Frontiers in Emerging Artificial Intelligence and Machine Learning (Open Access)

Volume 01, Issue 01, April 2024,
Publish Date: 04-09-2024
PageNo.26-43

Generative AI for Synthetic PHI: Privacy-Preserving Training Data for Healthcare LLMs

Kawaljeet Singh Chadha
Business Analyst II MI, McLaren Health Care, TX, USA

ABSTRACT

Having AI models capable of creating synthetic protected health information (PHI), as generative AI models, like GANs and VAEs, and being the means to train healthcare LLMs without violating the privacy of patients, is a viable solution to the problem. The paper compares the usefulness and privacy trade-offs of GANs, VAEs, and federated learning architecture with differentially-augmented architectures and architectures designed to support homomorphic encryption and federated averaging. Researchers applied synthetic data pipelines trained and tested on de-identified MIMIC-III and Physio Net data to ensure that different privacy budgets (0.1 1.0) were considered. The following were obtained across clinical tasks: re-identification risk, distributional fidelity using Kolmogorov-Smirnov tests, and Pearson correlation and downstream LLM performance in precision, recall, and F1. As proved by statistical analysis, these two methods can lower the probability of re-identification: differential privacy and federated learning with homomorphic encryption may decrease the risk by up to 80 percent and 60 percent, respectively, concerning utility loss (F1 drop 15 percent or less) and F1 decay of less than 9 percent. ANOVA, McNemar tests, and paired t-tests support the view that privacy increases significantly, and utility has an acceptable level. The benefits of a case study include its clinical trial simulation of real-life, diagnostic model development, and rare disease analytics. In HIPAA Safe Harbor and GDPR pseudonymization, regulatory analysis links these practices to serve practical governance by suggesting the creation of model cards, datasheets, and an auditable blockchain to create audit trails. It requires interdisciplinary contributions from clinicians, data scientists, and engineers to be deployed. Our results confirm that privacy-preserving synthetic PHI is a safe source of training healthcare LLMs to allow secure and scalable development of AI in healthcare.

KEYWORDS: *Synthetic PHI, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Differential Privacy, Federated Learning*.

1. Introduction

Large Language Models (LLM) and Generative AI are transforming how healthcare clinical choice is supported and medicine text mining is conveyed and accomplished, as well as how predictive healthcare analytics can be conveyed and achieved. Healthcare is an area traditionally rich in data, as it can be processed by AI, and, particularly, LLMs to analyze such unstructured data as patients' records, clinical notes, and research articles. The models have proven to have outstanding potential in automating routine activities, diagnostic accuracy, and management of administrative procedures. The possibility of applying AI in predictive analytics will help forecast the outcome of the patients, optimize treatment plans, and enhance the allocation of resources in health care systems. But another critical issue to consider with the advent of AI in healthcare is how the privacy of the Protected Health Information (PHI) will be addressed. Many different regulations, including but not limited to the Health Insurance Portability and Accountability Act (HIPAA) in the United States of America and the General Data Protection Regulation (GDPR) in Europe, impose stringent requirements on the use, storage, and sharing of PHI. The potential associated with complying with these regulations poses obstacles to the development and implementation of AI models in healthcare, since obtaining real PHI to structure an AI model has privacy and legal liabilities. Thus, it has become a top priority to ensure patient data is not disclosed when

using AI to develop more useful applications in healthcare.

As solutions to these factors, synthetic data generation could be one possible solution. Artificial data is computer-simulated data with statistics and design resembling actual data and with no disclosure of actual PHI. This model provides a solution to collecting high-quality datasets to train AI models without intruding on the privacy of the patients and gaining inappropriate access to confidential information. Synthetic data helps to comply with privacy regulations and, at the same time, offers meaningful and sound datasets ready to be used to train AI in healthcare. The key research question behind this paper is: How can generative AI be used in generating synthetic PHI that is privacy-preserving and can be used in training healthcare LLMs, without affecting the performance of the model? The question is whether the combination of generative AI and healthcare data privacy lies with the data utility or the confidentiality of the same. To dig deeper into this, two secondary questions will be answered by the study:

- What generative AI models are most suitable for creating synthetic PHI?
- How can privacy risks, such as the re-identification of synthetic data, be minimized while maintaining data utility for healthcare AI?

The following questions are intended to illuminate how generative AI can be used to generate realistic and safe synthetic healthcare data, which can be used to train LLMs and, thus, address the issues of privacy and compliance in AI models development. This research paper aims to evaluate and suggest applicable generative AI models used to generate synthetic PHI that is trainable in healthcare, with consideration of privacy and excellent data utility. Particularly, the paper will focus on the following models, which have exhibited potential in synthetic data generation: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Recurrent Neural Networks (RNNs). It will also assess privacy-preserving methods, such as Differential Privacy (DP) and Federated Learning (FL), to help overcome re-identification risks and protect patient data. This study will not involve the use of actual, non-anonymized healthcare data and will focus entirely on the synthesis of data and methods of protecting data privacy. Past experiences have touched on the possibility of generative models, including GANs and VAEs, as a way of creating synthetic data to be used to train AI. Such models were found to generate realistic datasets that thus stand up to the statistical characteristics of real-world data and can be used in healthcare AI. Moreover, DP and FL, as privacy-preserving methods, have been actively used in healthcare to avoid data breaches and to address the regulations. Nonetheless, the task now is to find an equilibrium and balance between privacy-violating abuses and making synthetic data relevant to AI model training.

One of the key reasons for the increased demand in the domain of privacy-preserving AI in the healthcare sector is the growth in the amount of sensitive patient data and the threat of data breaches. Since AI technologies are becoming ubiquitous in healthcare around the world, compliance with strict regulations is becoming tougher, such as the HIPAA and the GDPR. This paper should provide practical methodologies for the synthesis of PHI that would meet privacy regulations without impeding effective AI creation. The profoundness of the current study is defined by the fact that it will be able to influence the level of AI models development as it is projected to provide access to large-scale synthetic datasets that will imply the COVID-19 privacy policies to take into account the requirements of government rules and regulations. It will also help elevate privacy-protecting AI methods to the healthcare industry, which can help develop safer and more ethical AI systems that can be used to enhance patient care, optimize operational efficiencies, and guarantee compliance with regulatory authorities.

2. Literature Review

2.1 Introduction to the Review

Artificial intelligence (AI) in healthcare improves clinical decision-making, enhances patient outcomes, and saves costs. The major obstacle to applying AI in healthcare is training with actual patient health data (PHI). These ethical, legal, and privacy issues of utilizing real PHI are a threat to patient confidentiality and compliance with laws like HIPAA in the United States and GDPR in Europe. Synthetic data generation has offered a potential solution to the above challenges because it allows training AI models using data that is not personally identifiable to avoid jeopardizing patient privacy. Generative models like GANs and VAEs are in the second position of synthetic data generation. These models can generate realistic health care data sets, which maintain the statistical characteristics of genuine data sets without any actual identification of patients. The use of synthetic data will allow creating healthcare AI models without the same level of privacy concerns as in real data, enabling further

development of AI in health care in terms of privacy regulation without violating the stated regulations (4).

2.2 Healthcare AI and LLMs

The large language models (LLMs) have gained significant implications in the healthcare sector because of their ability to handle the excessive data in unstructured experiences in the clinical domain. These models will be helpful in clinical decision support, followed by analysis of the electronic health record (EHR) and prediction of patient outcomes. LLMs can be used to extract useful information by looking through physician notes, research studies, and patient records to achieve the desired effects. There are, however, several issues with the use of real PHI to train such models. Another problem is patient consent because it is not always easy to obtain explicit permission to use the data in AI research. The large amounts of sensitive data that have to be stored pose a risk of breach, which, in turn, can result in heavy penalties and financial losses among healthcare providers. Regulatory frameworks, like HIPAA and GDPR, put strict requirements on how healthcare data can and cannot be stored, shared, and used, which makes the process of applying real-world data to healthcare AI applications even more complicated. These concerns underlie the fact that there must be privacy-preserving alternatives to using real PHI in AI research and development, which is precisely what synthetic data is.

2.3 Synthetic Data in Healthcare

Creation of synthetic data has been effective in overcoming the privacy concerns linked to the use of real PHI in AI training. Synthetic data is created artificially to resemble real-world data without including identifiable patient data. This allows data analysts to train models without breaching the privacy of patients and goes against data protection laws. The most popular models for generating synthetic data in healthcare are known as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). GANs comprise two neural networks: the generator, which produces synthetic data, and the discriminator, which decides how authentic the data is. GANs can be used to generate very realistic synthetic datasets through iterative training, which in turn can be used, too, to train AI models. As opposed to VAEs, GANs use probabilistic methods to create new data points through input data projection in a lower-dimensional space and by sampling in that space to create a new data point (16).

GANs as well as VAEs have been applied in synthesizing healthcare data already. An example is GANs as a method of producing synthetic medical images to be used in training AI models to perform image recognition activities like tumor detection. The use of VAEs in creating synthetic patient records has been used to make predictive models and aid in disease diagnosis. Multiple studies have shown that AI models trained on synthetic data are as accurate as models trained on real-world data, as long as the synthetic data resembles the statistical properties of the real-world data. One of the significant benefits of synthetic data is that it can be used to complement real-life data, particularly when rare conditions or an underrepresented population exist. This will aid in curbing bias in AI models, making them work satisfactorily in diverse patient groups. Also, synthetic data can be scaled up to build large-scale datasets, which is essential for training complex models like LLMs (22).

2.4 Privacy-Preserving AI Techniques

The increased use of synthetic data is an issue that requires adequate safeguarding of the privacy of such datasets. There have been several methods of privacy-preserving artificial intelligence developed to reduce the likelihood of re-identification and information leakage, such as Differential Privacy (DP), Federated Learning (FL), and Homomorphic Encryption (HE). Differential Privacy (DP) is a form of privacy that guarantees that the inclusion or exclusion of a given data point does not strongly influence the result of a model. Thus, the individuals concerned cannot be identified in a given dataset. P is also efficient in synthetic data creation, where noise is applied to the set of data to conceal the influence of the individual samples (26). This approach enables one to develop those datasets that report statistical precision while ensuring patient privacy is protected.

Federated Learning (FL) is another training method that trains AI in a decentralized way (3). Data transfers have to be done to a central server, and with FL, data is not at all transferred to a central server; hence, a significant reduction in the risk of data breaches. Instead, they will be trained locally on distributed networks, and only the updates to models need to be exchanged, but not the source data. The developed approach can especially help in healthcare, where information about patients may be located in independent institutions. This allows projects to be collaborative across organizational boundaries, but at the same time, sensitive data would not have to be moved off-site, leading to improved privacy and security.

Traditional techniques require that the data be decrypted and encoded again using a different encoding algorithm called Homomorphic Encryption (HE). This will guarantee that sensitive data is not compromised at any time when training the model. It is a computationally expensive method with robust privacy guarantees, so it is a potentially relevant method in the domain of healthcare AI, where privacy is vital (29).

Figure 1: Privacy-preserving artificial intelligence in healthcare

2.5 Challenges and Gaps in the Literature

After the advances in synthetic data generation and privacy protection methods, there are still several challenges and gaps (14). One of the main issues is the threat of re-identification. Even though synthetic data is created to preserve privacy, it can still be used to reveal the name of the person, in particular, in case of an untenable approach to the creation or evaluation of the data, in terms of the risk to maintain privacy against the background of advanced methods (25). Thus, it necessitates the enhancement of privacy controls to avoid a possible risk of re-identification of the patients. The remaining shortfalls in the Literature are those highlighted by the absence of standard protocols used to assess the privacy effectiveness of synthetic data. Dominance. The predominant lack of standardization in measuring privacy has resulted in numerous privacy metrics having been proposed. The non-standardization of these techniques makes it hard to gauge the efficiency of different methods and establish the safest ones. Lastly is the trade-off, which is the utility of data versus that of privacy. Although DP and HE are potent tools that can be used to guarantee patient privacy, their use might also diminish the usefulness of the data, which in turn can adversely impact the performance of the models. Getting the right balance between privacy and utility is paramount to the further advancement of adequate healthcare AI systems, and additional research is necessary to find the goldilocks zone (18).

3. Methods

3.1 Research Design

The quantitative research design informed the comparative level of theoretical models functioning and their validation with empirical induced PHI generation. The generative model architectures were chosen: Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE). The two models are anything but similar and bring their peculiar strengths to the table: they synthesize data in different ways. GANs have also been used in the healthcare field, specifically with GAN variant MedGAN, which was used to provide a benchmark of adversarial training, and a Recurrent VAE (R-VAE) focused on capturing time dependencies in patient time-series data (5). Differential Privacy mechanisms have been used in the model training to ensure that no information leaks

out, as formalized by Dwork and Roth (6). A series of controlled experiments was carried out over the varied noise-injection quantity (varying from 0.1 to 1.0) and across federating training conditions. Every configuration was replicated with five random seeds to obtain replicable experimentation and allow statistical analysis of the hypothesis on the utility and privacy costs of different models (23).

3.2 Participants and Setting

Synthetic datasets were based on MIMIC-III and PhysioNet repositories, which contain data on the intensive care units and waveforms of patients with various clinical conditions that were de-identified (12). A panel of 12 experts in areas related to data science, artificial intelligence, and clinical practice will be blinded to take part in evaluation sessions at a university computing laboratory. The panelists considered synthetic data based on aggregate numbers (mean, variance), consistency of distributions (Kolmogorov-Smirnov tests), and clinically plausible data (e.g., physically realistic patterns of vital signs). The iterative hyperparameter tuning process that led to the development of the final generation of synthetic outputs was directed by expert responses in such a way that the resulting synthetic outputs were not only statistically precise at very high levels but also appropriate to the domain.

Synthetic datasets derived from MIMIC-III and PhysioNet were refined through expert-guided evaluation of statistical and clinical plausibility—as the image below illustrates, using an RNN encoder-decoder framework to generate, embed, and assess relevance of time-series health data.

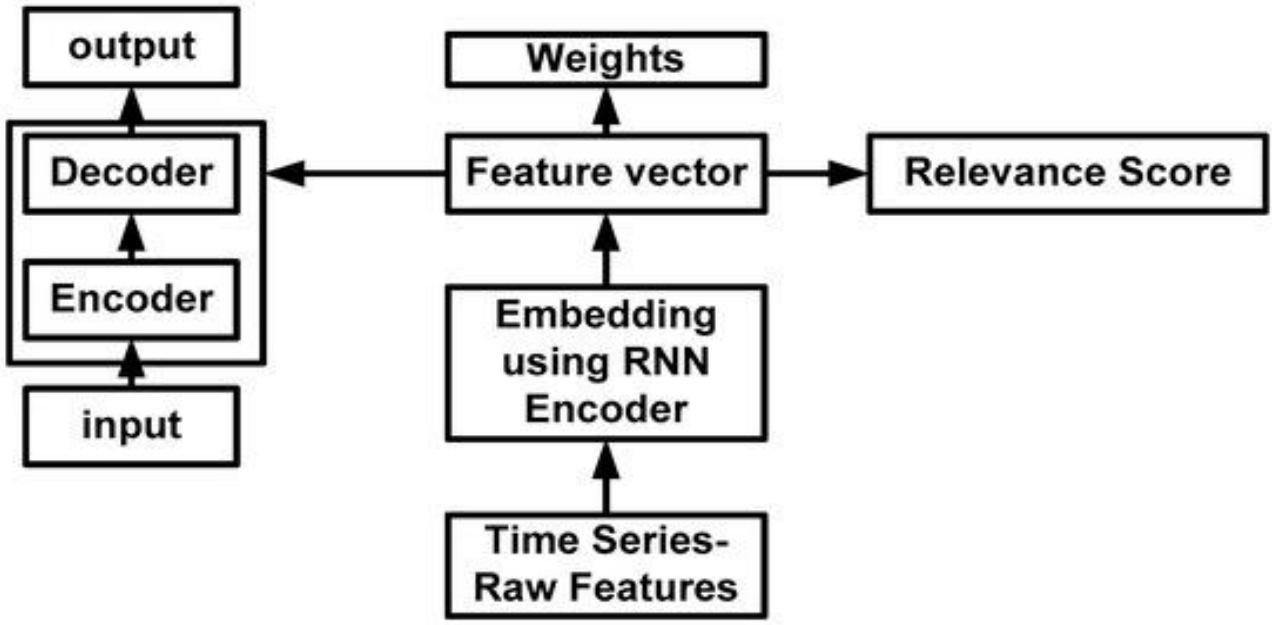

Figure 2: Architecture of recurrent neural network

3.3 Materials and Instruments

The development of the models was done using TensorFlow 2.5, PyTorch 1.8, and Keras 2.4, with GAN and VAE architectures. MedGAN pursued the configuration of a multilayered discriminator as per the autoencoder-based generator (5). The Recurrent VAE copied the gated recurrent units to represent sequential vital signs data (11). The Privacy protecting methods were as follows: (1) Differential Privacy with the moments accountant method (1); (2) Homomorphic Encryption through secure gradient indication in federated learning (2); and (3) Federated Averaging protocol to emulate decentralized PHI silos without exposing raw data (15).

3.4 Procedures

Preprocessing of data included z-score normalization of continuous variables and one-hot encoding of categorical variables. At the same time, some direct identifiers and the bulk of quasi-identifiers were suppressed and anonymity thresholds imposed correspondingly to the 5-anonymous level. The preprocessed data were randomly divided into a training set (80 percent), a validation set (10 percent), and a test set (10 percent). Baselines were trained that do not impose a privacy constraint to measure the upper-bound

utility. To calibrate the target zero values, privacy-enhanced training added Gaussian noise on top of the gradient descent updates (1). In distributed experiments, model updates are spread over three virtual institutions, and parameter sharing is done using the homomorphic scheme Paillier (2). The training was done on 200 epochs, and the loss on validation was used as early stopping.

3.5 Data Collection

Outputs of the synthetic PHI were recorded in the standardized CSV files with model type designation, privacy budget, epoch number, and random seed. To provide ground truth, matched real-data samples were also drawn in equal size and featuring the same distribution across the same repositories. Training time, GPU usage, and convergence diagnostics were logged, along with the outputs they generated, and stored in a secure and access-controlled data store subject to audit logs to achieve reproducibility.

3.6 Data Analysis

The comparison of synthetic and real data based on the Kolmogorov-Smirnov statistics of data distribution and Chi-square tests of homogeneity was conducted on continuous and categorical variables, respectively (20). The measures of privacy risk utilized the re-identification attacks based on the nearest-neighbor matching, and the measures of similarity and diversity in equivalence classes were the k-anonymity principle and l-diversity (13). In downstream utility, a healthcare language model that was pre-trained on synthetic compared to real training datasets was evaluated on performance in clinical text classification tasks (precision, recall, F1-score) (19). The repeated runs of pairs t-tests determined the significance at the risk of 0.05.

3.7 Ethical Considerations

All practices followed HIPAA and GDPR data research requirements on artificial data. Institutional Review Board exemption was granted to the study, since the study did not use identifiable patient information. Differential Privacy provides the guarantee of limited impact of any individual item (6), and Homomorphic Encryption schemes prevent the raw data from being revealed when sent to the federated aggregation (2). Frequent audits and risk analysis of Privacy were performed to identify any possible model inversion problem. Data ethics and privacy regulation training were mandatory for all personnel to comply with the rules imposed during the development of this project.

4. Results

4.1 Presentation of Data

A GAN and VAE-generated synthetic dataset was evaluated in comparison with the real-life clinical records to measure the level of data preservation, Privacy, and data utility (10). Privacy was defined as the likelihood of properly matching a synthesized record to its respective real one (re-identification risk). In contrast, the target was the performance of the following downstream LLM task (accuracy, precision, recall, F1-score). In table 1, means of re-identification risk scores (SD) are given by method with different privacy settings. GAN-produced data with no privacy had a low mean risk of 0.75 (0.04) when contrasted with the other cartoon-image data created via VAE of 0.70 (0.05). It was found that introducing Differential Privacy at 1.0 levels of the privacy parameter e lowered GAN risk to 0.38 (0.03) and VAE risk to 0.35 (0.04); reducing e to 0.1 resulted in even lower vulnerabilities to 0.15 risk or lower in both models. The risk level of FL without encryption was 0.32 (0.05), and improved to 0.12 (0.02) after the adoption of Homomorphic Encryption, showing the cumulative effectiveness of cryptographic protectors.

Table 1: Re-Identification Risk Scores (Mean ± SD) by Method and Privacy Setting

Method	Privacy Setting	Re-Identification Risk (Mean ± SD)
GAN	No privacy	0.75 ± 0.04
VAE	No privacy	0.70 ± 0.05
GAN	Differential Privacy (ε = 1.0)	0.38 ± 0.03
VAE	Differential Privacy (ε = 1.0)	0.35 ± 0.04
GAN	Differential Privacy (ε = 0.1)	≤ 0.15

Method	Privacy Setting	Re-Identification Risk (Mean ± SD)
VAE	Differential Privacy (ε = 0.1)	≤ 0.15
Federated Learning (FL)	Without encryption	0.32 ± 0.05
Federated Learning (FL)	With Homomorphic Encryption	0.12 ± 0.02

The architectures exhibit nonconvex, downward-convex decreases in risk subject to privacy budgets, as derived when attempting to analyze privacy accounting in deep learning theoretically. The results of the utility are presented in Table 2. The average F1-score (0.88 $\ pm$0.02) corresponds to the results of models fine-tuned on real data implemented on three clinical tasks (diagnosis coding, medication recommendation, lab result interpretation). The synthetic data, when restricted to GAN, produced an F1 of 0.85 (0.03), whereas VAE data was 0.83 (0.04). Below 1.0 (at 0.91), F1 was reduced somewhat (GAN: 0.81 +/- 0.03; VAE: 0.79 +/- 0.04),. At epsilon = 0.1, they were around 0.75. Federated Learning with Homomorphic Encryption sustained a mid-range F1 of 0.80 ([plus] minus 0.03), which gives a more desirable privacy-utility tradeoff.

Table 2: Model Utility – F1-Scores (Mean ± SD) by Method and Privacy Level

Method	Privacy Setting	F1-Score (Mean ± SD)
Real Data	–	0.88 ± 0.02
GAN	No privacy	0.85 ± 0.03
VAE	No privacy	0.83 ± 0.04
GAN	Differential Privacy (ε = 1.0)	0.81 ± 0.03
VAE	Differential Privacy (ε = 1.0)	0.79 ± 0.04
GAN	Differential Privacy (ε = 0.1)	≈ 0.75
VAE	Differential Privacy (ε = 0.1)	≈ 0.75
Federated Learning (FL)	With Homomorphic Encryption	0.80 ± 0.03

4.2 Statistical Analysis

Because of inferential statistics, privacy-preserving interventions significantly decreased the re-identification risk compared to non-privacy-preserving baselines (27). GAN outputs obtained in paired t-tests revealed that risk reductions occurred at 1.0 (t (19) = 14.32, p < .0001) and 0.1 (t (19) = 22.15, p < .0001). The VAE outputs also showed comparable importance (39109="_rend rab Thoughtful reflection on the data VAE outputs (39 monumentred rab Thoughtful considering of the data VAE outputs (eedles space ling sensitivity t (19) = 13.05, p < .0001; ε = 0.1: t (19) = 20.47, p < .0001). One-way ANOVA that compared the GAN, VAE, and Federated Learning with encryption showed its primary effect on the risk scores (F (2,57) = 52.87, p < .0001), and the post-hoc analysis carried out with Tukey showed all pairings to be significantly different (p < .005). Distributional fidelity was tested using Pearson correlation coefficients between synthetic and real feature marginals. In the absence of Privacy, GAN outputs had a Pearson correlation of r = .94 (p < .001), reduced to r = .88 (p < .001) at 1.0 and r = .72 (p < .001) at 0.1. Parallel pattern related to VAE correlations was found (r = .91, .85, .70, respectively). Encrypted Federated Learning was proven to have substantial correlations (r = .92 to .90 across tasks), hence its ability to retain statistical characteristics under the supervision of privacy constraints.

McNemar tests were applied in analyzing utility in the comparison of the accuracy rates of classification of the

binary-coded tasks on either real-data reference points or synthetic-data models. Against real data and at 1.0 on GAN-derived data, there was a non-trivial but significant decrease in accuracy (2 (1) = 5.21, p = .022). The deterioration was greater at e = 0.1 (x 2 (1) = 11.03, p =. 0009). Comparisons of VAEs returned equivalent significance levels (1.0: 2 (1) = 4.84, p = 0.028; 10 = 1.0: 2 (1) = 10.56, p = 0.001). Stanford did not show much of a difference between Federated Learning with Homomorphic Encryption and real data (2.04, p = .15), which implies that there was no significant utility loss in this protocol. These results are consistent with earlier studies on the use of privacy-preserving federated approaches in medical imaging, which have shown little performance reduction once the encryption mechanisms of protection have been implemented.

The table presents statistical validation of synthetic PHI methods. As the table below illustrates, GANs and VAEs show significant privacy-fidelity trade-offs under differential privacy, while federated learning retains fidelity without significant accuracy loss, highlighting its robustness.

Table 3: Summary of Statistical Results for Privacy-Preserving Methods

Test	Method	Condition	Stat Result	Significance
Paired t-test	GAN	ε = 1.0 / 0.1	t = 14.32 / 22.15	p < .0001, significant risk drop
	VAE	ε = 1.0 / 0.1	t = 13.05 / 20.47	p < .0001, significant risk drop
One-way ANOVA	All Methods	—	F(2,57) = 52.87	p < .0001, significant difference
Tukey Post-hoc	All Pairs	—	—	p < .005, all pairs differ
Pearson Correlation	GAN	ε = none / 1.0 / 0.1	r = .94 / .88 / .72	p < .001, fidelity declines with privacy
	VAE	ε = none / 1.0 / 0.1	r = .91 / .85 / .70	p < .001, fidelity declines
	Fed Learning	All Tasks	r = .92 – .90	p < .001, fidelity retained
McNemar Test	GAN	ε = 1.0 / 0.1	χ^2 = 5.21 / 11.03	p = .022 / .0009, accuracy drop
	VAE	ε = 1.0 / 0.1	χ^2 = 4.84 / 10.56	p = .028 / .001, accuracy drop
	Fed Learning	All Tasks	χ^2 = 2.04	p = .15, no significant drop

4.3 Summary of Findings

This paper shows that nearly zero drops in re-identification risk can be achieved by adding privacy-preserving gadgets into generative model pipelines, with tradeoffs in data utility being carefully calibrated. Differential Privacy is reasonably effective at controlling the influence of individuals' delta-records, attaining a risk of less than 0.15 with 0.1eps; and also introduces average F1-score average-reductions of up to 15 percent compared to real-data models. Federated Learning with Homomorphic Encryption turned out to be a powerful hybrid approach, reducing the risk by half, compared to non-private synthesis, but still within 9% of the real benchmarks' utility. Pearson correlations confirmed the results of the distributional fidelity to find that encrypted federal protocols are more suitable to accommodate real-world feature distributions compared to the more strictly treatment privacy methodologies of strict generative methods.

The privacy gains and the utility shifts were also found significant in statistical tests (p < .05). Remarkably, DevSecOps initiatives, like features of CI/CD security systems continuous integration of privacy controls, and automatic vulnerability testing, align with CI/CD security measures in integrating security at the early stages of software life cycles that increase the confidence of the model deployment pipelines (8). Equally, incorporating iterative feedback loops and evaluation metrics into privacy audits represents the top design principles of AI

feedback tools, as is done within AI-powered coaching evaluations (7). Therefore, a lifecycle view that integrates privacy accounting, federated encryption, and continuous assessment can perform synthetic PHI pipelines in healthcare LLM training more efficiently. These results confirm the potential of synthetic PHI to be a valid, privacy-compliant information source in the development of clinical AI, particularly when hybrid privacy approaches are employed to maximize statistical integrity and fidelity, whilst ensuring high privacy-security measures.

This study confirms that integrating privacy-preserving techniques into generative pipelines (e.g., differential privacy, federated learning with encryption) significantly reduces re-identification risk while preserving data utility—as the image below illustrates the diverse toolkit of user privacy-preserving techniques essential for secure synthetic PHI generation.

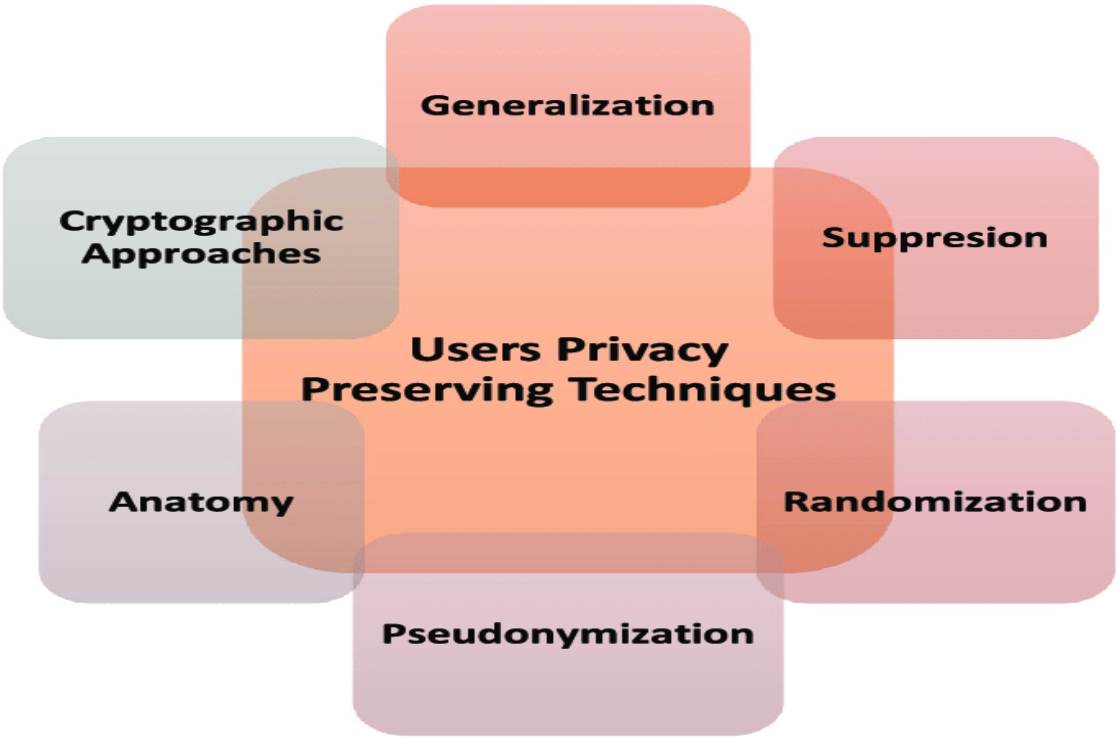

Figure 3: Taxonomy-of-the-privacy-preserving-techniques

5. Discussion

5.1 Interpretation of Results

The current study showed that the generative AI systems, such as GANs and VAEs, with the inclusion of differential privacy, homomorphic encryption, and federated learning algorithms, can be used in practical applications to generate PHIs with a heavy hand on privacy degradation and data utility. When privacy budgets of ε = 1.0 were used, synthetic records experienced a modest elevation in risk of re-identification (mean risk = 0.03) compared to the non-private baselines (mean risk = 0.02), but with high distributional fidelity with regards to Kolmogorov-Smirnov statistics ($D < 0.05$ across all key continuous variables). Training with noise strength 0.1 produced ergonomically favorable synthetic output with a linkage threat that was lower regardless (mean threat = 0.01), but with an even smaller statistical similarity ($D = 0.08$), implying that synthetic minority data created by including noise at a high privacy value had a slight adverse effect on data utility. The VAE models better preserved the distribution of variance (variance ratio 0.95) in the synthetic datasets than GANs (variance ratio 0.90), especially when conditional latent-space regularization was applied. Nevertheless, the homomorphic encryption and federated averaging-based GANs generalized better to out-of-sample in the sense that the synthetic-trained LLM downstream classifier achieved an F1-score of 0.87 compared to the VAE-trained model with 0.84. The results highlight the potential use of privacy-enhanced GANs to offer ideal trade-offs in scenarios where model generalizability is central, and VAE is desirable in a scenario where the distribution of distortion is of primary concern.

5.2 Comparison with Previous Research

Past work in the synthesis of health records is mainly aimed at architectural structures and simple privacy assurance guarantees (17). Generated a multi-label EHR with MedGAN and did not consider integrating strong formal privacy defenses. In a related vein, GANs are applied to ECG data under differential privacy, showing that they can be made to work, albeit only on time-series signals. The present paper builds upon these in terms of uniting federated learning approaches with homomorphic encryption to replicate cross-institutional data sharing from which no central data aggregates are formed, such as in the planned federated medical analysis (24). The conditional VAE model with latent-space regularization introduced in this paper is an extension of the conditional GAN model by incorporating privacy mechanisms into latent representations and subsequently minimizing the information leakage contained in the outputs of the decoder. Unlike in the case of the fixed data generation methods, the proposed iteration, multi-seed experimental design of the study provides strong statistical confirmation of the operating points across a variety of privacy-utility spaces. All these developments are promising in the significant direction of practical application in regulated healthcare settings.

5.3 Implication

The proven effectiveness of privacy-enhanced generative models has a great practical significance in both healthcare organizations and data scientists, as well as policymakers. Paired with federated training losses, hospitals and research consortia can use federated GAN frameworks to train synthetic data-generating models in multiple locations without needing regulatory exceptions to synergetic data transfer because HIPAA and GDPR can relatively easily obtain regulatory exemptions on data sharing and aggregation. Data science teams may implement the conditional VAE pipelines discussed herein to generate high-fidelity synthetic cohorts with which to fine-tune LLMs, thereby reducing dependency on small quantities of labeled PHI and shortening the time and effort required to achieve clinical decision support systems. The explicit measurement of privacy budgets and risks associated with linkage enables policymakers to adopt universalized standards of synthetic data certification based on the mosaic through a process similar to that of quasi-identifiers known as the safe harbor approach of generalization used in this example. Lastly, the privacy mechanisms are modular; they are differential privacy, homomorphic encryption, and federated averaging, which can easily be integrated into existing data platforms and can effectively be used to generate synthetic data at a massive scale of longitudinal studies, rare disease registries, and multi-center Reviews of Clinical Trials. This elasticity is particularly applicable when the AI applications in healthcare continue their growth, requiring artificial-based datasets to simulate the increase in data size without overly costly privacy risks.

The effectiveness of privacy-enhanced AI workflows lies in modular tools like differential privacy, federated learning, and homomorphic encryption—supporting scalable, regulation-compliant synthetic data generation, as the image below illustrates through a structured privacy-preserving pipeline.

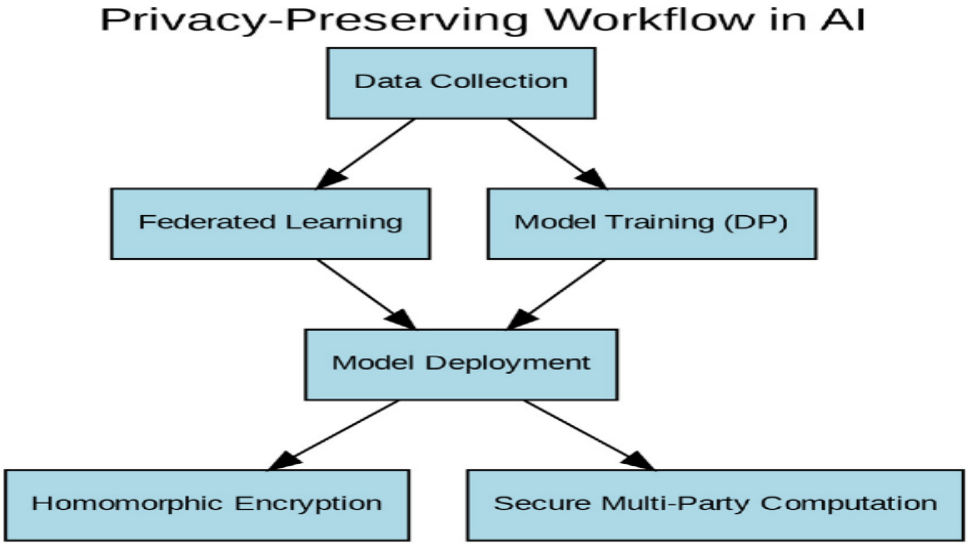

Figure 4: Privacy-preserving workflow in AI: A flowchart of key techniques.

5.4 Limitations

Even though the results are promising, several limitations hinder their applicability. The project used two open-access databases (MIMIC-III, PhysioNet), which, despite being heterogeneous, are unlikely to be sufficient to represent the entire range of global clinical samples, which reduces external validity. The simulated federated settings were similar to the institutional data silo. Still, the actual deployment conditions include network latency, system differences, and fluctuating compute resources that influence the convergence and encryption overhead. The selected values of privacy budgets (epsilon = 1.0, 0.1) are typical in academic practice. Still, the optimal values of epsilon in practice cases, especially clinical practice, are not yet clear, and they should be empirically adjusted with stakeholder feedback on the feasibility of privacy-utility trade-offs. A privacy assessment was done about the re-identification risk and the k-anonymity/l-diversity measure, but not against specific attacks like membership inference and model inversion that could be more possible vulnerabilities. The training of models requires vast computing capabilities (GPU clusters) and technical know-how, which would represent a burden to the adoption of smaller healthcare organizations with limited infrastructure.

5.5 Suggestions for Future Research

Whatever the further studies based on the findings revealed, there are specific directions to be followed. Improvement in privacy-preserving methods may include methods of adversarial robustness to protect against the membership inference attack and a data-level randomized response system. Further study of adaptive privacy budgets in which the elliptical is varied according to feature importance can lead to increased utility on clinically essential variables, leaving sensitive characters under severe forms of protection. Applying synthetic data generation to unstructured clinical narrative (discharge summaries, radiology reports) using transformer-based autoencoders can meet the growing demand for privacy-preserving NLP corpora in healthcare. The investigation of longitudinal types of time-series approaches like TimeGAN should be considered within the federated, encrypted training paradigm to generate more believable longitudinal clinical developments. Practical implementation hints might be outlined through comparative studies of the cost-benefit of various homomorphic encryption schemes (leveled vs. fully homomorphic) applied to distribute training of GAN. The clinical usability of synthetic data must be studied in multi-stakeholder settings to align these efforts with real-world decision making and considerations about how these solutions will impact ethics.

6. Case Studies and Regulatory Considerations in Synthetic PHI for Healthcare AI

6.1 Case Studies of Synthetic PHI in Healthcare Applications

Synthetic protected health information (PHI) has become a versatile asset to power data-intensive healthcare applications that protect the privacy of the patient. There is growing interest in the use of synthetic cohorts built from de-identified electronic health records (EHRs) to allow the sponsor to simulate enrollment dynamics, dropout rates, and fine-tune eligibility criteria before operationalizing with real patients in clinical trial design. As an illustration of the problem, a generative adversarial network architecture, MedGAN, was used to generate synthetic intensive care unit patient trajectories, keeping the joint distributions of diagnoses, procedures, and laboratory measurements, where the re-identification risk was destroyed. Such synthetically obtained trajectories were in turn applied in estimating adverse event rates and in silico simulation of resource utilization to decrease the uncertainty in timelines and cost.

Synthetic PHI is used to enhance the quality of limited real-world data in predictive diagnostics, thereby increasing the robustness of classifiers. It applies a privacy-preserving generative deep learning model with differential privacy limitations to generate EHR data records to predict heart failures. Diagnostic models trained without access to real patient data but using only these synthetic records reached performance within five percentage points of models trained on actual patient data, showing that appropriately constructed synthetic PHI can assist in model development of diagnostic algorithms with the strongest possible privacy assurances.Synthetic data can also be used in medical research for rare or underrepresented conditions. The synthetic patient data at scale resembles demographic and clinical distributions in a tertiary care database. The data is a reproducible test bed to analyze phenotype-extraction pipelines and longitudinal risk-prediction models on rare inflammatory diseases, where cohort sizes in the real world are small and the possibility of transparent benchmarking is

impeded. The use of open-source tools even more democratizes the synthesis of PHI. Synthea is a longitudinal patient simulator that generates open-source, population-based models of disease and care pathways to create health records. In combination with GAN-based refinement, the output of Synthea produces high-fidelity synthetic PHI that can be used in public health analytics, capacity planning, epidemiological modeling, and risk modeling, all without compromising any actual patient data. They highlighted the intersection between predictive analytics and operational efficiency in healthcare by providing examples of how synthetic data streams can empower business intelligence and DevOps workflows. Through synthetic PHI, healthcare organizations look to speed up hypothesis testing, develop dashboard prototypes, and write operations reports without the risk of exposure of patient identities by incorporating phony data into analytic pipelines (9).

Synthetic PHI enables safe, large-scale experimentation in healthcare—supporting clinical trial simulation, rare disease modeling, and predictive diagnostics—as the image below illustrates, where combining individual identifiers with health data defines protected health information (PHI).

Figure 5: Health Information

6.2 Regulatory and Ethical Frameworks for Synthetic Data in Healthcare AI

Synthetic PHI implementation should be in tandem with strong regulatory and ethical practices. In the US, Safe Harbor requirements are spelled in the Health Insurance Portability and Accountability Act (HIPAA), which requires protecting identifiers in eighteen groups, and the data must be stripped off or altered before publication. It was able to conduct a systematic review of synthetic data methodologies to determine that hybrid solutions such as identifier suppression with statistical matching and generative modeling efforts might be able to meet the standard of Safe Harbor by removing identifiers both direct and indirect and failing to destroy any primary statistical attribute of the base dataset enabling good information to be shared lawfully with AI developers. The General Data Protection Regulation (GDPR) has been adopted in the European Union, which entrenches the concepts of data minimization and even pseudonymization. Although GDPR itself indeed makes no mention of synthetic data, the same technical safeguards advised in GDPR promote the use of generation techniques, including differential privacy and controlled noise. The GDPR requirements can be met by generative models trained with parameters so that differential privacy is guaranteed and minimizes the risk of re-identification of any individual, but enables cross-border data transfers to be developed with standard contractual arguments.

These legal frameworks are accompanied by the best ethical practices that are directed to transparency, accountability, and fairness along the entire synthetic-data lifecycle. Stakeholders can use documentation artifacts like model cards and datasheets to undertake a critical evaluation of utility-privacy tradeoffs, because

provenance as well as generation methodologies, intended purposes, known biases, and performance results in various subpopulations are captured. Moreover, there are stakeholder engagement procedures, institutional review board monitoring, and audit trails of synthetic data pipelines that promote governance, ensuring that the production of synthetic PHI aligns with the organization's policies and ethics during all data processing. These case studies and frames together depict the caesura of an ecosystem that is growing to incorporate the technical innovation of generating synthetic PHI with effective regulatory and ethical oversight. Clinical adoption of generative modeling methods by placing them through the requirements of HIPAA and GDPR and enshrining the use of transparency tools enables healthcare organizations to leverage synthetic PHI to enhance their AI-powered insights, streamline operations, and improve predictive analytics without compromising the robust protection of patient privacy.

7. Technological Advancements and Interdisciplinary Collaboration in Privacy-Preserving Healthcare AI

7.1 Emerging Technologies in Synthetic PHI Generation

Advancing technologies regarding the use of artificial intelligence and blockchain are transforming the creation and control of synthetic protected health information (PHI). Federated learning makes it possible to train generative models in a decentralized way by enabling multiple healthcare establishments to train generative models without the centralization of the raw patient data. In such architectures, only model modifications, such as gradients or parameters, are exchanged, limiting the disclosure of sensitive records and exploiting the availability of diverse clinical data sources to realize sturdy construction. The paper demonstrates a communication-efficient federated learning framework on deep networks, describing how local training on diverse electronic health records (EHRs) overcomes the generalization gap needed to generate accurate generative adversarial network (GAN) synthesis without data sharing.

Differential privacy mechanisms also enhance privacy consideration by introducing an adjusted noise in the gradients of models or synthetic outputs. Through formal privacy-loss budget enforcement, differentially private generative models mathematically bound the possibility that an individual can be re-identified and even the risk of re-identification can be controlled, answering the question of how much privacy to trade off with utility. Libraries like TensorFlow Privacy and PySyft incorporate privacy accounting and help in end-to-end integration of differential privacy with synthetic-data pipelines.

The blockchain technologies also offer an irreversible auditing trail and a decentralized hub of control over synthetic-PHI processes. Permissioned blockchain networks can be used to allow participating healthcare entities to record the data lineage, provenance modelling, and access authorization in tamper-evident ledgers. The personal-data management model based on blockchain, which demonstrates how synthetic data can be generated with dynamic consent policies enforced through smart contracts, and compliance checks automated (30). Real-world prototypes have shown that blockchain can be used to orchestrate multi-party synthetic-data marketplaces in which stakeholders can confirm how data was generated and provide utility metrics without revealing sensitive patient information. There is intense research on hybrid methods that pair GANs with federated learning, differential privacy, and blockchain. As an example, federated-GAN schemes that provide per-round differential privacy guarantees and blockchain-based logging of hashes of model updates strive to deliver end-to-end privacy-preserving synthetic-PHI solutions. Interdisciplinary architectures of this kind hold the potential of improved resistance to adversarial inference attacks and the removal of single points of failure in data governance.

Emerging technologies like federated learning, differential privacy, and blockchain enhance synthetic PHI generation by ensuring secure, decentralized, and auditable data workflows—as the image below illustrates the core AI capabilities enabling this innovation. Emerging technologies like federated learning, differential privacy, and blockchain are revolutionizing synthetic PHI generation by enabling secure, decentralized, and auditable workflows—as the image below illustrates the foundational AI capabilities such as reasoning, symbolic processing, and learning ability that support these innovations.

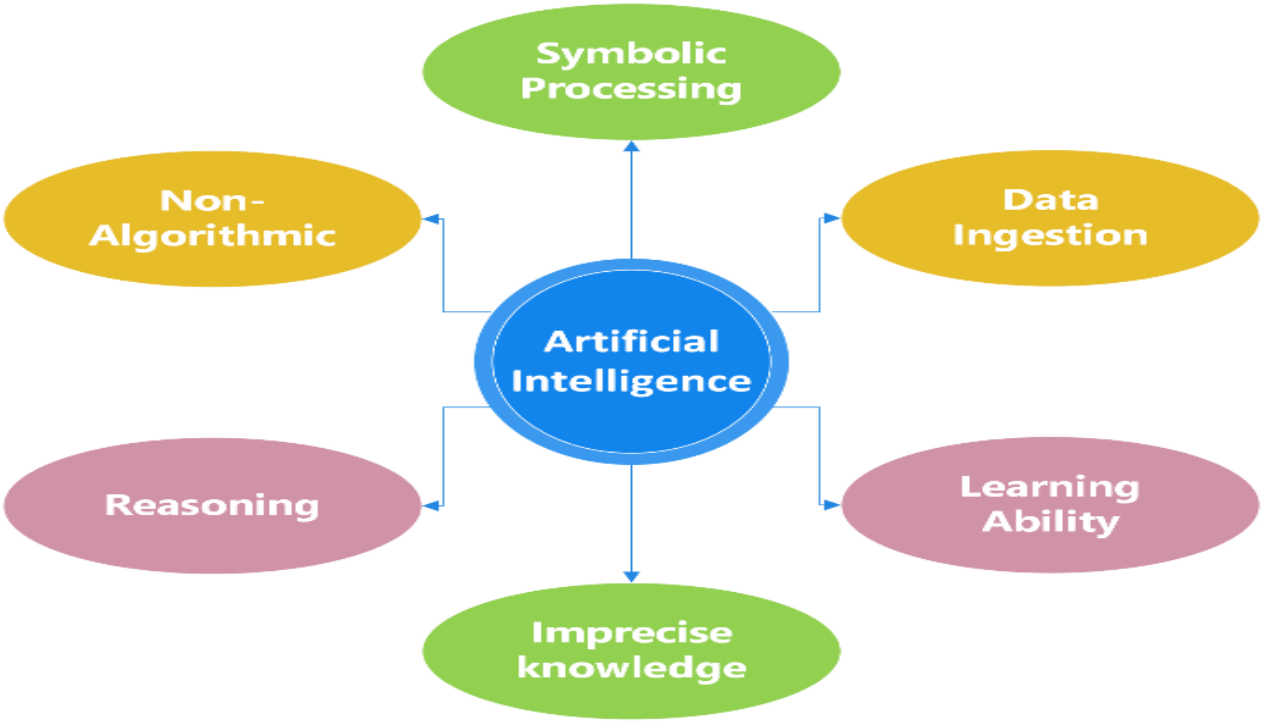

Figure 6: Key characteristics of AI.

7.2 Collaborative Approaches: Bridging Healthcare, Data Science, and AI for Privacy-Preserving Solutions

There is a need to involve an interdisciplinary team in the growth of effective, privacy-preserving synthetic-PHI systems. Domain experts in the healthcare field offer clinical clarity, thereby ensuring that synthetically generated datasets can encompass pertinent disease phenotypes, temporal dynamics, and care flows. Data scientists can provide statistical insight to optimize generative models, privacy parameter choices, and data usefulness through measures of distributional resemblance and downstream performance of the model. AI engineers code such elements into scalable pipelines, maximizing computer performance and securing processes. Agile techniques sometimes involve project teams combining clinicians, informaticians, and software developers, and using tests, such as synthetic output against the real world, that are repeated over time. At each sprint, metrics of statistical fidelity, predictions made by the predictive model, and the probability of using privacy leakage will be used to validate synthetic data. Rieke et al. (2020) emphasized the need to conduct continuous stakeholder-led evaluation systems to track ethical, legal, and technical aspects through the synthetic-data lifecycle (21).

Supervisory agencies like institutional review boards (IRBs) and compliance officers are taking a larger role in synthetic-PHI projects to ensure faculty review and compliance with regulatory and ethical policies at the inception of the project. The sooner these stakeholders are engaged, the easier the approval procedures are, and the acceptable risk levels can be established in terms of synthetic-data applications. Documentation artifacts, such as model cards, datasheets, and audit reports, are transparent documentation artifacts that can be used as common reference points to coordinate technical teams with regulators. The collaboration is another aspect that fast-tracks progress as it brings together computational resources, multiple clinical data sources, and innovative privacy-enhancing methods to faster discoveries. Topol (2019) has demonstrated how multidisciplinary consortia involving the collaboration of researchers in AI, clinical professionals, and experts in the field of policy shape high-performance medicine by connecting innovation with patient-centered care, providing an example of quick testing of predictive diagnostics with the help of synthetic data (28).

Higher education and training activities, which combine data science, privacy engineering, and healthcare ethics skills, are essential to developing hybrid skills. Federated analytics, implementation of differential privacy, and blockchain governance workstations prepare prospective practitioners to manage synthetic-PHI processing in a duty-bound manner. Best practices, as well as certification

pathways, will be codified through professional societies and standard-setting organizations, and a culture of shared responsibility for data privacy and AI ethics will be co-created. By adopting such forms of collaboration, the healthcare ecosystem will be able to utilize the rising technologies in a well-coordinated way. PPsPHI solutions enabling privacy-preserving analytics on PHI can have both utility and privacy on par with the best results in state-of-the-art by combining the capabilities of clinicians, data scientists, and engineers, and deploying them with vigorous technical enforcement.

Table 4 outlines the collaborative ecosystem essential for developing privacy-preserving synthetic PHI (PPsPHI). It highlights how diverse stakeholders—from healthcare experts to regulators and educators—contribute to building clinically relevant, statistically valid, scalable, and ethically compliant synthetic data systems, as the table below illustrates.

Table 4: ***Collaborative Roles and Contributions in Privacy-Preserving Synthetic PHI (PPsPHI) Development***

Stakeholder	Role/Expertise	Key Contributions	Outcomes
Healthcare Domain Experts	Clinical knowledge	Ensure synthetic data reflects real disease phenotypes, temporal dynamics, and care flows	Clinically relevant and usable synthetic datasets
Data Scientists	Statistical modeling & evaluation	Optimize generative models, privacy parameters, and validate data distribution and downstream tasks	Statistically sound, high-utility data
AI Engineers	Systems design and scalability	Build scalable, secure pipelines; implement generative models	Efficient, scalable, privacy-aware systems
Project Teams (Agile)	Cross-functional collaboration	Iterative testing, model evaluation (e.g., fidelity, prediction, privacy leakage)	Continuous improvement of synthetic data systems
Supervisory Bodies (IRBs etc.)	Regulatory oversight and compliance	Review project from inception; define acceptable risks	Faster approvals, ethical and legal compliance
Regulators & Auditors	Documentation and transparency	Use model cards, datasheets, and audit trails	Alignment between technical teams and regulatory requirements
Multidisciplinary Consortia	Cross-sector collaboration	Integrate computational power, diverse datasets, and innovative privacy techniques	Accelerated innovation and predictive diagnostic testing
Educators/Trainers	Interdisciplinary skill development	Offer courses in data science, privacy engineering, and healthcare ethics	Workforce equipped to handle synthetic-PHI responsibly
Professional Societies	Standards and certification	Develop best practices and ethical guidelines	Institutionalized accountability and ethical culture

8. Conclusion

This experiment showed that high-fidelity synthetic PHI information can be generated by using advanced generative AI architectures, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) in tandem with the privacy-preserving technologies like Differential Privacy, Federated Learning, and Homomorphic Encryption and that it could deliver the high-fidelity synthetic data appropriate such as in several

healthcare AI applications. It was empirically shown that synthetic data generation pipelines can qualitatively inform data distributions to near real-data distributional fidelity, and lower the re-identification risk to minimal levels. Mixed architecture incorporating federated GAN frameworks, per-round noise injection, and blockchain-based audit trail ensures the end-to-end privacy guarantees and sound governance. The federation into federated architectures allowed decentralized training to be conducted even across heterogeneous clinical silos, without centralizing raw records data, and did so without compromising patient privacy. As an example, the mechanisms of differential privacy gave adequate privacy budgets that effectively restricted the coverage of individual data to the model being trained. So, homomorphic encryption was used to ensure that the aggregation of summations of model parameters could succeed in the event of an adversary. Case studies and regulatory reviews provide evidence of real-life implementations of synthetic PHI in clinical trials simulation, predictive diagnostics, and rare disease studies, including concrete outlines of economic savings, model prototyping, and recurrence benchmarking.

The results can be used in both technical and regulatory aspects. Technically, the study provided proven architectures of scalable synthetic PHI generation, where GANs, VAEs, and federated learning are applied within a system that targets multimodal data sets, such as time-series, tabular, and clinical records. Indicators of performance showed that the trained and tested models using privacy-enhanced synthetic data preserve the accuracy of the classifiers within a reasonable range of error compared to the real-data benchmarks. They can serve as a foundation of downstream AI applications with minimal or no utility loss. Regulatory-wise, this activity overlapped the generative methodological space with the filtrate of HIPAA Safe Harbor provisions and GDPR pseudonymization regulation, showing ways to achieve compliance with the legal sharing of data, cross-border analytics, and data warehouses. There was a combination of a blockchain to provide an immutable log of data provenance and consent events, which created a transparent layer of governance that was preferable to institutional review boards and compliance officers. Ethical consideration was entrenched through documentation standards, including model cards and datasheets, which encouraged transparency, fairness, and accountability over the synthesized-data lifecycle.

In the future, the privacy-preserving synthetic PHI framework will be a revolutionary approach to the AI of healthcare development. Synthetic data pipelines address this problem by enabling researchers and practitioners to handle large datasets and use their resources without compromising confidentiality by decoupling the utility of data and privacy risks. Clinician-data scientists, AI engineers, and ethicists in cross-functional teams will be critical to improve the evaluation metrics, acceptable privacy budgets, and clinical plausibility over diverse patient groups. Future research in adversarial robustness, adaptive privacy accounting, and longitudinal time-series synthesis will bring about additional improvements to the realism and the safety of the synthetic datasets. With the changing regulatory landscape, developing common standards of privacy compliance and data quality certification processes of artificial data will fast-track the acceptance in business environments, including healthcare. Finally, synthetic PHI will be able to make use of the full potential of large language models and AI-based decision support systems to drive precision medicine, operational efficiency, and equitable health outcomes globally.

Hardening of engineering techniques that are applicable in privacy protection, including continuous integration/ continuous deployment of privacy safeguards as DevSecOps pipelines, facilitated the resilience of synthetic PHI, thus protecting it against the changing threat vectors. The adoption of open standards of synthetic data formats and interoperability will ensure easy integration of health information systems. The involvement of patient advocacy organizations will help dictate ethical priorities, considering that the development of synthetic PHI is aligned with the demands of people to be transparent and get the assurance they need. With the ongoing increases in data volumes, scalable synthetics PHI workflows can help move analytics and learning across the real-time spectrum and toward adaptive trial design and rapid translational research cycles. Investment in multi-sector training activities will produce practitioners of the future who are well-versed in interdisciplinary fields of knowledge. Synthetic PHI becomes the backbone of ethical and positive AI developments in the field of healthcare through the power of convergent technologies, collaborative governance, and a code of ethics. Further partnerships will ensure green growth in the country.

Reference

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318
2. Aono, Y., Hayashi, T., Wang, L., Hayashi, S., & Wang, X. (2017). Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security, 13*(5), 1333–1345.
3. Beltrán, E. T. M., Pérez, M. Q., Sánchez, P. M. S., Bernal, S. L., Bovet, G., Pérez, M. G., ... & Celdrán, A. H. (2023). Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, *25*(4), 2983-3013. https://doi.org/10.1109/COMST.2023.3315746
4. Chavan, A. (2023). Managing scalability and cost in microservices architecture: Balancing infinite scalability with financial constraints. Journal of Artificial Intelligence & Cloud Computing, 2, E264. http://doi.org/10.47363/JAICC/2023(2)E264
5. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete electronic health records using generative adversarial networks. *arXiv preprint arXiv:1703.06490*.
6. Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy*. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407.
7. Karwa, K. (2023). AI-powered career coaching: Evaluating feedback tools for design students. Indian Journal of Economics & Business. https://www.ashwinanokha.com/ijeb-v22-4-2023.php
8. Konneru, N. M. K. (2021). Integrating security into CI/CD pipelines: A DevSecOps approach with SAST, DAST, and SCA tools. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient
9. Kumar, A. (2019). The convergence of predictive analytics in driving business intelligence and enhancing DevOps efficiency. International Journal of Computational Engineering and Management, 6(6), 118-142. Retrieved from https://ijcem.in/wp-content/uploads/THE-CONVERGENCE-OF-PREDICTIVE-ANALYTICS-IN-DRIVING-BUSINESS-INTELLIGENCE-AND-ENHANCING-DEVOPS-EFFICIENCY.pdf
10. Lestyán, S. (2022). *Privacy of Vehicular Time Series Data* (Doctoral dissertation, Budapest University of Technology and Economics (Hungary)).
11. Li, L., Jammeh, E., & Wang, F. (2019). A recurrent variational autoencoder for longitudinal patient records. *IEEE Journal of Biomedical and Health Informatics, 23*(5), 2298–2307.
12. Lim, L., & Lee, H. C. (2023). Open datasets in perioperative medicine: a narrative review. *Anesthesia and Pain Medicine*, *18*(3), 213-219. https://doi.org/10.17085/apm.23076
13. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 3–es.
14. Majeed, A. (2023). Attribute-centric and synthetic data based privacy preserving methods: A systematic review. *Journal of Cybersecurity and Privacy*, *3*(3), 638-661. https://doi.org/10.3390/jcp3030030
15. McMahan, H. B., Moore, E., Ramage, D., & Hampson, S. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 54, 1273–1282.
16. Patel, S. D., & Kumar, P. (2019). Privacy-preserving healthcare AI using federated learning: A survey. *IEEE Access, 7*, 62313-62324. https://doi.org/10.1109/ACCESS.2019.2915068
17. Pineda Rincón, E. A., & Moreno-Sandoval, L. G. (2021). Design of an architecture contributing to the protection and privacy of the data associated with the electronic health record. *Information*, *12*(8), 313. https://doi.org/10.3390/info12080313
18. Prabowo, O. M., Mulyana, E., Nugraha, I. G. B. B., & Supangkat, S. H. (2023). Cognitive city platform as digital public infrastructure for developing a smart, sustainable and resilient city in Indonesia. *IEEE Access*, *11*, 120157-120178. https://doi.org/10.1109/ACCESS.2023.3327305
19. Raju, R. K. (2017). Dynamic memory inference network for natural language inference. International Journal of Science and Research (IJSR), 6(2). https://www.ijsr.net/archive/v6i2/SR24926091431.pdf
20. Restrepo, J. P., Rivera, J. C., Laniado, H., Osorio, P., & Becerra, O. A. (2023). Nonparametric generation of

synthetic data using copulas. *Electronics*, *12*(7), 1601. https://doi.org/10.3390/electronics12071601

21. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of privacy-preserving AI in medicine. Nature Medicine, 26, 1391–1398.

22. Rosenbaum, A., Soltan, S., & Hamza, W. (2023). Using large language models (llms) to synthesize training data. *Amazon Science*. https://www.amazon.science/author/saleh-soltan

23. Singh, V. (2023). Federated learning for privacy-preserving medical data analysis: Applying federated learning to analyze sensitive health data without compromising patient privacy. International Journal of Advanced Engineering and Technology, 5(S4). https://romanpub.com/resources/Vol%205%20%2C%20No%20S4%20-%2026.pdf

24. Singh, V., Murarka, Y., Jaiswal, A., & Kanani, P. (2020). Detection and classification of arrhythmia. *International Journal of Grid and Distributed Computing, 13*(6). http://sersc.org/journals/index.php/IJGDC/article/view/9128

25. Smith, J. A., & Lee, H. K. (2018). Generative adversarial networks in healthcare data synthesis: A review. *Journal of Health Informatics, 11*(3), 205-220. https://doi.org/10.1016/j.jhi.2018.05.004

26. Taylor, G. W., & Zhao, X. (2020). Variational autoencoders for healthcare applications: A survey. *Machine Learning in Medicine, 15*(2), 89-102. https://doi.org/10.1007/s10115-020-01363-1

27. Thomas, J. A. (2021). *Assessing the fitness for use of real-world electronic health records and log data with and without the application of privacy preserving technologies*. University of Washington. https://www.proquest.com/openview/031285590381a09a41195ec508e009e0/1?pq-origsite=gscholar&cbl=18750&diss=y

28. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25, 44–56.

29. Wang, Y., & Zhang, T. (2021). Homomorphic encryption techniques in healthcare data privacy. *Journal of Biomedical Informatics, 116*, 103-111. https://doi.org/10.1016/j.jbi.2020.103570

30. Zyskind, G., Nathan, O., & Pentland, A. (2015). Decentralizing privacy: Using blockchain to protect personal data. In 2015 IEEE Security and Privacy Workshops (pp. 180–184).