International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Synthetic data generation in healthcare: Using GANs to overcome data scarcity and bias in machine learning

Muhammad Faheem [1,*] and Aqib Iqbal [2]

[1] *Department of Information Technology Management, Cumberland University, Lebanon Tennessee United States.*
[2] *Department of Project Management, University of Law, Birmingham, United Kingdom.*

International Journal of Science and Research Archive, 2025, 16(01), 628-639

Publication history: Received on 31 May 2025; revised on 05 July 2025; accepted on 08 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2022

Abstract

The growing use of machine learning (ML) in healthcare is constrained by data scarcity, privacy regulations, fragmented data systems, and demographic imbalances. These limitations reduce model accuracy, hinder generalizability, and contribute to algorithmic bias, particularly affecting minority populations and underrepresented disease categories. Generative Adversarial Networks (GANs) have emerged as a promising solution by enabling the creation of synthetic datasets that preserve data utility while enhancing privacy and fairness. This paper explores the use of GAN-based synthetic data in addressing data limitations within healthcare ML pipelines. It examines key GAN architectures suited for structured clinical data, electronic health records (EHRs), and medical imaging, highlighting their training processes and privacy-preserving capabilities. Applications across clinical research, epidemiology, rare disease modeling, and privacy-conscious data sharing are reviewed. The paper further evaluates synthetic data quality using utility metrics, privacy risk assessments, and fidelity–privacy tradeoffs. While synthetic data offers transformative potential, challenges remain in GAN stability, ethical governance, and validation standards. Future directions include integrating federated learning, enhancing explainability, and advancing differential privacy to ensure ethical and inclusive AI development in healthcare.

Keywords: Synthetic Data; GANs (Generative Adversarial Networks); Healthcare; Machine Learning; Data Scarcity

1. Introduction

The healthcare sector using data has witnessed an exponentially increasing rate of growth which led to the implementation of machine learning (ML) models into clinical diagnostics, predictive analytical medicine, personalized healthcare, and exploring public health. Nevertheless, even though these advancements have been made, lack of information and the existing bias in the available high-quality data, remains one of the most urgent areas of concern in the field of ML implementation into the healthcare industry. Health data is very sensitive with strict privacy laws like Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. Moreover, health data is gravely fragmented at various institutions resulting in variable, sparse, dysfunctional data environments. Machine learning models cannot be generalized and fair because of a dearth of sufficiently large, varied, and representative datasets. Algorithmic bias can be eliminated by addressing bias in the data, whether in the demographic, clinical, or socioeconomic nature. The use of the traditional methods of anonymization has been shown to be ineffective given that they lead to a considerable loss of information and yet they still have some risk of re-identification. Generative Adversarial Networks (GANs) as proposed by Goodfellow et al. (2014) have been shown as a promising method of creating synthetic data. GANs may generate realistic, but artificial, datasets after learning the joint probability distribution of real-world data. This technology has a potential to ameliorate data scarcity in healthcare, diminish bias, and improve privacy-preserving data sharing. This paper would attempt to evaluate these fundamental problems in healthcare machine learning comprehensively and list out their resolution by

* Corresponding author: Muhammad Faheem.

GAN-based synthetic data generation. A line of recent literature, such as the works by IEEE Transactions on Knowledge and Data Engineering, BMC Public Health, ACM Computing Surveys, European Journal of Radiology Health, and JAMIA Open, is used in this paper to shed light on methodological developments, assess practical use cases, and outline the potentiality and the limitations of the revolutionary method.

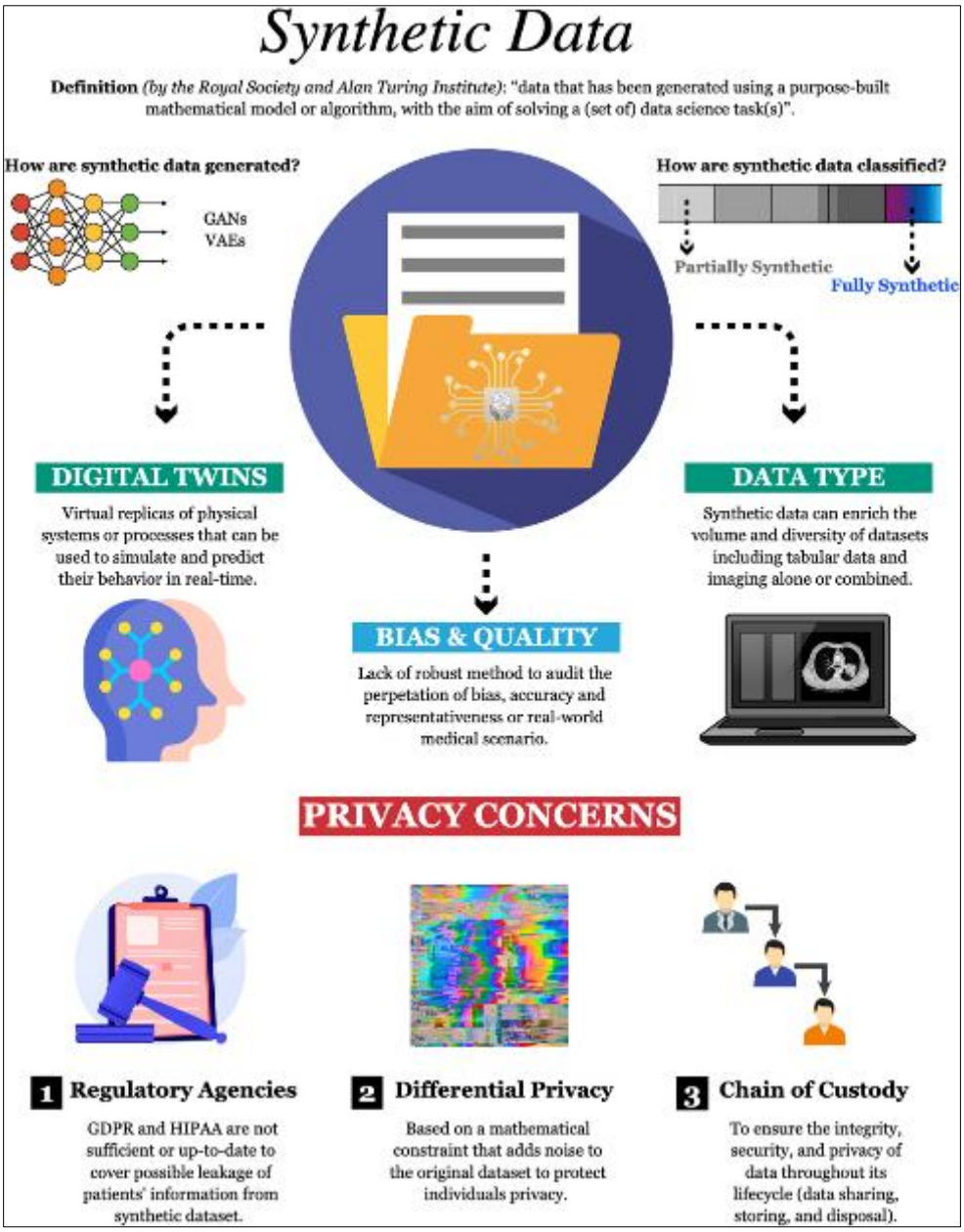

Figure 1 A conceptual diagram illustrating the problem of data scarcity and bias in healthcare ML pipelines and how synthetic data offers a solution

2. Background and Literature Review

2.1. Data Scarcity and Bias in Healthcare Machine Learning

Under the constraints of privacy and ethical dimensions as well as operational silos, healthcare data is frequently small-scale and disjointed. These data are often highly biased even when available. Some populations, like ethnic minorities, rural residents, or patients with a rare disease are less likely to be represented in electronic health records (EHRs) (Zhang et al., 2021). This prejudice might distort machine learning models and make them unfair and unsafe in AI-driven clinical tooling.

Adding the issue of data scarcity is the issue of data heterogeneity. Medical information comes in a variety of sorts and shapes--such as organized numerical inputs in EHRs but also high-dimensional imaging information such as MRI or CT scans, or unstructured clinical notes. This diversity makes this task even more challenging since it is hard to create viable models with a chance to generalize their findings to other populations of patients and even health systems.

2.2. Synthetic Data: Definition and Applications

Synthetic data is data that are created artificially and satisfying the same statistical characteristics of real data but not identifiable personal information. Synthetic data can be roughly divided into three main categories:

- Fully synthetic: Entirely fabricated data generated from learned distributions.
- Partially synthetic: A mix of real and synthetic values where sensitive fields are replaced.
- Hybrid synthetic: Combines synthetic records with real data subsets.

Synthetic data is largely applied in other spheres including finance, autonomous driving, and cybersecurity to expand datasets, perform simulations, and test algorithms in different settings. Synthetic data are promising in healthcare to break privacy walls, improve model training and solve the problems of underrepresentation.

Table 1 carries out a comparative analysis of real data, standard anonymization, and synthetic data on three dimensions of privacy, utility, and bias minimization.

Table 1 Comparison Between Real Data, Anonymized Data, and Synthetic Data in Healthcare

Feature	**Real Data**	**Anonymized Data**	**Synthetic Data**
Privacy Risk	High	Medium (can be re-identified)	Low (no real patients represented)
Data Utility	High	Medium (loss of granularity)	Medium to High (depends on model quality)
Bias Preservation	Retains real-world biases	Retains biases	Can correct or balance biases
Regulatory Barrier	High (requires consent/approval)	High	Low to Medium (fewer restrictions)
Reproducibility	Limited	Limited	High (can share without privacy concern)

2.3. Generative Models and GANs

GANs are made of two neural networks: the generator and the discriminator that play a minimax game. The generator is used to generate synthetic data based on random noise and the discriminator determines whether the data used is actual or artificial. By means of iterative adversarial training, the generator can fine-tune its output until the discriminator is not able to identify real and synthetic samples anymore.

GANs have been successfully adapted for various healthcare data types, including

- MedGAN: For generating high-dimensional discrete EHR data.
- ImageGAN: For synthesizing medical images like X-rays and MRIs.
- TimeGAN: For longitudinal patient data.

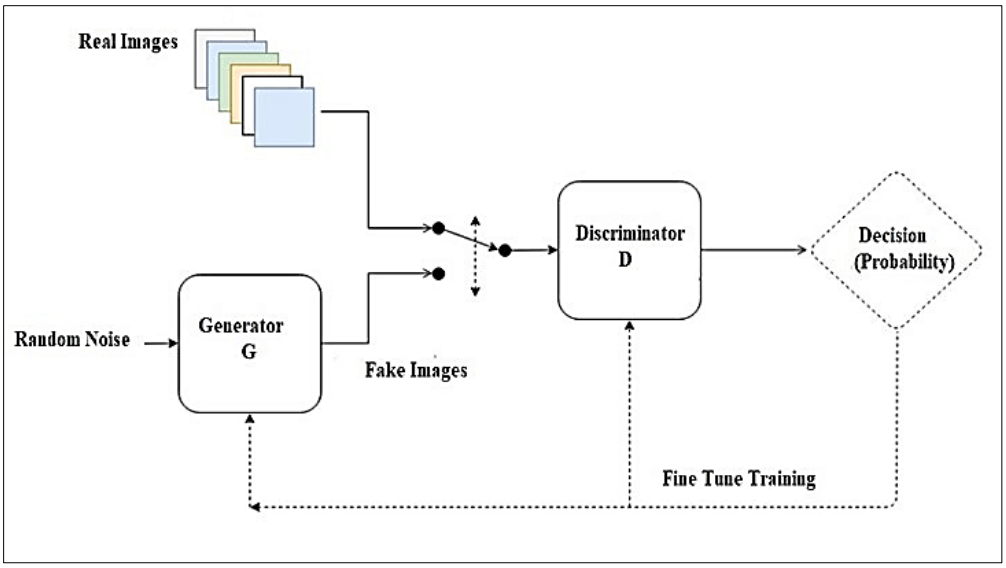

Figure 2 A typical GAN architecture showing the generator-discriminator workflow

3. Methodology: Synthetic Data Generation Using gans in healthcare

3.1. GAN Architectures Suitable for Healthcare Data

Standard GANs have difficulties working on some kinds of healthcare data especially tabular and longitudinal data. These challenges have been tackled by developing modified architectures

- Conditional GAN (cGAN): Allows data generation conditioned on class labels (e.g., age, disease type).
- Wasserstein GAN (WGAN): Uses Earth Mover's Distance to improve training stability.
- DP-GAN: Integrates differential privacy to ensure that generated data cannot be traced back to individual patients.

These models are designed to maintain the balance between data utility and privacy, making them suitable for healthcare applications where privacy preservation is paramount (Chen et al., 2021; Park et al., 2018).

3.2. Training GANs with Healthcare Data

The process of effective training of GANs in the task of generating healthcare data requires several essential preprocessing steps to prepare the data to fit the trained model and dynamic training process. Normalization is one of the basic procedures in which continuous variables that are numerical, are scaled to common range, which is usually between 0 and 1 or to a normal distribution. It is this standardization that will help to ensure that, all the input features are evenly employed in the training process to make sure that no single variable is employed with a larger magnitude to try to dominate the learning dynamics.

In the cases of data with categorical variables like diagnostic codes, type of medication or demographic terms, categorical encoding will be required. Popular methods are one-hot encoding or embedding-based representations. Encoding categorical features in a one-hot manner represents the feature in a binary vector space, whereas the embedding layers convert discrete features to continuous, dense vector space and preserve the semantic links between the categories. This practice is especially essential with the electronic health records (EHR) data since structured data such as gender, ethnicity, disease codes (e.g., ICD codes) and types of medication among others should be translated into the numerical form that can be effectively operated by the GAN.

Another essential preprocessing procedure is how to deal with the missing data. There is always an incidence of missing health data in healthcare datasets given the diversity in clinical application, patients, or the confusion of data collection. Ignoring missing data may bring a bias or deteriorate the performance of the GAN. The most common methods that the researchers used to counter it are imputation or replacement of missing values with statistical values such as the mean, median, or estimation via models or build the GAN in such a way that the model knows that missing information exists via mask vectors that are used during training.

As far as architecture is concerned, it all depends on the kind of GAN that will be used depending on the nature of the data. DCGANs are popularly employed in the case of medical imaging data. DCGANs exploit convolutional layers which can easily represent hierarchies and patterns of images. This facilitates creation of high resolutions, real looking synthetic medical images such as MRI scan, X rays, and CT images, which are essential in training diagnostic algorithms.

In case of EHR data that are expected to be mixed type and tabular, GAN architectures incorporate embedding layers that encode discrete variables where numeric representation (of diagnosis code or treatment type) is desired. These embeddings enable the GAN to learn semantic similarities and extraordinarily complicated relations between variables in structured medical records. The issues of high-dimensional tabular healthcare data being inherent have led to the development of specialized variants of GAN namely medGAN, tableGAN, and ctGAN.

Since healthcare data is very sensitive, privacy-preserving methods are essential during the GAN training procedure. Gradient clipping is one of such methods gradient clipping limits the magnitude of gradients used in backpropagation to ensure that the model does not over-fit to individual data points that may cause privacy leakage. Differential privacy which involves adding noise to gradients or model parameters during training is another popular method. This noise makes sure that the impact of the data of everyone on the trained model is mathematically limited and insignificant, which greatly diminishes the probability of re-identification or data leakage. Such privacy-preserving measures play a central role in reconciling synthetic data generation processes with legal regulations, GDPR and HIPAA, but with still high utility in downstream machine learning tasks.

3.3. Bias Mitigation with GANs

Through conditional GANs in which they condition the GANs on minority-labeled classes or underrepresented features, researchers have been able to explicitly predetermine the generation of their data or samples that will concentrate on creating synthetic data that reflects these underrepresented groups. This class conditioned method makes use of actual incorporated attributes such as classes labels, demographic and clinically conditions into the GAN. As an example, in case the given healthcare dataset has underrepresentation of some ethnicities, age groups, or patients with uncommon medical conditions, the GAN may be trained to produce more instances that will represent these groups. This targeted synthesis will help in dealing with the disparity which is very common in the actual health care data sets, which have majority classes dominating the training set.

The process will not only add realistic and varied cases of the marginalized population to the dataset but reduce the chances of algorithmic bias. The application of machine learning models in testing, prediction, and the prescription of treatments is biased because it under-performs when trained on majority group populations leading to the creation of disparate groups in diagnosis, risk prediction, and treatment recommendations. The model should have a better usage of data in diverse forms where entire classes are randomly generated synthetically, thus experiencing more balanced distribution that shows better generalizability and fairness in use of the model among demographic groups. This approach therefore strengthens the ethical soundness of machine learning models in healthcare, making the models more inclusive, less performing disparities and catering more to all patient groups including historically marginalized or underrepresented individuals in the medical data repositories.

4. Applications of Synthetic Data in Healthcare

The machine-learning feature of synthetic creation of data, especially using the GAN-based models, is becoming a paramount necessity in present-day healthcare. The many uses include clinical research, machine learning development, privacy-preserving collaborations and rare disease modeling advancements. This part discusses these applications through specific examples as well as the use of recent research.

4.1. Clinical Research and Epidemiology

In epidemiology and clinical research, it is important to gain access to big, diversified, and representative data. Nonetheless, the privacy laws and separate systems of data usually make it difficult for researchers to use such data. One tool has been found in synthetic data to break this barrier. Researchers in the BMC Public Health study were able to create synthetic population to simulate the effect of a public health intervention without use of the original patient data. This was beneficial in the simulation of diseases transmission, evaluation of immunization measures, and the evaluation of healthcare policies in a much controlled but realistic setting. The original populations were represented statistically in the synthetic datasets, which did not give away the anonymity of the patients. Such a process not only speeds up the process of epidemiological studies but also allows results in it to be of a general nature and can be extended to other demographics and healthcare environments.

4.2. Machine Learning Model Training and Validation

In healthcare, development of reliable and accurate machine learning models is always impacted by the barrier of data scarcity especially those relating to rare diseases or minority subpopulations. Synthetic data adds tremendously to training of the models, and gives more, diversified, and balanced samples. Synthetic data with augmented real-world datasets allows avoiding overfitting, which is generally brought about by small, imbalanced datasets by training the models. As proved in several studies, models, trained on a combination of real-world and synthetic data, have a better quality of generalization and become more robust in the research of external datasets. As an example, synthetic CT and MRI images have been synthesized in radiology to augment real data and enhance the performance of the model in activities like identification of the tumor. In an analogous situation, when synthetic records are used in the EHR-based approaches, the demography is balanced, and the predictive models should work efficiently in multiple patient populations.

4.3. Privacy-Preserving Data Sharing

The use of synthetic data can be one of the most significant in the transfer of privacy data between institutions. Strict policies like GDPR and HIPAA demand additional comprehensive legal contracts and privacy protection of shared data on healthcare, which makes it difficult to share across organizational boundaries. This problem is solved with synthetic data which generates data exhibiting similar statistical profiles to those of real data but also lacking any identifiable patient data. The JAMIA Open article finds several successful applications in which a synthetic subset of EHR data was used to allow cross-institutional research and model development without breaching privacy. This has especially been useful in joint efforts among academic centers and industrial partners where synthetic data may be used to test algorithms, benchmark, and validate them without exposing sensitive patient data to leakage.

4.4. Rare Disease Research

The nature of rare diseases has only limited datasets because they are rare, which in turn is a major problem in the application of machine learning. The synthetic data will be an opportunity because it will allow creating detailed profiles of realistic patients who simulate the features of patients with rare cases. The increase brought about by this augmentation will allow the creation of diagnostics tools that will be more accurate in the diagnosis of unusual diseases even with small amounts of real patient data. Moreover, the synthetic datasets can enable researchers to test different conditions of treatment leading to better comprehension of the disease course and how it can be addressed. A potential to generate evenly matched datasets on these rare conditions will go a long way in expediting research, which would otherwise be obstructed by insufficient information.

Table 2 Applications of Synthetic Data Across Healthcare Domains

Healthcare Domain	Application	Benefits	Example Use Case
Clinical Research	Population simulation for public health studies	Supports large-scale modeling without privacy concerns	Synthetic populations for COVID-19 spread modeling
Epidemiology	Disease intervention simulations	Enables policy testing without real patient data	Vaccination strategy assessment
ML Model Development	Training augmentation and validation	Prevents overfitting; improves model generalizability	Tumor detection in MRI using synthetic images
Privacy-Preserving Sharing	Cross-institutional data sharing	Facilitates collaborations without data breach risks	Synthetic EHR used between hospitals and AI developers
Rare Disease Research	Synthetic rare patient data generation	Enables research on rare diseases; balances underrepresented data	Rare genetic disorder diagnosis model development

5. Evaluation of Synthetic Data Quality

The concern of synthetic data protecting their utility and privacy is a primary issue in healthcare applications. The assessment scheme will be composed of several parts that will measure the utility, the privacy, and intrinsic trade-off between the two.

5.1. Utility Metrics

The first indicator of utility is the extent to which the synthetic data will maintain statistical characteristics of the original dataset and will be applicable in model development and research. Other steps conducted to check statistical similarities such as distribution histogram, covariance matrix and marginal distributions are usually performed to ensure that the synthetic data remain similar when compared with the real data. Principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) visualizations show further confirmation that synthetic records can be clustered in a similar way as real ones.

Machine learning performance testing forms another crucial aspect of utility. Synthetic trained models are tested using their predictive-accuracy, precision, recall and F1-score on real dataset. In case the performance of models does not differ greatly, the utility of the synthetic data is deemed to be high. Specific measures are used too; in the case of imaging, structural similarity indices (SSIM) are measured to compare real and synthetic images, to ensure their quality.

5.2. Privacy Metrics

Privacy metrics compare how well the underlying persons have been anonymized by synthetic data that was used in training. A typical assessment is the membership inference attack in which an attacker tries to determine whether a given subject belonged to the initial training cohort. To be successful, a privacy-related synthetic data must have low susceptibility to these manipulations. Attribute disclosure risks are another issue, where the malicious user makes attempts to infer about some sensitive details about particular people basing on synthetic data. To address such risks, the techniques of differential privacy are used more frequently during GAN training. Differential privacy expressly guarantees a mathematical quality on the impact of modifying or adding an individual to the outcomes of a query to examine the indifference of the impact of the absence or presence of a specific person on the data given back.

5.3. Fidelity vs. Privacy Trade-offs

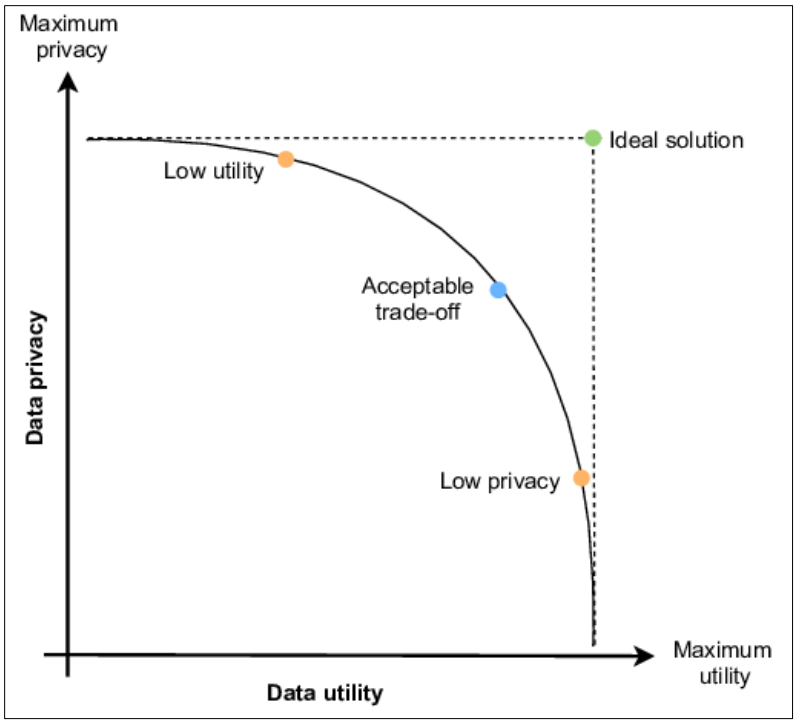

Figure 3 A trade-off curve illustrating how increasing privacy protections reduces data utility and vice versa

One of the main issues that remain pertinent in generating synthetic data is the need to accommodate the tension between fidelity and privacy. Increased fidelity would guarantee that the synthetics exhibit a strong resemblance to the real-life data, which increases the performance of the models and the research results. But the closer it is, the more the chances of unintentionally disclosing confidential information about the patient. In contrast, at the extreme end where privacy-preserving mechanisms are very strong, e.g. a form of differential privacy, the utility of the data can be compromised by the introduction of noise, or loss of a feature granularity. This exchange can be seen clearly in the ACM computing survey article, which shows that privacy-advancing models compromise data richness in the name of confidentiality. The best position on this trade-off curve is determined by the use case, that is, it is ---for high-stakes clinical decision or general research, or exploratory data analysis.

6. Challenges and Limitations

While synthetic data offers substantial benefits, several challenges and limitations must be addressed for it to reach its full potential in healthcare.

6.1. Technical Challenges

Generating synthetic data from GANs poses serious technical challenges. Mode collapse is one of the most widespread issues which cause a generator to not produce enough variations in the output and create monotone and less valuable synthetic records. Also, GANs can have a challenging time on high dimensional data that is tabular in nature as is the case with the EHR data or any other mixed data, i.e. numerical, categorical and temporal data. To address such complexities sufficient advanced architecture designs are required e.g. conditional GANs or table-specific GAN variants, which however are still computationally expensive and technically demanding. In addition to that, there lack uniform evaluation guidelines that would help to confirm that a synthetic dataset is adequate and heterogeneous enough.

6.2. Ethical and Legal Considerations

There is an ethical and legal concern about the use of synthetic data. Even though theoretically, synthetic data removes possible identifiable data of a patient, questions are raised about any possible privacy risk and the possible misuse of synthetic data sets. To give an example, the wrong conclusions might be drawn in case of using synthetic data in the decision-making that has a lot at stake without correct validation. The legal issues additionally drive a wedge, with regulations like GDPR and HIPAA yet to make direct efforts on the control of synthetic information. As the current literature of the European Journal of Radiology Health draws our attention, there is not a universal stance on whether synthetic data produced with the help of real patient data should be left to the same laws as the ordinary health information about a patient. In addition, the question of the informed consent is rather questionable; patients might be unaware that their data are involved in creating synthetic pools, which might be further distributed or sold to a wide audience.

6.3. Validation and Trust

A general principle that still needs to be overcome is trust in synthetic data as being used in clinical practice. Synthetic data is generally received with a form of suspicion among clinicians and regulators especially where validation processes lack rigorousness' and transparency. Contrary to traditional data where provenance is evident, synthetic data should be recorded thoroughly how the data were created including details of the GAN architecture, the characteristics of the training data, and privacy mechanisms. The adoption of synthetic data in the clinical environment is therefore set to be restricted in absence of normalized metrics and validation instruments.

7. Future Directions

As synthetic data technology continues to evolve, several promising directions are poised to shape its future in healthcare.

7.1. Federated Learning and Synthetic Data

Federated learning is the inability to synthesize the model of several institutions but train it in a distributed form without exchanging any raw data. The possible future improvement of privacy is the incorporation of federated learning in combination with synthetic data generation. Local GANs can be trained to make synthetic datasets using their own data and those synthetic datasets are pooled together to collectively analyze it. Such method keeps the privacy intact, and at the same time, it prevents the problems connected with data heterogeneity and institutional biases.

7.2. Explainability in Synthetic Data Generation

One of the major drawbacks of the modern GAN models is that they are not transparent and explainable, creating a sizeable obstacle to using them in medical practice in general. GANs are black box models that are complex since, internally, they are deep neural networks whose inner dynamics are difficult to understand by human users. Although these examples of models are similarly fantastic in screening complicated data distributions and creating convincing artificial information, the murkiness of their decision-making procedures activates vital issues of reliability, accountability, and honesty, especially in delicate areas like healthcare.

Explainability serves as a key element that must be incorporated so that stakeholders (such as clinicians, data scientists, regulatory agencies, and patients) feel confident that the statistics of generated synthetic data possibly represent that of real-world populations of patients. Until the inner mechanisms of the model and the process by which it masters the given input data and produces the synthetic counterparts could be observed with better visibility, there is high probability of such biases captive in the training data would be replicated and even increased in the synthetic ones. As an example, when the original data lack the minority demographic groups, the GAN can unintentionally produce insufficient instances of these groups, which will result in biased downstream machine learning algorithms. In addition, there is a possibility of openly discriminating against certain characteristics (e.g., race, gender or socio-economic status) being suppressed, or overrepresented, which gives rise to fairness concerns.

The issue is also complicated by the fact that GANs lack transparent tools to trace the way specific features affect the outcome. The process of generation of synthetic data cannot be easily satisfied in relation to the fairly representation of the diverse patient subgroups or excessive effect by some variables on the generated samples. This ambiguity detracts the possibility to identify and compensate poor performance of algorithm biases or data leakage, or presentational harms and their aspects, which are vital in healthcare applications, where a decision might result in changes of life.

This drawback has turned out to be the center of research attention of explainable GAN frameworks in future research. The new practices seek to come up with techniques that shed some light on how GANs learn data distributions, how they strike the balance between the sensitivity on certain aspects and, how we can institute fairness constraints on training. Such methods are currently being studied, such as disentangled representations, whereby the GAN learns to infector the various components generating the data, enabling the researcher to manipulate or examine how such specific factors, e.g. of age, gender or disease type, shape the generated outputs. Further, attention mechanisms are under investigation to provide visibility as to what features the GAN concentrates on when being trained and generating outputs, which provides a degree of interpretability comparable to the already demonstrated successful application to natural language processing and computer vision applications.

The next possible direction is including post-hoc explanation tools like SHAP (SHapley Additive explanations) or LIME (Local Interpretable Model-Agnostic Explanations) to evaluate input feature to output relationships rather post-factum. Nevertheless, the approaches have limitations per se as they are not intrinsic to the GAN system. Thus, the topic of injecting explainability directly into the model training process is also under research. As an example, FairGAN and XAI-GAN architecture add fairness constraints and transparency layers which enables a real-time monitoring of how different groups are reflected in the generated synthetic data.

Finally, explainable GANs will play critical roles in ensuring that generating synthetic data is not only technically sound but also ethical due to the idea of fairness, accountability, and transparency applied to healthcare. With these developments, more trust between clinicians, researchers, and regulators will be achieved, thus facilitating safe utilization of synthetic data in clinical research, development of AI models and policymaking.

7.3. Advances in Differential Privacy with GANs

The integration of stronger differential privacy guarantees within GAN models is an active area of research. Emerging models like PATE-GAN and DP-WGAN embed privacy constraints directly into the training process, providing formal mathematical assurances that the synthetic data do not expose any individual's information. As regulatory frameworks evolve, these advancements will become essential for ensuring that synthetic data complies with stringent privacy laws while retaining high utility.

8. Conclusion

Medical data generation based on GANs is one of the paradigms changing breakthroughs in the domain of healthcare data science. The technology provides a disruptive approach to overcome some of the worst problems in machine learning in healthcare, such as data scarcity, privacy, and bias. When it comes to the traditional healthcare environment,

privacy laws like HIPAA, GDPR, and other data governance policies of the institution are dramatically limiting the availability of big, heterogeneous, and high-quality data. What is more, the inequitable distribution of the data, especially concerning the lower representation of some groups of population and uncommon diseases, poses substantial hindrances to the creation of fair and generalized forms of AI. Synthetic data generation provides a direct solution to such shortcomings because GANs allow generating extremely realistic robots without privacy that mimics the statistical properties of real data without any personally identifiable information.

The capacity of GANs to create synthetic data makes an efficient data sharing across institutions and geographic through shared data possible in a safe manner. Such ability enables researchers, clinicians, and the developers of technology, to work on machine learning projects without jeopardizing patient privacy. Moreover, the synthetic data is an effective method of training, testing, and validation of AI models, particularly where real-world data does not exist (or is unbalanced or has tight access restrictions). Synthetic data generation is especially helpful in such domains as rare disease research that is crippled by limited patient data. It can be used to build strong datasets, which can reflect on the clinical character of patients with rare diseases and faster development of diagnostic methods, prognosis models, and therapeutic procedures. In addition, the ability of supporting cross-institutional collaborations, which have so far been slowed down by legal and technical data-sharing restrictions, becomes a more viable alternative when pushed forward by synthetic data pipelines.

Regardless of its potential revolutionization of healthcare, the implementation of synthetic data in the healthcare sector does not lack considerable technical, ethical, and regulatory challenges. The instability of GAN training is one of the important technical limitations that have accompanied GANs with challenges of mode collapses, non-convergence, and the hyperparameter sensitivity problems. Such issues may result in synthetic data that is either lacking in diversity or otherwise does not reflect complex distributions common to healthcare data. Also, it is critical to solve the “case of data fidelity” according to which the synthetic data should not lose its integrity on the intricate relationships and time dependence in the original patient data. Lack of preservation of these relationships would undermine reliability of the AI models using synthetic datasets.

Regarding the privacy issues, synthetic data is constructed to remove direct identifiers but not genuinely immune to privacy threats like membership inference attack or attribute inference attack especially when models are trained on small scale or vastly sensitive datasets. Moreover, the lack of universal and standardized validation procedures and widely agreed assessment standards in terms of evaluating quality, utility and privacy assurance of synthetic data makes its implementation even more difficult. Not only does this create problems with technical credibility, but it will also lead regulators, clinicians, and patients to question the validity and ethical-integrity of synthetic data-driven research.

However, as more privacy-preserving methods, including differential privacy-enhanced GANs, and explainable AI- (XAI-) based synthetic data generation frameworks are developed, these are gradually being overcome. Synthetic data generation can also be complemented with the addition of federated learning paradigm which enables decentralized model training, with no exchange of raw data, supporting data utility across multiple institutions, and protection of their privacy. Such synergy opens the door to future healthcare data ecosystem, which is much more collaborative and privacy-respecting.

With the development of this field, interdisciplinary cooperation will become the key to success. Data scientists are mandated to join forces with clinicians who have the required field knowledge to sign off on the suitability of synthetic data or lack thereof to reflect the true clinical scenarios. At the same time, the scientific community, legal experts, and ethicists are responsible to make sure that the process of synthetic data generation follows the developing regulatory environment and the ethical norms. Policy makers on their part are obliged to come up with clear guideline that balances between ethical use of synthetic data and protection of patient rights.

Conclusively, synthetic data is leading the new wave of innovation in healthcare. It provides scalable, privacy-preserving, and ethically sustainable solution to utilizing the complete possibility of machine learning in medicine. Nevertheless, to achieve this promise, there needs to be a shared commitment to limiting the technical shortcomings, instilling approaches of fairness and openness when generating data and coming up with effective regulatory laws. When such efforts are taken seriously, synthetic data will no doubt become the pillar of the future healthcare data economy, boosting innovation in clinical research, personalized medicine, population screening, and AI-based healthcare solutions across the globe.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Pianykh, O. S., Guitron, S., Parke, D., Zhang, C., Pandharipande, P., Brink, J., and Rosenthal, D. (2018). Big data and machine learning in radiology: Opportunities and challenges. European Journal of Radiology Health, 102, 152–158. https://doi.org/10.1016/j.ejrh.2017.04.004

[2] Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., Ghassemi, M., ... and Rajkomar, A. (2024). Ethical machine learning in health care. Nature Reviews Methods Primers, 4(1), 1–24. https://doi.org/10.1038/s43017-024-00516-2

[3] Feng, S., and Hu, Y. (2022). A survey of synthetic data generation methods for privacy-preserving data publishing. Sustainability, 15(1), 70. https://doi.org/10.3390/su15010070

[4] Mathews, A., Foo, M., and Lim, E. (2023). A survey of synthetic data generation for tabular data in machine learning. Journal of Big Data, 10(1), 1–35. https://doi.org/10.1186/s40537-023-00727-2

[5] El Emam, K., Mosquera, L., Bass, J., Buckeridge, D. L., and Snider, J. (2023). Utility metrics for evaluating synthetic health data generation methods: A survey. ACM Computing Surveys, 56(4), 1–36. https://doi.org/10.1145/3639063

[6] Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2021). Modeling tabular data using conditional GAN. IEEE Transactions on Knowledge and Data Engineering, 35(9), 8493–8504. https://doi.org/10.1109/TKDE.2021.3079836

[7] Kaur, M., and Kumari, S. (2020). Applications of GANs in cybersecurity: Challenges and solutions. In M. I. Khalil (Ed.), Cybersecurity: Issues, Challenges, and Solutions (pp. 227–248). CRC Press. https://doi.org/10.1201/9780367816377-16

[8] Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and evaluation of synthetic patient data. NPJ Digital Medicine, 3(1), 1–13. https://doi.org/10.1145/3485128

[9] van der Valk, H. W., van der Aalst, W. M., and Kumar, A. (2021). Privacy-preserving process mining: A systematic review. Electronic Markets, 31(2), 295–320. https://doi.org/10.1007/s12525-021-00475-2

[10] Kumar, A., and van der Aalst, W. (2023). Foundations of process mining: Accelerating performance with artificial intelligence. In Process Mining Handbook (pp. 31–51). CRC Press. https://doi.org/10.1201/9781003347484-2

[11] Raghupathi, W., and Raghupathi, V. (2022). Leveraging data mining for healthcare analytics: Techniques and applications. BMC Public Health, 22, 1–10. https://doi.org/10.1186/s12889-022-13122-y

[12] Goncalves, A., and Ray, P. (2020). Privacy-preserving synthetic data: A review of approaches, methods, and applications. Health and Technology, 10(6), 1595–1607. https://doi.org/10.1007/s41666-020-00082-4

[13] Ahmed, F., and Litchfield, I. (2022). Ethical and privacy concerns in the use of synthetic healthcare data: A systematic review. BMC Health Services Research, 22(1), 1–16. https://doi.org/10.1186/s12913-022-08215-8

[14] Alharbi, A., and Qureshi, K. N. (2022). Privacy-preserving machine learning using synthetic data: A survey. IEEE Access, 10, 117755–117775. https://doi.org/10.1109/ACCESS.2022.3219845

[15] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2023). Generating multi-label discrete patient data using generative adversarial networks. Neurocomputing, 523, 127017. https://doi.org/10.1016/j.neucom.2023.127017

[16] Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A. (2020). General and specific utility measures for synthetic data. ACM Computing Surveys, 53(1), 1–30. https://doi.org/10.1145/3422622

[17] Baowaly, M. K., Lin, C. C., Liu, C. L., and Chen, K. T. (2023). Synthesizing tabular data using generative adversarial networks: A comprehensive review. Machine Learning, 112(3), 1419–1452. https://doi.org/10.1007/s10994-023-06367-0

[18] Islam, M. R., Islam, R., Noor, N., and Baek, S. Y. (2019). A comprehensive survey on machine learning for cybersecurity: Advances, challenges, and future directions. Archives of Computational Methods in Engineering, 26(5), 1359–1381. https://doi.org/10.1007/s11831-019-09388-y

[19] Costa, P., Ferranti, L., Gazzola, M., and Masera, G. (2019). Privacy-preserving machine learning through federated learning and differential privacy: A survey. IEEE Access, 7, 155550–155574. https://doi.org/10.1109/ACCESS.2019.2905015

[20] Bittorf, V., Hirschfeld, D., Kennedy, O., Kraska, T., and Madden, S. (2021). Qbeast: Towards scalable and efficient data generation. Proceedings of the 2021 International Conference on Management of Data (SIGMOD), 2661–2664. https://doi.org/10.1145/3446374

[21] Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., and Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. JAMIA Open, 4(3), ooab012. https://doi.org/10.1093/jamiaopen/ooab012

[22] Mathews, M., and Jain, R. (2020). Blockchain in healthcare: Applications, challenges, and future perspectives. Computational Intelligence, 36(1), 1–20. https://doi.org/10.1111/coin.12427

[23] Kaiser, J., and Hüsig, S. (2020). The influence of deep learning on sustainable business models of artificial intelligence start-ups. Entropy, 23(9), 1165. https://doi.org/10.3390/e23091165

[24] Umer, S. F., and Zhai, Q. (2021). Applications of artificial intelligence in manufacturing: A review. Applied Sciences, 11(5), 2158. https://doi.org/10.3390/app11052158