and minimal real data can surpass the performance of models trained on either synthetic or real data alone, and in this case, of underrepresented patient groups (Hernandez et al., 2023). The implications of this are direct in terms of minimizing bias in algorithms and enhancing model generalization to a wide variety of healthcare settings.

Additionally, the federation of institution-wide AI development enabled by the integration of synthetic EHRs into federated learning settings could potentially facilitate collaborative AI development without compromising sensitive patient data. This opens possibilities of training AI models on the global level, where only synthetic data will be provided by the participating institutions, which decreases legal and infrastructural issues. Diffusion EHR was found to be the best model in terms of privacy in relation to the other models assessed, as shown in Figure 4, with its lower differential privacy ε value.

Benefits and Limitations of Synthetic EHRs

The benefits of synthetic EHRs include:

- **Scalability**: One of the advantages of large-scale AI training is the capability to create datasets that are practically limitless in size.
- **Privacy Preservation**: The artificial nature of the information removes all explicitly identifying characteristics of patients, thus lowering the risk of re-identification.
- **Bias Mitigation**: Minority demographics and rare conditions may be oversampled to establish more balanced datasets.
- **Faster Iteration**: By avoiding bottlenecks in data access, research and model development cycles are significantly accelerated.

But some limitations exist. Although the strength of statistical similarity metrics was considerable, some temporal and causal relationships in patient trajectories might be inaccurately reconstructed, particularly for complex clinical events. Moreover, practitioners may rely too heavily on synthetic data and not validate it sufficiently against reality, which can lead to a set of performance gaps when the model is put into practice. Additionally, the computational costs for training cutting-edge

generative models, especially those based on diffusion architectures, are still relatively high, which means that advanced facilities, which are not always available in resource-limited locations, are needed.

Integration with Emerging Technologies (Federated Learning, Blockchain)

Combining federated learning and synthetic data opens the possibility of utilizing the synergy between these two technologies to train AI on heterogeneous systems without the need for centralized data. In such cases, synthetic data can be created on-site at each member location and merged via standard data models, allowing for the sharing of models to be jointly trained through federated protocols.

This reduces both the privacy risks and the governance burden, which are typically associated with federated AI. The aid of blockchain technology will be available for this, as it provides a transparent and immutable record of the creation of synthetic data, the activities of model training, and the achievement of privacy compliance (Lan et al., 2020). By deploying blockchain-anchored smart contracts, organizations will have the ability to control usage, verify the source of datasets, and ensure agreement with privacy consents that are mutually agreed upon. The new combined technologies may become the basis of a secure, global synthetic data trade system.

Ethical, Legal, and Social Implications (ELSI)

Ethically, synthetic EHRs offer solutions to numerous privacy issues surrounding patients, although they do not eliminate all risks of misuse. Synthetic datasets might capture biases even without the direct inclusion of identifiers or mirror patterns that serve to underrepresent a particular group of people. It requires constant surveillance and bias auditing. Synthetic data constitutes a grey area in legal terms, as most jurisdictions do not consider it to be protected health information (PHI). However, regulators are increasingly demanding disclosure of synthetic data generation techniques (Dave et al., 2024). Since researchers using synthetic EHRs have a greater representation in democratizing AI research, less well-funded institutions, non-profits, and low-resource organizations may easily utilize high-quality training data without the need for overly restrictive data-sharing contracts. The availability of synthetic data can only be accepted by people when there are proper explanations of how that data was developed, checked, and used.

Conclusion

Summary of Key Contributions

This study provides an in-depth analysis of the scope and achievements of generative AI in fabricating synthetic Electronic Health Records (EHRs) for medical AI training by 2025. The results show that the state-of-the-art generative architectures, in particular, GANs, VAEs, and diffusion models, are capable of generating synthetic data, which is similar to the actual patient data and aids in privacy protection.

Key contributions include:

- **Quantitative Validation**: Empirical evidence suggests that artificially generated electronic health record (EHR) models trained on synthetic data can achieve predictive performance comparable to that of models trained on real data, reaching 96-98% of the latter's accuracy across various clinical tasks.

- **Privacy Preservation**: Showed resilience to re-identification attempts, as risk values were near the range of statistical randomness after the application of privacy-preserving techniques.

- **Hybrid Dataset Advantage**: Findings underscore that the use of synthetic data, along with real data, enables higher model generalization, most notably in those subpopulations that have been less represented.

- **Applicability to Low-Resource Settings**: Synthetic EHRs provide a means to practically expand the research area of AI in places where

health digital infrastructure is limited or absent. In this way, they back health equity worldwide.

- **Integration Pathways**: Described how to integrate synthetic data generation into federated learning and blockchain ecosystems to have a safe, collaborative AI development.

Recommendations for Future Work

Despite the evidence highlighting the capability of artificial EHRs for a wide range of AI training applications, they still require groundbreaking research and improvements in their quality:

1. **Enhancing Temporal Fidelity**: Enhance the representation of complex changes over time and causal relationships in synthetic patient trajectories to more accurately reflect rare clinical events.

2. **Bias Auditing Frameworks**: Establish standardized tools for identifying biases in demographics, socioeconomics, and clinical areas and for generating a bias-free dataset.

3. **Regulatory Standards**: Collaborate with legislators to develop official regulations that incorporate healthcare data generated through the use of synthetic methods, verification, and ethical considerations.

4. **Multi-Modal Expansion**: Enhance generation capabilities to facilitate the seamless integration of imaging, genomic, and sensor data, as well as structured electronic health records, for precision medicine research.

5. **Real-World Clinical Trials**: Perform evaluations of AI models that are trained on synthetic or hybrid data with a view to their safety and performance when introduced into the healthcare environment.

These are the steps that the community of AI in healthcare could take to realize the complete capacity of generative AI to produce a synthetic EHR. The pace of that change would be phenomenal, the outreach broader, and the systems would be not only useful but fair in terms of distribution across different patient groups.

Acknowledgment

The author declares no external funding or institutional support for this study.

References

Achterberg, J. L., Haas, M. R., & Spruit, M. R. (2024). On the evaluation of synthetic longitudinal electronic health records. *BMC Medical Research Methodology, 24*(1), 181. https://doi.org/10.1186/s12874-024-02304-4

Chen, Y., & Esmaeilzadeh, P. (2024). Generative AI in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, *26*, e53008. https://doi.org/10.2196/53008

Dave, T., Athaluri, S. A., & Singh, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science, 19*(1), 27. https://doi.org/10.1186/s13012-024-01357-9

Ghosheh, G. O., Li, J., & Zhu, T. (2024). A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Computing Surveys, 56*(6), 1–34. https://doi.org/10.1145/3636424

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Boyd, A. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology, 20*(1), 108. https://doi.org/10.1186/s12874-020-00977-1

Hernandez, M., Epelde, G., Alberdi, A., & Cilla, R. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review, 48*, 100546. https://doi.org/10.1016/j.cosrev.2023.100546

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing, 493*, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

Lan, L., You, L., Zhang, Z., Fan, Z., Zhao, W., Zeng, N., Chen, Y., & Zhou, X. (2020). Generative adversarial networks and its applications in biomedical informatics. *Frontiers in Public Health, 8*, 164. https://doi.org/10.3389/fpubh.2020.00164

Li, J., Cairns, B. J., Li, J., & Zhu, T. (2023). Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine, 6*(1), 98. https://doi.org/10.1038/s41746-023-00834-7

Li, R., Tian, Y., Shen, Z., Li, J., Li, J., Ding, K., & Li, J. (2023). Improving an electronic health record–based clinical prediction model under label deficiency: Network-based generative adversarial semisupervised approach. *JMIR Medical Informatics, 11*(1), e47862. https://doi.org/10.2196/47862

Loni, M., Kangavari, M. R., & Alinejad-Rokny, H. (2025). A review on generative AI models for synthetic medical text, time series, and longitudinal data. *npj Digital Medicine, 8*(1), 95. https://doi.org/10.1038/s41746-024-01409-w

Pezoulas, V. C., Zaridis, D. I., Mylona, E., Androutsos, C., & Fotiadis, D. I. (2024). Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal, 25*, 2892–2910. https://doi.org/10.1016/j.csbj.2024.07.005

Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., & Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications, 13*(1), 7609. https://doi.org/10.1038/s41467-022-35295-1

Yan, C., Zhang, Z., Nyemba, S., & Li, Z. (2024). Generating synthetic electronic health record data using generative adversarial networks: Tutorial. *JMIR AI, 3*, e52615. https://doi.org/10.2196/52615

Zhang, P., & Kamel Boulos, M. N. (2023). Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet*, *15*(9), 286. https://doi.org/10.3390/fi15090286