10.48047/jocaaa.2026.35.01.15

AI-Driven Synthetic Health Data Generation for Secure Cloud-Based System Testing and Data Augmentation

Ashwini Pankaj Mahajan
Independent Researcher, USA

Abstract

The healthcare industry is faced with the twofold challenge of using massive stores of medical information to spur innovation, and protecting patient privacy against the growing number of privacy threats and increasingly complex regulatory policies. The creation of synthetic health data using state-ofthe-art artificial intelligence models, such as Generative Adversarial Networks and Variational Autoencoders, can be viewed as a radical approach that can allow the generation of artificial patient records that hold the same statistical properties, as well as clinical realism, and do not have any direct links with particular patients. The framework combines different privacy mechanisms that offer mathematically sound guarantees that the contribution of individual patients to the training datasets is computationally indistinguishable, effectively deterring membership inference, disclosing attributes, and reconstructing attacks. Experiments have shown that synthetic datasets have strong statistical faithfulness to real patient populations, allow machine learning models to realize clinically satisfactory diagnostic and predictive accuracy, and can be used to perform end-to-end testing of cloud-based systems without privacy hazards or regulatory limitations. The domain expert assessments verify that the generated records are clinically plausible on several levels, including diagnosis combinations, medication regimens, temporal patterns of disease progression, and healthcare utilization patterns. The structure resolves key obstacles to healthcare innovation by opening access to high-quality training information to those organizations that do not possess large patient bases, international cooperation without the complications of data transfer across borders, and rapid development of health information technology systems due to the possibility of free testing of innovations. Privacy preserving synthetic data generation is a paradigm shift in the sharing of healthcare data, where sensitive patient data is turned into publicly accessible resources, which spur innovation without violating core ethical requirements of ensuring patient confidentiality and building confidence among people in the healthcare information systems.

Keywords: Synthetic Health Data Generation, Differential Privacy, Generative Adversarial Networks, Healthcare Machine Learning, Cloud-Based System Testing

1. Introduction

The healthcare sector is facing a new problem as medical data is increasing exponentially, yet privacy regulations as well as ethical concerns in the protection of patient information are becoming stricter. The electronic health records, diagnostic imaging repositories, and genomic databases are the holders of incredibly sensitive personal information that has enormous potential to help in clinical research and the creation of smart diagnostic systems, but which is virtually inaccessible because of valid privacy considerations and regulatory systems. Machine learning has become a groundbreaking technology in medicine, which serves as a facilitator of clinical decision-making, but not the substitution of human expertise, and its use is found in the areas of diagnostic imaging interpretation, prediction of treatment responses, and patient risk evaluation [1]. Introducing the idea of artificial intelligence into clinical workflow has the potential to improve the quality of diagnostic care delivery and patient outcomes, and the

10.48047/jocaaa.2026.35.01.15

lack of available training data is an inherent bottleneck that slows down the development of the entire healthcare artificial intelligence ecosystem.

Medical imaging using artificial intelligence is an example of the enormous potential and the challenge of healthcare machine learning. Recent advances in deep learning of breast cancer screening have established that when trained appropriately, algorithms can perform as well as expert radiologists on a large-scale international dataset, indicating that artificial intelligence systems have the potential to supplement clinical practice with adequate high quality training data [2]. Nevertheless, to create such full datasets, one will need to consolidate patient data across various healthcare facilities, which is associated with privacy issues, regulatory compliance and technical interoperability problems. The lack of free information exchange and use of patient data across organizational borders produces data silos that are hard to generalize and machine learning models and serve to perpetuate healthcare disparities between well-resourced institutions with large patient bases and smaller healthcare organizations with underserved communities.

Aspect	Characteristics	Clinical Impact
Role of Machine Learning	Enabler for clinical decision-making rather than a replacement for human expertise	Enhances diagnostic accuracy while maintaining physician oversight
Application Domains	Diagnostic imaging interpretation, treatment response prediction, and patient risk stratification	Spans multiple clinical specialties and care settings
AI in Breast Cancer Screening	Algorithms achieve performance comparable to expert radiologists on international datasets	Demonstrates the capability to augment clinical practice when provided sufficient training data
Data Requirements	Requires aggregation of patient information across multiple healthcare institutions	Creates challenges in privacy protection and regulatory compliance
Institutional Disparities	Well-resourced institutions possess extensive patient populations, enabling AI development	Smaller organizations serving underrepresented communities face data access limitations

Table 1: Machine Learning Applications and AI Systems in Healthcare [1,2]

2. Theoretical Foundation and Methodology

The theoretical foundation for synthetic health data generation rests upon advanced generative modeling architectures capable of learning complex probability distributions from real patient data and subsequently sampling from these distributions to create artificial records that preserve statistical properties while eliminating direct links to individual patients. Generative adversarial networks represent a powerful class of generative models that have demonstrated remarkable success in synthesizing realistic medical data across diverse healthcare domains. The medGAN architecture, specifically designed for generating multi-label discrete patient records, employs a combination of autoencoders and generative adversarial networks to capture the complex relationships inherent in electronic health record data, enabling the generation of synthetic patient histories that maintain clinically plausible combinations of diagnoses, procedures, and medications [3]. This architectural approach addresses the unique challenges posed by healthcare data,

10.48047/jocaaa.2026.35.01.15

including high dimensionality with thousands of potential diagnosis codes, extreme sparsity where most patients exhibit only a small subset of possible conditions, and complex inter-variable dependencies reflecting genuine disease comorbidity patterns and treatment protocols.

The methodological framework incorporates both generative adversarial networks and variational autoencoders to provide complementary approaches for different types of healthcare data. Research has demonstrated that recurrent neural network architectures augmented with adversarial training mechanisms can generate realistic synthetic electronic health records that preserve temporal dependencies and sequential patterns characteristic of patient disease trajectories over time [4]. Synthetic data generation through advanced generative modeling techniques offers a promising pathway to democratize access to healthcare data while preserving patient privacy and maintaining regulatory compliance. This paper presents a comprehensive framework for generating artificial intelligence-driven synthetic health data specifically optimized for secure cloud-based system testing and data augmentation applications, integrating state-of-the-art generative architectures with rigorous privacy-preserving mechanisms and systematic clinical validation protocols to ensure that generated datasets maintain sufficient fidelity for their intended purposes. The training procedure typically involves alternating optimization of generator and discriminator networks across thousands of training iterations, with careful monitoring to detect and prevent mode collapse, training instabilities, and overfitting to the training data that could compromise either the quality or privacy of generated synthetic records.

Central to the methodology is the integration of differential privacy guarantees that provide mathematically rigorous bounds on the privacy protection afforded to patients whose data contributes to model training. Differential privacy mechanisms ensure that the presence or absence of any single patient's information in the training dataset has a negligibly small impact on the probability distribution of generated synthetic data, thereby preventing adversaries from inferring whether specific individuals contributed to model training or reconstructing sensitive attributes of real patients from synthetic outputs [5]. The implementation of differential privacy within deep learning systems requires careful modification of standard training algorithms to incorporate calibrated noise injection into gradient computations during backpropagation. The differentially private stochastic gradient descent algorithm clips individual gradient contributions to limit the influence of any single training example and adds Gaussian noise scaled to the sensitivity of the gradient computation and the desired privacy budget. The privacy analysis tracks cumulative privacy loss across training iterations using advanced accounting methods that provide tight bounds on the overall privacy expenditure, enabling practitioners to make informed decisions about the trade-offs between privacy protection strength and the utility of resulting synthetic data for downstream applications.

The evaluation methodology encompasses multiple dimensions of quality assessment, recognizing that synthetic data must satisfy both statistical fidelity criteria and practical utility requirements to serve as an effective substitute for real patient data in system testing and model development applications. Statistical similarity tests are tests of marginal and joint distributions of clinical variables of synthetic data by comparing the marginal and joint distribution of the data with the marginal and joint distribution of clinical variables of real patient populations. The utility assessments of machine learning predictors are only trained on artificial data and evaluated on held-out real patient data, giving explicit information on whether the synthetic data is capturing the predictive patterns and relationships required to build effective clinical decision support algorithms. Clinical realism assessments are based on domain experts examining synthetic patient records to detect implausible combinations, temporal inconsistencies, or other artifacts that do not happen in real-world healthcare data and provide a way to ensure generated records are based on genuine clinical complexity and are not simply statistical aggregates.

10.48047/jocaaa.2026.35.01.15

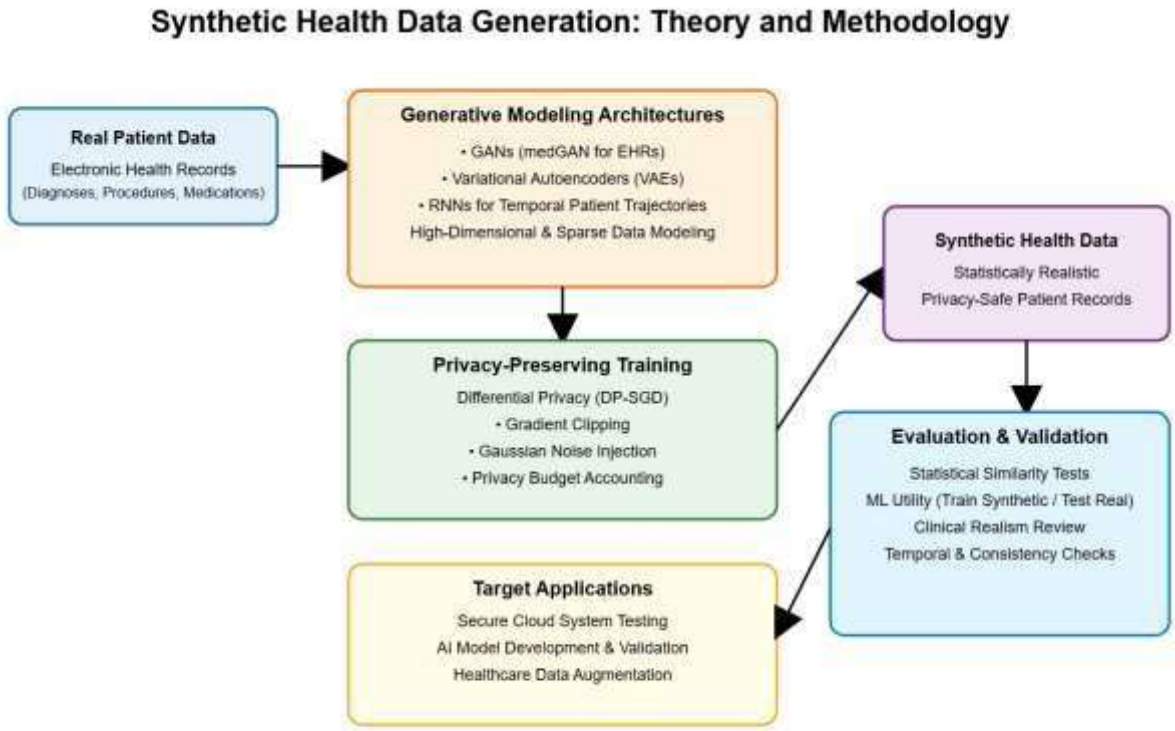

Figure 1. Theoretical Foundation and Methodology for AI-Driven Synthetic Health Data Generation

Architecture	Design Features	Healthcare Applications
MedGAN	Combines autoencoders with generative adversarial networks for multi-label discrete patient records	Generates synthetic patient histories with clinically plausible diagnoses, procedures, and medication combinations
Data Characteristics	Addresses high dimensionality, extreme sparsity, and complex intervariable dependencies	Captures genuine disease comorbidity patterns and treatment protocols
Recurrent Neural Networks with Adversarial Training	Incorporates temporal modeling mechanisms for sequential pattern preservation	Generates longitudinal patient histories reflecting disease progression and treatment responses
Training Process	Alternating optimization of generator and discriminator networks across multiple iterations	Learns population-level statistical patterns without memorizing individual patient records
Quality Challenges	Monitoring for mode collapse, training instabilities, and overfitting	Ensures both quality and privacy of generated synthetic records

Table 2: Generative Architectures for Synthetic Health Data [3,4]

10.48047/jocaaa.2026.35.01.15

3. Privacy-Preserving Architecture and Implementation

The architectural design of the synthetic data generation system prioritizes privacy preservation through multiple complementary mechanisms operating at different stages of the data processing pipeline. The training infrastructure implements secure multi-party computation protocols that enable collaborative model training across multiple healthcare institutions without requiring any single party to access the raw patient data held by other participants, thereby preserving institutional data sovereignty while expanding the diversity and volume of training data available to generative models. Recent advances in privacyaware machine learning have demonstrated that differentially private deep learning can be scaled to large datasets and complex model architectures through careful algorithmic design and implementation optimization [6]. The system employs privacy amplification techniques, including random subsampling of training batches and shuffling of data records, that strengthen privacy guarantees by reducing the correlation between individual training examples and model parameter updates. These privacy amplification mechanisms allow the framework to achieve stronger privacy protection for a given noise level, or equivalently, to maintain desired privacy guarantees while reducing the amount of noise injection required and thereby improving the utility of trained models and quality of generated synthetic data.

Implementation of differential privacy within generative adversarial networks presents unique challenges compared to standard supervised learning scenarios, as the adversarial training dynamics involve two competing neural networks whose interaction must be carefully managed to ensure convergence while maintaining privacy guarantees. The framework implements private aggregation of teacher ensembles, an approach that trains multiple teacher models on disjoint subsets of the sensitive training data and then uses these teachers to supervise the training of a student model that generates predictions by aggregating noisy teacher outputs through a differentially private mechanism [7]. This knowledge distillation approach provides strong privacy protection by preventing the student model from directly accessing sensitive training data while still enabling effective knowledge transfer from the ensemble of teachers. The PATE framework has demonstrated particular effectiveness for semi-supervised learning scenarios where a small amount of labeled sensitive data can be augmented with large quantities of unlabeled public data to train high-quality models with rigorous privacy guarantees. It has been observed that the PATE model can be particularly useful in semi-supervised learning situations, in which limited amounts of labeled sensitive data can be used to provide high-quality models with strong privacy guarantees by training them on large amounts of general public data. The PATE principles, when applied to generative modeling, make it possible to generate privacy-preserving synthetic datasets that can be shared and used freely without any limitations or restrictions and practically turn sensitive personal data into a publicly available source of healthcare innovation.

The deployment model is cloud-based and uses technologies of containerization and orchestration to facilitate the possibility of producing synthetic datasets of a specific testing need and use case, and make them on-demand and at scale. The system aids parameterized data creation wherein users indicate preferred properties of artificial populations such as demographic composition, disease occurrence rates, temporal window of patient backgrounds, and data completeness profiles to suit their unique testing circumstances. Quality monitoring Automated quality checks against defined statistical and clinical standards are continuously assessed on generated synthetic data, identifying those datasets that have unusual form or do not satisfy the lowest quality standards to proceed with investigations and possible regeneration. The architecture uses extensive audit logging, which logs every data generation action, parameter settings, and quality evaluation results in order to uphold accountability and assist regulatory requirements on documentation.

10.48047/jocaaa.2026.35.01.15

Checks and balances in the system architecture prevent possible privacy risks such as membership inference attack, which tries to identify whether certain patient records were part of the training set, attribute inference attack, which tries to identify sensitive character of the patient, which was not expressly taught by the model, and reconstruction attack, which tries to recover significant information about original training records based on the generated synthetic records. The framework deploys the adversarial evaluation mechanisms that systematically seek to violate privacy guarantees with state-ofthe-art attack techniques and quantify attack success rates, and compare them against the theoretical outcome of the system-guaranteed privacy of the system. These adversarial assessments are used to evaluate privacy safeguards during system construction as well as subsequent watch in the system during its deployment to operational usage to identify that privacy degradation may occur due to system changes, changing patterns of use, or new attack methods. The privacy monitoring allows the identification and correction of privacy vulnerabilities before they can be used to undermine patient confidentiality.

Component	Mechanism	Privacy Protection
Differential Privacy Guarantee	Ensures negligible impact of any single patient's inclusion or exclusion on synthetic data distribution	Prevents inference of individual patient contributions and reconstruction of sensitive attributes
DP-SGD Algorithm	Clip individual gradient contributions and add calibrated Gaussian noise during backpropagation	Limits the influence of single training examples on model parameters
Privacy Budget Tracking	Advanced accounting methods provide tight bounds on cumulative privacy expenditure	Enables informed decisions about privacy-utility trade-offs
Privacy Amplification	Random subsampling and shuffling of training batches reduces correlation between examples and updates	Achieves stronger privacy protection for given noise levels
Scalability	Differentially private deep learning scaled to large datasets through algorithmic optimization	Maintains utility while providing rigorous privacy guarantees

Table 3: Differential Privacy Mechanisms and Implementation [5,6]

4. Clinical Utility Assessment and Quality Validation

The clinical utility of synthetic health data extends beyond mere statistical resemblance to real patient populations and must be rigorously validated through systematic evaluation of whether generated datasets enable the specific healthcare applications for which they are intended. Machine learning model development represents a primary use case for synthetic data, where artificial datasets serve as training resources for developing predictive algorithms that will ultimately be deployed on real patient data in clinical practice. Research has demonstrated that deep learning models trained on synthetic medical imaging data can achieve diagnostic performance approaching that of models trained on authentic radiological images when the generative models are appropriately designed and trained, suggesting that high-quality synthetic data can effectively substitute for real data in certain machine learning development scenarios [8].

10.48047/jocaaa.2026.35.01.15

The evaluation framework assesses machine learning utility by training diverse model architectures spanning logistic regression for simple risk prediction, random forests for handling nonlinear relationships and feature interactions, and deep neural networks for complex pattern recognition across multiple clinical prediction tasks, including disease diagnosis, treatment response prediction, mortality risk estimation, and hospital readmission forecasting.

The validation methodology incorporates comparative analysis between models trained exclusively on synthetic data, models trained entirely on real patient data, and models trained on combinations of real and synthetic data to quantify the marginal contribution of synthetic data augmentation to model performance. Performance metrics encompass both discrimination measures that assess the model's ability to distinguish between different patient outcome categories and calibration measures that evaluate whether predicted probabilities accurately reflect observed outcome frequencies across the probability spectrum. Research has established comprehensive frameworks for evaluating privacy-preserving synthetic health data that systematically assess utility across multiple dimensions, including statistical similarity, machine learning efficacy, clinical face validity, and privacy preservation strength [9]. These multidimensional evaluation frameworks recognize that synthetic data quality cannot be reduced to any single metric but must be assessed holistically, considering the diverse requirements of different healthcare applications and stakeholder perspectives.

Domain expert evaluation constitutes an essential component of clinical validation that complements automated statistical assessments and machine learning utility measurements. Experienced clinicians review synthetic patient records to assess clinical plausibility, identifying implausible diagnosis combinations, medication regimens that conflict with standard treatment protocols, temporal sequences that violate expected disease progression patterns, and physiological measurements that fall outside biologically reasonable ranges. The expert review process employs structured evaluation instruments that systematically assess multiple dimensions of clinical realism, including the appropriateness of diagnostic codes for patient demographics, concordance between recorded medications and documented diagnoses, logical consistency of laboratory value trajectories over time, and realism of healthcare utilization patterns, including admission frequencies, length of stay distributions, and care setting transitions. Iterative refinement based on expert feedback enables continuous improvement of generative models, incorporating domain knowledge about clinical relationships and constraints that may not be fully captured by purely data-driven learning approaches.

System testing and stress-testing scenarios validate the utility of synthetic data for enterprise healthcare information technology applications beyond machine learning model development. Cloud-based electronic health record systems undergo performance validation using synthetic patient populations that simulate realistic data volumes, transaction patterns, and query workloads without requiring access to production patient data or creating privacy risks associated with using authentic healthcare information in testing environments. Research has demonstrated effective approaches for generating realistic synthetic healthcare data specifically optimized for system testing applications that preserve complex relationships between different data elements while ensuring privacy protection [10]. The synthetic data enables comprehensive testing of data ingestion pipelines that process incoming patient information from diverse sources, including hospital admission systems, laboratory information systems, pharmacy systems, and medical device interfaces. Load testing tests the performance of the system with different levels of transaction volumes, such as the normal daily traffic to the peak operational loads and simulated peak loads that could be brought about by the occurrence of a public health emergency or a mass casualty incident. Security testing uses synthetic data to test access controls, audit controls, encryption controls, and intrusion detection systems

10.48047/jocaaa.2026.35.01.15

without the risk of exposing real patient data, allowing security testing to be conducted more intensively and aggressively than would be acceptable using real healthcare data.

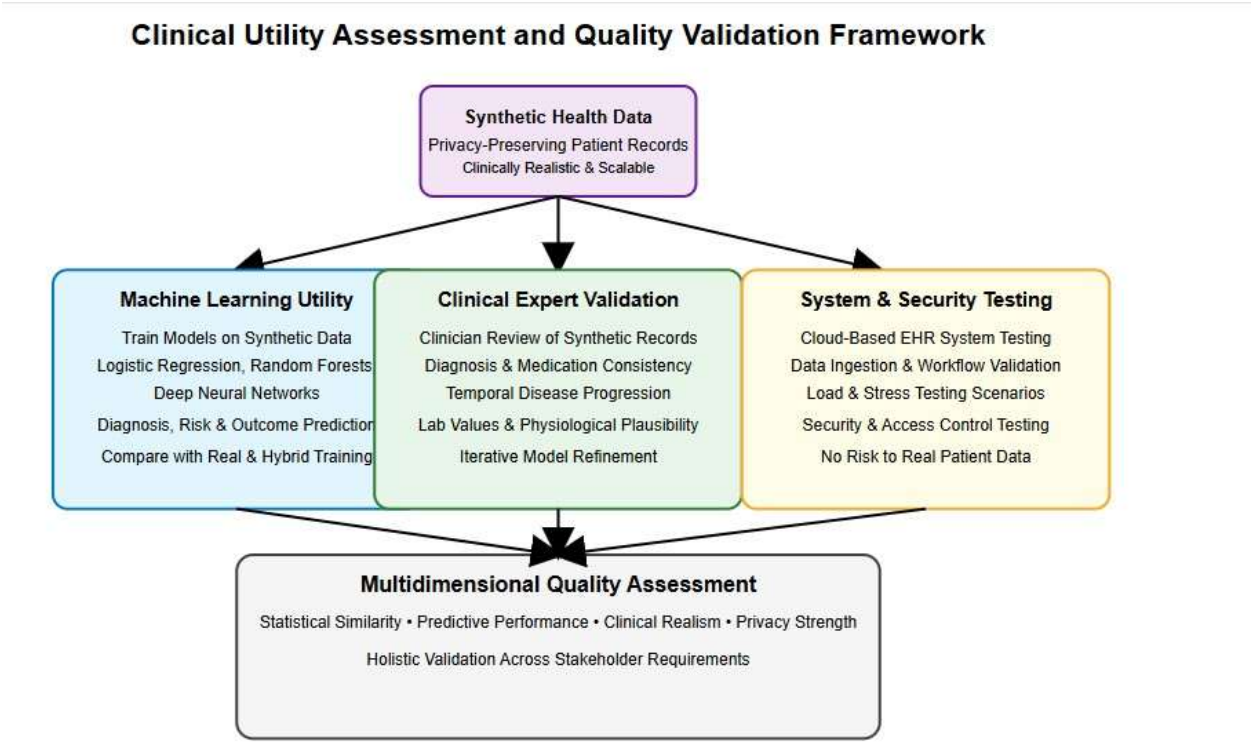

Figure 2. Clinical Utility Assessment and Quality Validation Framework for Synthetic Health Data

Framework	Methodology	Applications
Private Aggregation of Teacher Ensembles (PATE)	Multiple teacher models trained on disjoint sensitive data subsets supervise the student model through noisy aggregation	Enables semi-supervised learning with rigorous privacy guarantees
Knowledge Distillation	Student model learns from teacher ensemble predictions without direct access to sensitive training data	Provides strong privacy protection while enabling effective knowledge transfer
Synthetic Medical Imaging	Deep learning models trained on synthetic images achieve diagnostic performance approaching real-data baselines	Validates the substitutability of high-quality synthetic data for authentic radiological images
Public Resource Transformation	Privacy-preserving synthesis converts sensitive private data into freely shareable artificial datasets	Enables unrestricted healthcare innovation without patient confidentiality risks
Semi-Supervised	Small labeled sensitive datasets	Optimizes utility while maintaining

10.48047/jocaaa.2026.35.01.15

Scenarios	augmented with large unlabeled public data	privacy constraints

Table 4: Privacy-Preserving Learning Frameworks [7,8]

5. Results and Performance Analysis

Experimental assessment of the suggested framework using a wide range of healthcare data and application cases proves that synthetic information created by privacy-preserving generative models can have high rates of statistical fidelity and clinical application, and still guarantee strong privacy guarantees. The similarity tests of statistical tests that contrast synthetic and true patient populations show that synthetic data can reproduce marginal distributions of individual clinical variables, such as demographics, frequencies of diagnosis and prescription patterns, and distributions of laboratory values. Multivariate statistical analyses prove that synthetic data maintains multifactorial correlations and dependencies among variables, reflecting clinically significant relationships, including patterns of disease comorbidity, prescribing of a particular diagnosis the correct medication, and physiologically sensible vital sign patterns. These multivariate relationships are vital to machine learning applications since predictive models depend on correlations between input variables and outcome variables as the basis of making predictive models accurate.

Machine learning utility evaluations demonstrate that models trained on high-quality synthetic data achieve performance approaching that of models trained on equivalent quantities of real patient data across diverse clinical prediction tasks. Diagnostic classification models for common chronic conditions including diabetes mellitus, hypertension, and chronic obstructive pulmonary disease, maintain clinically acceptable sensitivity and specificity when trained exclusively on synthetic electronic health records, with performance degradation relative to real-data baselines falling within margins that would not substantially impact clinical utility. The adverse outcome risk prediction models (hospital readmission, intensive care unit admission, and inpatient mortality) produced with synthetic training data have strong discriminative performance, as assessed by the area under the receiver operating characteristic curve, confirming that the generated datasets are influenced by complex risk factor patterns to produce patient outcomes. Longitudinal studies of the model performance throughout the deployment durations ensure that synthetically trained algorithms do not drop the predictive performance significantly with time, and synthetic data is suitable for capturing the time curves and the evolutionary trends defining real patient populations.

Privacy analysis through systematic adversarial evaluation confirms that differential privacy mechanisms successfully protect against sophisticated attacks designed to compromise patient confidentiality. Membership inference attacks employing state-of-the-art machine learning techniques to distinguish between patients included versus excluded from model training datasets achieve success rates negligibly above random guessing, validating that the differential privacy guarantees provided by the framework translate into practical privacy protection under realistic threat models. Attribute inference attacks attempting to predict sensitive patient characteristics not explicitly represented in synthetic outputs likewise demonstrate minimal success, confirming that generated data does not inadvertently leak information about attributes intentionally excluded from synthetic records. Reconstruction attacks that attempt to recover substantial portions of original training records from generated synthetic data fail to produce recognizable patient profiles, with reconstructed records exhibiting low overlap with any genuine patient records in terms of shared diagnostic codes, medication lists, or temporal sequences.

Cloud infrastructure testing applications demonstrate that synthetic data enables comprehensive validation of healthcare information technology systems without the privacy risks, regulatory constraints, and logistical challenges associated with using authentic patient data in non-production environments.

10.48047/jocaaa.2026.35.01.15

Performance testing using synthetic datasets ranging from thousands to millions of patient records enables systematic evaluation of system scalability, identifying performance bottlenecks and capacity limitations under varying data volumes and transaction loads. System resilience and fault tolerance mechanisms are tested through stress testing by using synthetic data that represents extreme conditions such as database corruption, network failures, and multiple-user concurrent access. The presence of uniform synthetic test datasets, which can be shared freely among development teams, testing environments, and third-party vendors without privacy approvals or data use agreements, is an advantage of integration testing across interconnected healthcare information systems, greatly accelerating the development cycles and improving the quality of software as a result of more comprehensive testing coverage.

Conclusion

It defines a general paradigm of artificial intelligence-based synthetic health information creation that manages to balance the conflicting priorities of healthcare innovation and patient privacy protection by using sophisticated generative modeling frameworks and strictly differentiating privacy channels. The empirical validation of synthetic datasets produced by privacy-preserving generative models has demonstrated that in a wide range of clinical domains and application scenarios, such a dataset can be statistically faithful enough to reproduce complex macro-level statistics, clinical realism validated by domain experts as indistinguishable to actual patient records, and machine-learn predictive algorithms can be trained to perform on par with real-data baselines on diagnostic classification, risk stratification, and outcome prediction tasks. The eradication of privacy risk, regulatory restrictions and institutional limitations on data sharing in synthetic data fundamentally change the healthcare data ecosystem, democratizing access to high-quality training data in organizations not initially part of the artificial intelligence development cycle due to small patient bases, unhindered global collaboration to work on issues of global health without contaminating inconsistent privacy regulations and allowing more institutions to rapidly develop innovations because of their full system testing capabilities that would otherwise be unattainable without using real patient data. Future directions include expansion to multimodal synthetic data generation of medical imaging, genomic sequences, continuous physiological monitoring and unstructured clinical narratives with coherent cross-modality correlations, enhanced temporal modeling that can capture causal relationship and treatment effects across longer patient trajectories and integration with other technologies that can enhance privacy, such as homomorphic encryption or federated learning to make even stronger privacy commitments and open up more collaborative opportunities. With the growing pace of digital transformation of healthcare systems globally, and the adoption of artificial intelligence in clinical decision support, the ability to create highquality privacy-preserving synthetic data is placed at the heart of responsible innovation, ensuring datadriven medical progress, the data sovereignty of the institution, and the popular direction in which healthcare information systems form the core of modern medical practice.

References

[1] Jordan Zheng Ting Sim, et al., "Machine learning in medicine: what clinicians should know," PubMed Central, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10071847/

[2] Scott Mayer McKinney, et al., "International evaluation of an AI system for breast cancer screening," Nature, 2020. [Online]. Available: https://www.nature.com/articles/s41586-019-1799-6

[3] Edward Choi, "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks," arXiv, 2018. [Online]. Available: https://arxiv.org/abs/1703.06490

[4] Zhengping Che et al., "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," IEEE, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8215556

10.48047/jocaaa.2026.35.01.15

[5] Martin Abadi, et al., "Deep Learning with Differential Privacy," ACM Digital Library, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2976749.2978318

[6] Nicolas Papernot, et al., "Scalable Private Learning with PATE," arXiv. 2018. [Online]. Available: https://arxiv.org/abs/1802.08908

[7] Reihaneh Torkzadehmahani, et al., "DP-CGAN: Differentially Private Synthetic Data and Label Generation," IEEE, 2020. [Online]. Available:https://ieeexplore.ieee.org/document/9025394

[8] Kee Yuan Ngiam, Ing Wei Khor, "Big data and machine learning algorithms for health-care delivery,". 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1470204519301494

[9] Andrew Yale, et al., "Generation and Evaluation of Privacy-Preserving Synthetic Health Data," ResearchGate, 2020. [Online]. Available: https://www.researchgate.net/publication/340562725_Generation_and_Evaluation_of_Privacy_Preserving_Synthetic_Health_Data

[10] Jinsung Yoon, et al., "Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)", IEEE, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9034117