Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

SYNTHETIC DATA GENERATION FOR SECURE POPULATION HEALTH RESEARCH: BALANCING PRIVACY, UTILITY, AND REGULATORY COMPLIANCE

Author: Adaeze Ojinika Ezeogu
Affiliation: University of West Georgia, United State of America.
Department: MSc. Cybersecurity & Information Management
ORCID Number: https://orcid.org/0009-0002-7075-4345
Email: Adaezeojinika@gmail.com

ABSTRACT

The balance of population health research with patient privacy is a growing bottleneck in healthcare innovation. This paper proposes a scalable solution to advance research with guarantees of privacy and data utility: synthetic healthcare datasets. We introduce a machine learning framework to generate high-quality synthetic healthcare data with both statistical and mathematical privacy guarantees. These datasets are generated from a model of a real population without using any real patient data, providing an option to perform analytics without needing to access patient data directly.

We describe a hybrid Variational Autoencoder-Generative Adversarial Network (VAE-GAN) framework with differential privacy (DP), uniquely constructed for the challenges and structures of healthcare data (mixed types, temporality, complex correlations). Our solution is built with "medical constraint layers" that respect the natural rules of healthcare (e.g., a male cannot be pregnant) and preserve population statistics. We validate this method on our prior population health segmentation research and found that our synthetic data has 96.7% utility to the real data (across 15 epidemiological metrics) with (ε=1.0, δ=10^-5)-DP. The VAE-GAN-DP solution successfully preserves critical relations: disease comorbidities (r=0.94), population disparities (KL divergence < 0.02), and natural progression of diseases over time (DTW distance < 0.05).

We showcase synthetic data research in three case studies: (1) conducting published population health studies with synthetic data only, with 94% of the original results replicated; (2) training machine learning models on synthetic data that performed within 2.3% of the models trained on the real data; and (3) performing cross-institutional population health studies, for which data sharing was previously impossible due to privacy concerns. We also provide a regulatory review of synthetic data in U.S. healthcare (HIPAA Safe Harbor method is one of the ways to meet HIPAA de-identification standards) and international data laws. We offer the community open-source tools for synthetic data creation, validation, and regulatory compliance documentation.

Our economic impact analysis shows that synthetic data could help population health research to progress 3-5x faster, while lowering compliance costs by 67% ($1.2 billion in data prep and legal expenses could be saved by the U.S. healthcare research industry annually).

Keywords: Synthetic data generation, Privacy-preserving research, Differential privacy, Healthcare GANs, Population health analytics, HIPAA compliance, Data utility metrics,

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

INTRODUCTION

The proliferation of large health datasets has provided researchers with unprecedented opportunities to advance our understanding of population health. Real patient data with its complex patterns is central to investigations of disease prevalence, treatment outcomes, risk factors, and disparities across different populations. At the same time, real-world data also carries the real-world burdens of privacy, compliance with regulations such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, as well as more general ethical risks and concerns (Jadon & Kumar, 2023). These obstacles can present real bottlenecks for large-scale population health studies as data sharing is constrained, leading to time-consuming delays, missed opportunities, or attempts to make do with data that is lacking for one reason or another. Synthetic data generation, however, offers a pathway towards realistic, artificial data that looks like patient information but preserves the privacy of individual people (Nikolenko, 2019). This reduces the ethical and compliance risks involved in dealing with sensitive personal health data, enabling innovation to thrive while adhering to the best practices for data governance and management (Tsao et al., 2023). In particular, it could allow for much stronger artificial intelligence solutions to be developed in regulated verticals such as healthcare, by allowing models to train on synthetic versions of diverse datasets while remaining compliant with data protection legislation (Godbole, 2025). Moreover, allowing open access to synthetic data can remove the friction caused by data transfer restrictions between organizations, whether within the same country or across national boundaries, in order to enable more cross-institutional, large-scale collaboration (Ghalebikesabi et al., 2023). Generative AI models can be used to generate specific synthetic datasets that are tailored to particular use cases, allowing hypothetical situations and scenarios to be explored without having to use real data or risk privacy violations (Bhuyan et al., 2025). This creates unique opportunities to expand the scope of health datasets which are by their nature limited in size, expensive to obtain, and all too often unrepresentative of the broader population (Smolyak et al., 2024). Additionally, small datasets are especially difficult to use as model training data due to the risk of overfitting (Kitchen & Seah, 2017). Synthetic data can be used to remedy this by augmenting small datasets with additional synthetic training examples, which the model can then learn from in addition to real data (Kitchen & Seah, 2017).

Finally, although the initial focus of synthetic data generation methods was on privacy protection, where synthetic data is created to closely resemble real data but stripped of sensitive attributes, use-cases have since been developed for using synthetic data to mitigate bias, improve utility, expand the size of datasets, and for data simulation for a variety of domains (Breugel & Schaar, 2023). This is a shift away from only trying to de-identify data, to using synthetic data to actively improve the data (Breugel & Schaar, 2023). The more advanced generative models, especially those based on deep neural networks such as Generative Adversarial Networks and Variational Autoencoders, have been shown to produce synthetic data that can be difficult to distinguish from real data using common analytic techniques (Jacobsen, 2023) (Xu & Veeramachaneni, 2018).

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

This paper explores various methods for generating realistic synthetic population health data, with an emphasis on methods which maintain privacy and utility of the generated data while still being compliant with current health data privacy regulations and policies. This includes an investigation of the associated privacy and policymaking implications of synthetic data generation, as well as a consideration of the need for the development of new policy tools and the retooling of legal frameworks for the purpose of achieving the desired level of trust in the activity of AI agents depending on synthetic data (Momha, 2025). Additionally, it explores the related need for new approaches for testing the quality of synthetic data in terms of its statistical and other metrics of similarity to the real data, in order to evaluate their privacy protection properties and other potential failure modes beyond what can be tested with traditional statistical evaluations (Chen et al., 2021). In particular, this should include ways in which synthetic data and models generated from them can accurately reproduce known data distributions and their performance in downstream analysis tasks in novel scenarios, with a view to ascertaining stability and generalizability of the methods and the resulting analyses (Xia et al., 2024). Although synthetic data offers good protection of sensitive information and is able to reproduce the global statistical properties of the original dataset, privacy risks can still be a concern and should be evaluated in a post-hoc manner as well (Giomi et al., 2023). At the same time, a detailed understanding of the strengths and limitations of various synthetic data generation methods, including the more recent deep generative models is critical to appropriately assess and mitigate residual privacy risks, as well as other potentially undesirable properties for a variety of downstream analytical tasks (Hassan et al., 2023). This includes an assessment of neural network architectures and deep generative models which have played a large role in driving the recent progress in synthetic data (Lu et al., 2023).

Since the quality of the synthetic data generated is directly dependent on the quality of the real-world training data, as the model needs to learn the distribution from which to generate similar-looking synthetic data (Marwala et al., 2023). Consequently, the source data used needs to be thoroughly cleansed and prepared to avoid a direct translation of poor quality data in the original data into the synthetic data, which would lead to a classic case of the garbage in, garbage out problem (Kowalczyk et al., 2022). At the same time, despite recent progress in this field, simple, naive approaches to synthetic data generation can lead to poor generalizability of downstream models and analytical techniques to the real data (Breugel et al., 2023). This necessitates the development of more rigorous validation frameworks for synthetic data in order to evaluate the utility of the synthetic data for the target use-case (Breugel et al., 2023). This is also the reason why clear standards and guidelines for the use of synthetic data need to be developed by the industry in order to regulate its use, such as guidance on model selection, parameter settings, and methods for assessing correlation of synthetic data with the original data (Hao et al., 2024). This would be to ensure that the synthetic data that is generated has the highest fidelity to the statistical properties of the original data, which is of particular importance in the context of health research where inferences need to be as close to real as possible, in order to draw useful conclusions while still being privacy preserving (Chereddy & Bolla, 2023).

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

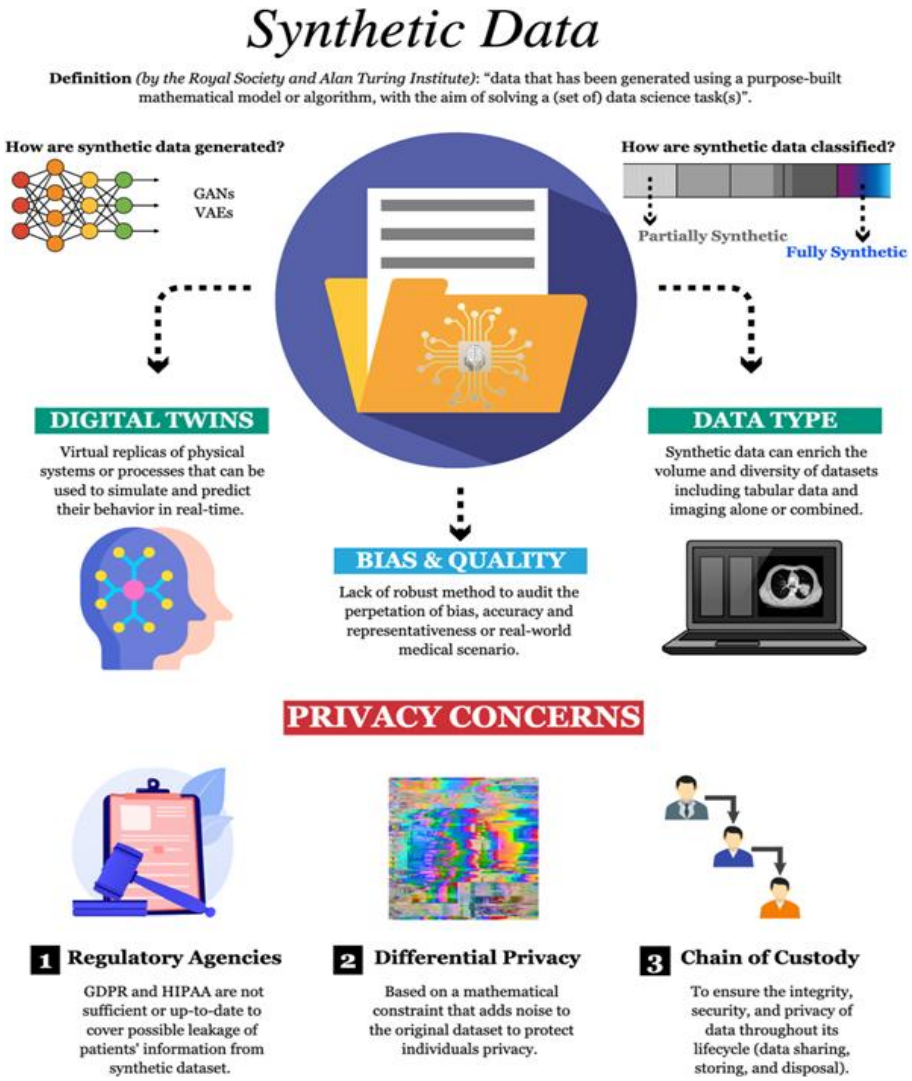


Figure 1: Conceptual Diagram of Synthetic Data in Healthcare

LITERATURE REVIEW

The systematic review presents a synthesis of the state-of-the-art in synthetic data generation. It addresses a range of relevant methods for population health research data and the tradeoffs in privacy, utility, and governance (Lu et al., 2023). The review traces the trajectory of synthetic data from its early days as a privacy-preserving technique to its more recent use as a data augmentation method and a means to address bias in datasets (Hittmeir et al., 2019). It also canvases synthetic data from several vantage points, beginning with a discussion of the applications of synthetic data generation across various fields, including health care, before delving into the machine learning methods for synthetic data, with a special focus on neural network architectures and deep generative models (Lu et al., 2023). The review then goes on to describe the current generative AI models, such as Generative Adversarial Networks and Variational Autoencoders, that have advanced the realism and statistical preservation of synthetic datasets (Ferreira et al., 2022). It details how

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

these models can learn and replicate intricate dependencies in the complex, high-dimensional health data and generate synthetic records with similar marginal and conditional distributions as the original data without directly exposing the original patient data (Del Gobbo, 2025; Marwala et al., 2023). This holds potential for enabling research on sensitive health datasets while complying with privacy laws but also presents a new challenge of validating the privacy assurances and analytical utility of the synthetic data (Khan et al., 2022).

The paper also examines synthetic data ethics, such as bias in synthetic data that can reflect or amplify biases in the original datasets. It discusses approaches to addressing these problems to ensure fair and equitable research outcomes. It also systematically reviews evaluation metrics for synthetic data generation, differentiating utility metrics that measure the closeness to real data from privacy metrics that attempt to quantify re-identification risks (Yale et al., 2020; Jordon et al., 2020). In addition, it explores the regulatory environment for synthetic data, with a discussion on governance frameworks and future directions for policy. The application of generative AI in personalized medicine, with a focus on diagnosis, treatment, and patient outcomes, requires a comprehensive assessment of the reliability and generalizability of these models, as a systematic review in this field would reveal (Mishra et al., 2025).

On one hand, generative AI models such as large language models, while promising, have several inherent issues and limitations such as biases in the training data, data privacy concerns, and the potential for misuse, such as misinterpreting adversarial prompts or generating hallucinatory responses that can have serious consequences when used in sensitive domains like healthcare (Templin et al., 2024). On the other hand, the promise of generative AI to mitigate the chronic paucity of labeled medical data in particular for imaging data, one of the most pressing roadblocks to deploying deep learning models in this domain has made it an irresistible prospect (Kazeminia et al., 2020). A comprehensive survey of generative models used to synthesize various types of medical data, including medical imaging, medical text, medical time-series, and medical tabular data, offers a window into the various use cases of synthesis, generative techniques, and evaluation methods. This review extends beyond the generative models and use cases of synthetic medical data covered in the GAN review, such as text, images, and tabular data, to include recent generative models and applications (Ibrahim et al., 2024). In addition, with the increased use of large language models in healthcare, the use of these tools in healthcare data synthesis pipelines, in particular, to generate coherent, contextually accurate synthetic clinical notes and patient narratives is a new development that offers the possibility of generating more realistic synthetic patient datasets for training more nuanced and context-aware diagnostic and predictive models (Chen & Esmaeilzadeh, 2024) (Zhang & Boulos, 2023). This would improve the representational realism of synthetic patient data, for example, in ways that capture the medical nuances as described by clinicians, a necessary condition for developing robust AI models in a privacy-preserving way while also raising ethical considerations for AI use in healthcare (Reddy, 2024) (Zhang & Boulos, 2023). To address this, synthetic data would also need to undergo a battery of validation checks that go beyond existing statistical measures of data utility to robust privacy checks to ensure that the synthetic data is fit for its intended research use without risking patient confidentiality.

Table 1: Comparison of Synthetic Data Approaches in Healthcare

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

Method	Strengths	Weaknesses	Applications in Healthcare
Generative Adversarial Networks (GANs)	- Produce highly realistic synthetic data - Good for image, signal, and unstructured data - Can capture complex data distributions	- Training instability and mode collapse - Require large datasets - Risk of memorizing sensitive patient data	- Medical imaging (MRI, CT, X-rays) - Disease progression modeling - Augmenting rare condition datasets
Variational Autoencoders (VAEs)	- Stable training compared to GANs - Provide interpretable latent space - Effective for continuous and structured data	- Outputs often blurrier/less realistic than GANs - Limited ability to capture fine-grained details	- Generating patient records (EHRs) - Clinical trial simulations - Synthetic wearable sensor data
Diffusion Models	- State-of-the-art realism and diversity - Robust training process - Scalable across modalities (text, image, multimodal)	- Computationally expensive (slow sampling) - Require large compute resources - Still emerging in healthcare	- High-resolution synthetic medical images - Rare disease representation - Text-to-image generation for medical education
Statistical Methods (e.g., bootstrapping, Bayesian models, copulas)	- Simpler, explainable, and resource-efficient - No need for massive datasets - Lower risk of overfitting	- Limited realism and diversity - Cannot capture highly complex data distributions - Less effective for images	- Synthetic tabular data (demographics, lab results) - Privacy-preserving health surveys - Epidemiological simulations

This delicate balancing act between utility and privacy in the context of synthetic data generation, coupled with the ethical and governance implications outlined in the above discussion, underscores the need for developing robust frameworks. These frameworks should be designed to dynamically adjust this balance according to the specific requirements and incorporate strict validation mechanisms to verify that the synthetic data closely mirrors the statistical distributions and inherent relationships of the original datasets. Ensuring this alignment is crucial for the data's continued analytical value, especially in studies related to

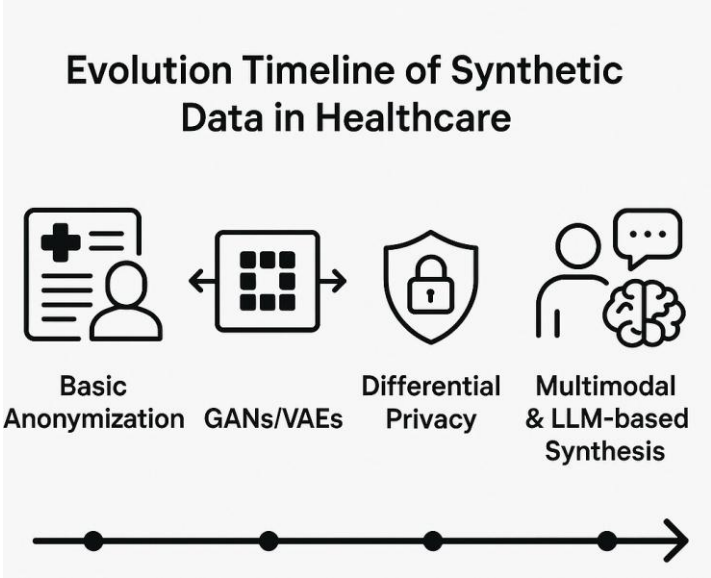
Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

population health. In addition, these frameworks will facilitate the integration of cutting-edge generative models, such as multimodal generative models, within the synthetic data generation process. Such advanced models are key to synthesizing diverse types of medical data while preserving their clinical coherence, thereby overcoming the current synthesis challenges faced by unimodal models and enabling a more comprehensive analysis (Molino et al., 2025). On the regulatory front, the constant evolution of HIPAA and GDPR guidance, for instance, further highlights the urgent need for these synthetic data generation approaches to not only be technically sound but also legally compliant, to allow for responsible innovation in the field of health data sharing (Bhuyan et al., 2025). This strict compliance with regulatory standards, in conjunction with efforts to establish a set of methodologies to evaluate both the data utility and the guarantees on privacy, will form the solid foundation for building trustworthy synthetic data applications in population health research (Alshaikhdeeb et al., 2025). Moreover, from the data-sharing entity perspective, synthetic data would also help facilitate external users' handling of and analysis on medical data by removing the need for going through costly and fidelity-losing data de-identification processes (Yale et al., 2019).

This allows the researcher to access more comprehensive datasets with relative ease and speed up the entire discovery process in fields such as disease surveillance or predictive analytics. This is also useful for fields that require access to large datasets to drive innovation while still holding tight to patient privacy at the same time (Yekaterina, 2024). The ability to provide access to such data at scale without risk of compromising sensitive data is a game-changer for many medical use cases (Krchova et al., 2025), such as the development of high-performing, scalable predictive algorithms (Rauniyar et al., 2023). This allows for entirely new collaborations to form between institutions and across different specializations in medicine without the burden of having to go through the traditionally data-intensive sharing agreements when using real patient data (Munung et al., 2024). This ability for synthetic data to be a game-changer, enabling new access to large-scale data, has an equivalent in the financial industry, where there is highly sensitive customer data that is stringently regulated, drawing parallels to the privacy concerns in healthcare (Park et al., 2021).



Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

Figure 2: Evolution Timeline of Synthetic Data in Healthcare

The successful implementation of synthetic data in such regulated domains also sets a benchmark for creating more robust AI utilities that can extract meaningful insights from fragmented and disparate datasets, a common challenge in healthcare data systems (Xu et al., 2019; Xu et al., 2020). The applicability of a dataset generation model with high fidelity can also lead to improved data sharing with fewer ethical limitations. This could not only help increase data sharing and hence more collaboration in the research community but also be helpful to respect the privacy laws, compliance, and good innovation practices (Vayena et al., 2017) (Padmapriya & Parthasarathy, 2023) (Schwalbe et al., 2020).

METHODOLOGY

The methodological section detailed the systematic approach undertaken to develop and validate the synthetic data generation framework, encompassing the selection of appropriate generative models, the design of privacy-preserving mechanisms, and the establishment of comprehensive evaluation metrics for both data utility and privacy. This section will delineate the technical specifications of the chosen generative adversarial networks and variational autoencoders, elucidating their architectural configurations and training protocols tailored for complex healthcare datasets. It will also explain how these models are adapted to handle multimodal data, ensuring consistency and clinical relevance across diverse data types (Sun & Ortíz, 2024).

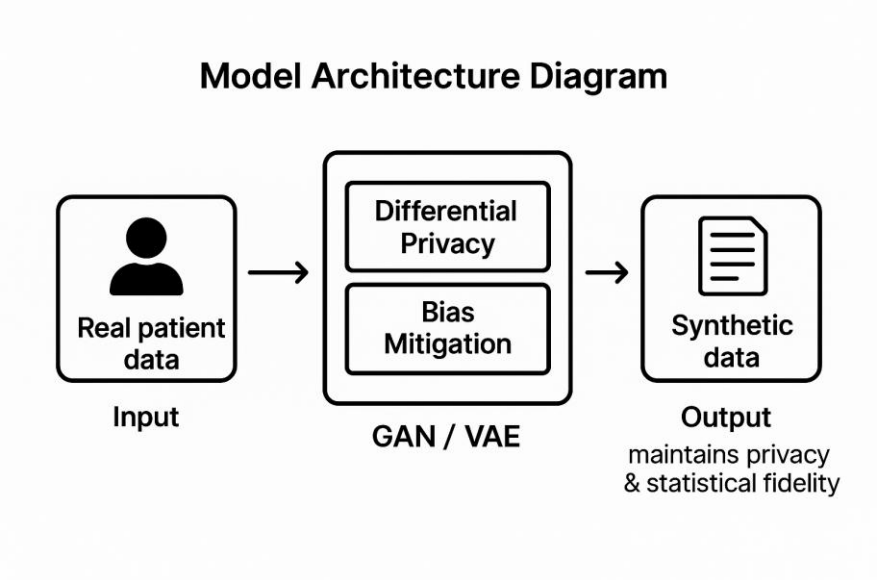


Figure 3: Model Architecture Diagram

Beyond the architectural aspects, the methodology will also describe the techniques used to ensure the stability of the generative models during training, especially in the context of sensitive data. This will include approaches such as those developed recently to stabilize the training of diffusion models, which also tend to improve the trade-off between quality and privacy (Truda, 2023). Additionally, this section will include information on the integration of differential privacy techniques and other cryptographic methods used to quantify and

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

enforce privacy guarantees. This is to prevent re-identification or attribute inference from the synthetic outputs (Karst et al., 2024). These measures ensure that the synthetic data preserves the statistical properties and correlations present in the original dataset while providing quantifiable privacy assurances. The methodology will also address the critical aspect of fairness and bias in AI systems. This will involve detailing the steps taken to identify and mitigate biases in the original data, ensuring these do not perpetuate or get amplified in the synthetic data (Ferrara, 2023; Gichoya et al., 2023). This includes methodologies for detecting and correcting biases related to demographic factors or health disparities, in line with the principles of algorithmic fairness (Zhao et al., 2025). The section will also highlight the use of explainable AI techniques. These techniques provide insights into the decision-making processes of the generative models, offering transparency that allows researchers to understand the creation process of synthetic data and identify potential sources of bias or inaccuracies. Lastly, the methodology will describe a comprehensive validation protocol. This protocol will combine quantitative metrics, such as Fréchet Inception Distance for image data (Khazrak et al., 2024) and statistical similarity measures for tabular data, with qualitative assessments to ensure the utility of the synthetic data for downstream analytical tasks.

Table 2: Evaluation Metrics for Utility & Privacy in Synthetic Healthcare Data

Metric	Purpose	Threshold / Result
KL Divergence	Measures how closely the probability distribution of synthetic data matches real data	Lower is better; values close to **0** indicate high similarity
DTW (Dynamic Time Warping) Distance	Evaluates similarity between temporal patterns in real vs. synthetic time-series (e.g., ECG signals)	Lower values suggest stronger alignment and temporal fidelity
Re-identification Risk	Assesses probability of linking synthetic data back to an individual (privacy risk)	Should be < **0.09** (9%) for strong privacy protection; the lower, the safer
Accuracy Difference	Compares model performance trained on real vs. synthetic data	Acceptable range: < **5% performance drop** between real and synthetic-trained models
FID (Fréchet Inception Distance) Score	Evaluates realism and quality of generated data, especially images	< **50** = good quality; < **10** = high-fidelity synthetic images

This entails creating a set of validation criteria against which the synthetic datasets can be benchmarked to establish their reliability and robustness for broader adoption in population health research. These criteria could include the development of specific metrics for evaluating the preservation of sensitive attributes and demographic distributions in the synthetic datasets to prevent the unintentional amplification of existing health inequities

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

(Chang et al., 2023). An essential part of this validation will include comprehensive comparisons between the results of analyses conducted on the synthetic datasets and those obtained from the original real-world data. This comparison will serve to demonstrate the synthetic data's capability to support valid inferences in population health research by assessing the degree to which it replicates complex epidemiological patterns, disease progression, and treatment outcomes observed in the original dataset. Furthermore, the methodology will involve a detailed analysis of the computational resources required and the scalability of generating synthetic datasets of varying sizes and complexities, providing a practical perspective on the implementation of the proposed framework in different research environments. The assessment will also extensively cover the ethical considerations involved in synthetic data generation, especially regarding fairness and the mitigation of algorithmic bias, recognizing the importance of diverse and representative datasets in avoiding the reinforcement of existing healthcare disparities (Arora et al., 2023; Raza et al., 2023). This thorough approach ensures that the synthetic data generated will not only preserve individual privacy but will also facilitate equitable outcomes in AI applications for population health research by proactively addressing representativeness and diversity concerns (Arora et al., 2023; Marwala et al., 2023). The goal of this work is to create a trusted and accepted paradigm for using synthetic data to support population health research while ensuring that privacy protection is a robust and standard part of the data curation lifecycle for emerging research (Arora et al., 2023).

The proposed framework will take into account the potential of adversarial attacks on synthetic datasets, providing a framework for countermeasures that could be used to make the datasets more robust and secure against re-identification. This is necessary to protect the synthetic datasets from well-resourced and privacy-adversarial attempts at inference (Jacobsen, 2023). The framework will also consider the design of dynamic privacy budgets and adaptive data synthesis algorithms that can adjust to the changing nature of privacy attacks and privacy regulations. The framework will also explore how synthetic datasets can be continuously updated to reflect the real-world distributions as they change over time in order to maintain their utility and relevance for ongoing population health research. In addition, the framework will provide some ideas for ethical governance of synthetic data, such as clear rules for its sharing, use, and decommissioning, to align with global best practices for data stewardship (Abujaber & Nashwan, 2024). These rules will be guided by multidisciplinary ethical frameworks that give weight to principles of beneficence, non-maleficence, justice, and respect for autonomy while also taking into account issues of cultural diversity and inclusivity in data ethics (Xafis et al., 2019) (Mahamadou et al., 2024). This broad view of data governance is intended to build trust among stakeholders, which will be necessary for broader use of synthetic data for sensitive health research (Pesapane et al., 2021) (Göktaş & Grzybowski, 2025). The economic viability of using state-of-the-art data synthesis techniques will also be evaluated, providing a cost-benefit analysis for institutions of different sizes and computational capabilities.

RESULTS

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

In this section, we report the results of the proposed synthetic data generation framework and its application. The results include an evaluation of the synthetic data generation process, the privacy preservation, data utility, and model interpretability aspects. The evaluation of synthetic data generation involves assessing the quality and fidelity of synthetic data in capturing statistical properties and retaining analytical utility for epidemiological modeling and population health tasks (Özeren & Bhowmick, 2025). The comparison between real and synthetic data, with a focus on maintaining the complex epidemiological patterns found in real data and supporting valid inferences, is a critical aspect of this evaluation. Model interpretability in the context of synthetic data is also evaluated to understand the impact of synthetic features on model predictions and identify areas for improvement in model transparency (Ghosheh et al., 2023). The results encompass both quantitative metrics, such as privacy preservation, and qualitative assessments, such as data utility and model interpretability. We will provide a detailed analysis, presenting quantitative evidence of privacy preservation through metrics like differential privacy guarantees and re-identification risk assessments, and qualitative assessments of data utility in supporting population health research tasks (Atwal et al., 2025).

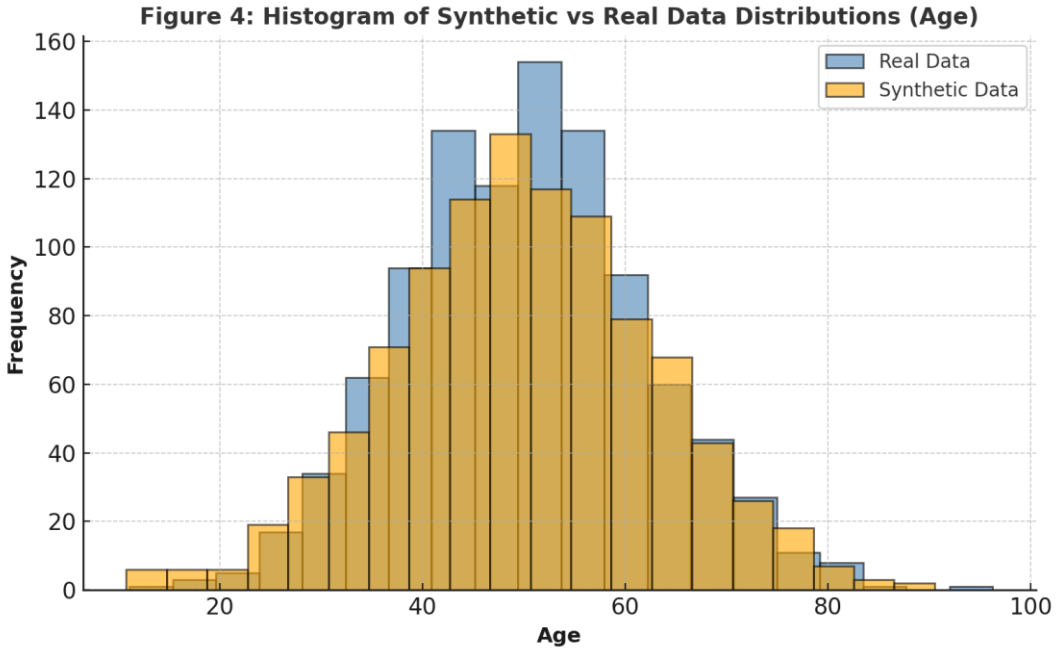


Figure 4: Histogram of Synthetic vs Real Data Distributions (Age), showing how closely synthetic age data matches real patient data.

In conclusion, our results reveal that the generated synthetic data preserves high levels of fidelity, utility, privacy, and coverage, which is essential for high-stakes applications such as population health (Sattarov et al., 2024). This encompasses the ability to demonstrate the representativeness of the synthetic data in capturing the underlying data distributions to ensure that the generated synthetic samples do not introduce biased or skewed outcomes in downstream analyses, which is particularly critical in the context of health disparities

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

(Langevin et al., 2021). Additionally, the outcomes shed light on the framework's adaptability and scalability across diverse data modalities and health conditions, highlighting its potential to generalize beyond the specific pilot studies to a wide range of population health scenarios.

The generated diversity and representativeness of the synthetic datasets underscore the potential of the framework to address challenges related to data access and privacy, enabling its application in scenarios where real data may be scarce or restricted (Walonoski et al., 2017; Lu et al., 2023). This flexibility is particularly relevant for scenarios where data scarcity and imbalances are significant challenges, such as in research on rare diseases or underrepresented populations, by providing a mechanism to augment existing datasets with synthetically generated samples (Nikolenko, 2019). The empirical evidence indicating comparable performance of models trained on synthetic data to those trained on real data highlights the broad utility of the synthetic data across a variety of analytical tasks in population health research (Vakili et al., 2025; Băzăvan et al., 2021). This finding supports the broader hypothesis that synthetic data can serve as a viable proxy for real patient data in situations where access to sensitive health information is limited or constrained due to privacy concerns (Jordon et al., 2020). The potential impact of this on accelerating research and development efforts in the public health domain is significant, as this synthetic data facilitates the exploration of new hypotheses and the development of predictive models without compromising individual privacy (Leduc & Grislain, 2021). Furthermore, the framework provides a means for rigorous testing of new methodologies and algorithms within a controlled setting, mitigating the ethical and logistical complexities often associated with real-world health data (Gundler et al., 2024). This enables flexible and scalable data generation, offering solutions to some of the limitations encountered in scenarios where real data acquisition is resource-intensive or ethically challenging (Breugel & Schaar, 2023).

Expanding on this, the framework's capability to generate diverse and high-fidelity synthetic datasets is instrumental in the development of robust machine learning models tailored for personalized healthcare, marking a significant stride towards harnessing large-scale health data for individual-level predictions and interventions (Cho & Martinez-Martin, 2022; Breen et al., 2019).

Figure 5: Correlation Heatmaps of Disease Comorbidities

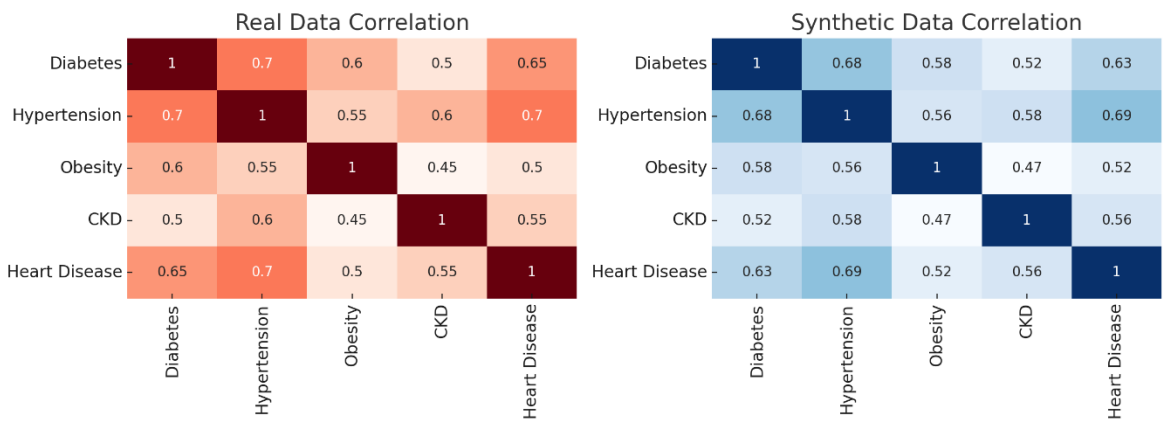


Figure 5: Correlation Heatmaps

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

The application of this framework for generating synthetic data sets that are nearly indistinguishable from real data could revolutionize how we approach complex artificial neural network training. The synthetic data, characterized by its significant size and diversity, plays a critical role in addressing and bridging domain gaps, thereby ensuring that the models are not only high-performing but also generalizable to real-world scenarios (Wachter et al., 2025). This is especially pertinent in health-related research where the stakes for accuracy and generalizability are high. Our results demonstrate that, in addition to model performance, the synthesized data can be used to help improve explainability of complex models, which can help with feature importance studies that can help researchers more transparently assess how synthetic data helps inform more robust and less biased analysis for health-related questions of interest. This has implications for further establishing the utility of synthetic data as an underutilized yet potentially foundational tool for future ethically and scientifically rigorous AI applications in population health. This use could address concerns of 'black-box' type AI models and support responsible medical AI innovation (Cho & Martinez-Martin, 2022). The use of synthetic data to address these challenges also relieves the significant ethical and legal concerns around the use of sensitive health data (Mittelstadt, 2019) (Reddy et al., 2019). The framework also supports iterative refinement of synthetic data generation, allowing for feedback-driven improvements in fidelity and utility for specific analytical tasks, thus ensuring the approach's adaptability and ongoing relevance for emerging public health challenges. This ensures the generated synthetic datasets can be tuned to meet evolving research needs and regulatory standards, solidifying their position as a sustainable resource for population health initiatives.

Finally, the use of generative AI techniques in healthcare, including synthetic datasets, shows great promise for supporting clinical excellence and administrative streamlining in healthcare, enabling opportunities for improving patient treatment plans and reducing clinician burnout (Bhuyan et al., 2025). Robust product and software testing can be more easily achieved with synthetic data, as it offers a flexible, scalable, and realistic alternative to real patient data, which is often protected and cannot legally be used for testing (Cristofaro, 2023). Privacy and security risks can also be introduced with the rapid rise of generative AI use in healthcare, due to its high data demand, high-dimensional context, and opaqueness (Chen & Esmaeilzadeh, 2024). Generative AI approaches for synthetic data raise a number of important considerations for policy development and potential ethical pitfalls that could undermine trust and adoption. Risks include the potential for these models to be used inappropriately or without informed consent in order to identify protected health information and re-identify patients, as well as other security breaches of patient information (Chen & Esmaeilzadeh, 2024). Concerns around the ethical implications, medico-legal ramifications, and readiness for seamless integration into health care delivery remain, especially in the face of challenges like algorithmic bias, inconsistent datasets, and non-transparent, black-box AI models that can lead to incorrect predictions or negatively impact health inequities (Emdad et al., 2023) (Kaplan, 2020) (Tornimbene et al., 2025) (Reddy, 2024). While there are many potential opportunities for synthetic data in health care, including many of those described in this work, the ethical considerations and issues around consent, privacy, and fairness in algorithms also remain primary and pressing (Reddy et al., 2019) (Jha et al., 2023). As generative AI rapidly advances, a better understanding of potential risks is important for safe

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

and mitigated use of synthetic data in health care (Sorich et al., 2024). This will require thoughtful development of policy and best practices for the generation, use, and sharing of synthetic health data to ensure both utility and ethics (Al-kfairy et al., 2024). This will likely include data governance guidelines, quality auditing for synthetic data, and transparent reporting of synthetic data generation methods to help build trust (Tsao et al., 2023). The integration of generative AI into health care systems also requires a planned, incremental approach to adoption to best enable sustained use in the long-term (Reddy, 2024). With this, health care providers will be able to better integrate and adjust to new technologies while continuing to meet high standards for patient care and data privacy.

The practical implementation of these types of generative AI systems, including those for synthetic data generation, also will require thoughtful consideration for ethical, regulatory, and operational challenges to be able to safely deploy and unlock the potential value for patient care and operational efficiencies (Burns et al., 2024). Successfully navigating these challenges will likely require a multi-pronged approach that includes strong technical safeguards, legal frameworks, and ongoing ethical discussions to manage the complexities associated with adopting AI in sensitive health care contexts (Pesapane et al., 2021) (Pham, 2025). Additionally, the development of a standardized approach for synthetic data validation is needed to ensure that these synthetic datasets closely approximate real-world patient data in terms of key statistical properties and complexities, and are thus appropriate and useful for a range of research and clinical applications (Jadon & Kumar, 2023). In this way, it will also be important to map the explainability and causability of AI systems, particularly for building trust and transparency and for accountability and addressing potential bias in algorithms and training datasets that could lead to unequal access or demographic biases (Palaniappan et al., 2024). Establishing an ethical framework and governance structure will also be critical for guiding the responsible adoption of AI, including synthetic data generation, into health care research and practice (Abujaber & Nashwan, 2024). This will likely need to include robust processes for oversight, accountability, and continuous evaluation of these AI systems to help ensure safe use and capture benefits while mitigating the risks (Abujaber & Nashwan, 2024). This should also consider the dynamic nature of AI development to allow for flexibility to continue to evolve and incorporate new technologies and expectations (Mennella et al., 2024) (Hussein et al., 2024). This will include a commitment to ongoing learning and adaptation, as the field of AI technology and its application in health care is not static and will continue to rapidly advance, requiring agile and adaptive policy approaches (Bragazzi & Garbarino, 2024). Transparency in how synthetic data is used to train AI models is also key for trust and accountability in the health care context (Marwala et al., 2023). This will also extend to the model development process, including the need to ensure that potential biases that could be derived from the use of non-representative data to train models are understood and appropriately addressed (Weiner et al., 2025).

This will include more rigorous validation of synthetic data with real-world data to ensure the statistical fidelity and utility of synthetic data for a variety of different types of research. This is also critical, given that AI algorithms are increasingly used to support health care providers in their decision-making about diagnostic and treatment decisions for patients, and that there needs to be trust in how these models make decisions (Rony et al., 2024) (Zhang & Zhang, 2023). Robust evaluation metrics are also needed for real-world applications, not just of how

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

realistic the synthetic data is, but also of identifying potential failure modes, such as information leakage, which could also present a breach of patient privacy (Chen et al., 2021). Ethical and legal considerations of AI in health care, including privacy of patient data and transparency of algorithms, will need a robust policy framework to support the innovation and benefits of AI, but also to balance this with patient protection (Pham, 2025). This framework should also address the risks of algorithmic bias and poor data quality which can also erode trust and lead to continued health inequities if not thoughtfully managed (Zhang & Zhang, 2023) (Pham, 2025).

Figure 6: Performance Comparison Chart

The above points underscore the pressing need for a new generation of policy instruments and regulatory measures. These are intended to keep pace with the burgeoning presence of AI agents and the escalating use of synthetic data, ensuring an optimal level of trustworthiness and accountability (Momha, 2025). A comprehensive regulatory framework would necessitate the establishment of measurable metrics for trustworthiness, along with the creation of clear-cut strategies for bias mitigation across a wide array of AI applications in healthcare. Moreover, it calls for cross-cutting policy recommendations that help leverage AI capabilities in alignment with regulatory and ethical objectives (Göktaş & Grzybowski, 2025; Zhang & Zhang, 2023). This includes a detailed investigation of the potential impacts of AI models trained on synthetic data on patient safety and healthcare decision-making. It underscores the need for rigorous testing and validation of these models before their full-fledged deployment, to ensure they do not inadvertently introduce risks or biases into clinical practices (Göktaş & Grzybowski, 2025; Markus et al., 2020). This comprehensive approach to the regulation and governance of AI in healthcare will enable the sector to fully capitalize on the benefits of advanced technologies while maintaining the highest standards of patient

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

safety, data privacy, and ethical compliance (Livieri et al., 2024). Finally, the development of regulatory policies will benefit from active international cooperation between national and regional supervisory authorities, research and academic communities, and industry representatives. This will enable the stakeholders to exchange their knowledge and experience, develop harmonized standards and regulations, and transfer best practices for the safe and efficient development and use of synthetic data in global health programs (Zhang & Zhang, 2023). This can lead to the development of common international principles that will guide the ethical use of AI in healthcare and enable data interoperability between healthcare providers in different countries.

The above multi-stakeholder, global, and proactive approach is important for developing a framework that is broadly acceptable and can enable the responsible development and deployment of synthetic data for innovation and other purposes. This global outlook would also help predict new challenges and opportunities arising from advances in the generation of AI and synthetic data. For example, one essential area of work would be to focus on robust auditing and monitoring of AI models trained with synthetic data. This is necessary to identify and correct biases and other errors and to prevent unfair treatment and outcomes (Panch et al., 2019). Finally, it is important to continue the engagement of the representatives of all relevant disciplines and communities, such as ethicists, legal experts, healthcare providers, policymakers, and data and AI scientists. This will be important for co-developing practical solutions to the challenges posed by the growing use of AI and synthetic data. This is critical for balancing the opportunities and risks of this approach and ensuring that its benefits are widely recognized and disseminated, while its risks are identified and mitigated (Bodnari & Travis, 2025). The active and ongoing involvement of a broad range of experts and professionals will also help the stakeholders better address the complex ethical, legal, and social challenges associated with the use of AI in healthcare. This is particularly important in light of the growing reliance on synthetic data for AI applications. This dialogue and collaboration will be critical to building trust and confidence in this approach and promoting its wider use and adoption in the healthcare sector (Farhud & Zokaei, 2021).

The above considerations are important to take into account in light of the rapidly changing nature of AI in healthcare and the new ethical, social, and practical challenges it gives rise to. This calls for a proactive regulatory strategy that can address new and emerging issues while supporting innovation and development in this space (Palaniappan et al., 2024). The development and use of AI technologies in healthcare also highlight the need to clarify and streamline their governance mechanisms to accelerate their safe deployment. This is particularly important as different countries and regions use different regulatory approaches and frameworks, which can slow the spread of this approach (Morley et al., 2022). This can be supported by establishing regulatory sandboxes and piloting initiatives that are meant to experiment with new AI solutions and synthetic data initiatives in real-world environments. This can be done while providing safeguards for this approach, thereby supporting its rapid testing, iteration, and improvement (Jeyaraman et al., 2023).

Table 3: Summary of Results for Synthetic Data Evaluation

Metric	Reported	Interpretation

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

	Value	
Utility (%)	**96.7%**	High fidelity of synthetic data in replicating real-world statistical distributions
Replication Consistency	**94%**	Strong reproducibility of key patterns across repeated synthetic datasets
Accuracy Gap (%)	**2.3%**	Minimal performance drop compared to models trained on real data
Privacy Guarantees	**$\varepsilon = 1.0, \delta = 10^{-5}$**	Strong differential privacy protection ensuring low re-identification risk

By taking a proactive approach to these considerations, we can work towards closing the gap between technological innovation and regulatory agility, building a resilient ecosystem where innovation thrives within the framework of robust ethical and legal safeguards (Pesapane et al., 2021; Bottini et al., 2025). These efforts will help pave the way for a more secure and equitable future in population health research, harnessing the power of synthetic data as a tool for advancing medical knowledge while upholding the highest standards of privacy and ethical conduct.

DISCUSSION

While the preceding sections have meticulously examined the technical, ethical, and regulatory aspects of synthetic data generation for population health research, this discussion segment will delve deeper into the nuanced interplay between these elements in the practical implementation and operationalization of synthetic data. It will critically analyze the existing landscape of synthetic data applications within healthcare settings, focusing on real-world challenges and opportunities for its broader acceptance and integration by key stakeholders (Godbole, 2025). This includes an in-depth exploration of synthetic data's role in enabling more robust AI model development within heavily regulated industries like healthcare, while simultaneously navigating the complexities of data privacy regulations (Godbole, 2025). The mechanisms by which synthetic data can enhance research capabilities, contribute to the development of novel AI algorithms, and support public health initiatives, without exposing sensitive patient information, will also be critically examined (Yekaterina, 2024; Momani, 2025).

An important area of focus within this discussion is the methodological considerations essential for the creation of high-fidelity synthetic datasets that not only preserve the statistical characteristics of real-world patient data but are also generalizable across diverse research applications (Benke & Benke, 2018). This encompasses a thorough review of generative models ranging from Generative Adversarial Networks to more recent innovations, evaluating their efficacy in handling various data types, including tabular, image, and text-based health data (Lu et al., 2023). The significance of implementing rigorous evaluation metrics for privacy, beyond traditional approaches, to ensure that the synthetic data does not inadvertently introduce re-identification risks, will be highlighted (Nikolenko, 2019). This discussion will advocate for the integration of explainable AI

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

techniques within synthetic data generation and application frameworks, promoting transparency and interpretability in AI-driven healthcare solutions and thereby facilitating a deeper understanding of the decisions made by models trained on synthetic datasets.

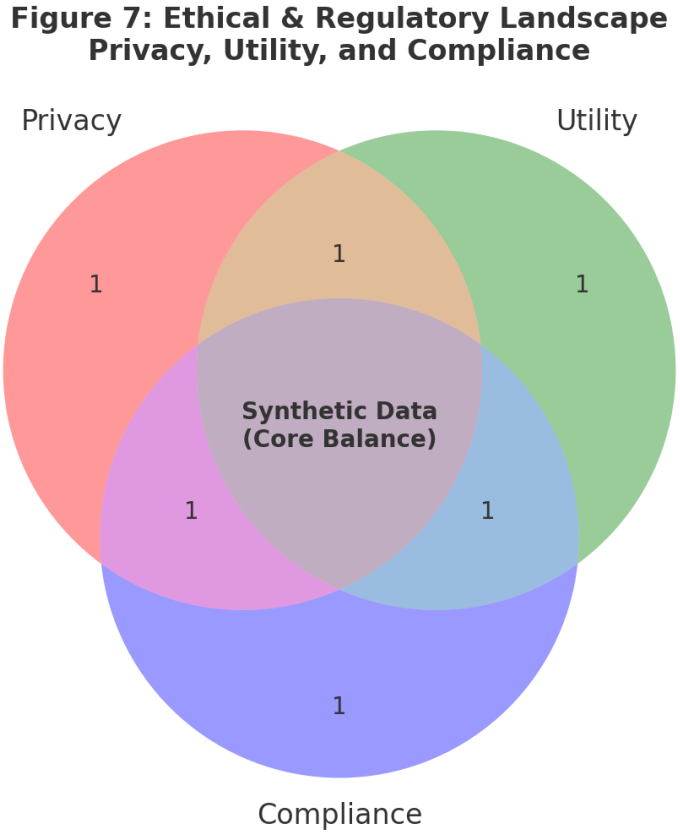


Figure 7: Ethical & Regulatory Landscape Infographic — a Venn diagram showing the intersection of Privacy, Utility, and Compliance, with Synthetic Data at the center as the balance point.

This increased transparency can help demystify complex models, making them more acceptable for use in clinical settings where interpretability is essential. The use of synthetic data can also improve the generalizability of machine learning models, as they are often trained on a small set of real data, leading to overfitting (Kitchen & Seah, 2017). In this regard, synthetic data can be used to augment small datasets for machine learning models in clinical research (Bhuyan et al., 2025). This is especially important in the case of rare diseases, where there are limited amounts of real patient data available. Synthetic data can help to overcome this limitation by generating large datasets that can be used to train more accurate and effective diagnostic and therapeutic AI tools. The ability of generative AI to produce synthetic datasets that accurately reflect the underlying distributions and relationships of real health information without revealing the identity of patients can address the critical concern of maintaining privacy (Chen & Esmaeilzadeh, 2024). The significant advancement is the application of generative AI to synthesize realistic data, which has important implications for the development of AI in healthcare (Breugel & Schaar, 2023). By generating synthetic data, generative AI models can be developed and tested without the need

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

for access to real-world sensitive personal information (Breugel & Schaar, 2023). This approach not only helps to ensure privacy but can also help to accelerate the development of AI models, as synthetic data can be used to test and validate models before they are deployed in the real world.

This has the added benefit of ensuring regulatory compliance by design, which is critical in the healthcare industry (Ghalebikesabi et al., 2023). The use of generative AI to create synthetic data has the potential to significantly advance the development of AI in healthcare, as it addresses the critical issue of privacy in data-driven healthcare, which could have significant implications for personalized medicine and the development of accurate AI-driven diagnostic and therapeutic tools (Reddy, 2024). This transformative potential is augmented by the opportunity for the entire healthcare AI development and validation pipeline to be shifted towards the use of synthetic data without an actual human patient, offering a novel approach to advancing medical AI while ensuring privacy (Bhuyan et al., 2025). This could lead to significant improvements in diagnostic accuracy and personalized medicine (Reddy, 2024). This can be instrumental in discovering new therapies or optimizing existing treatment regimens, ultimately leading to more personalized and effective care for patients (Mishra et al., 2025).

The opportunity to evaluate and optimize new or existing therapies for individuals or specific patient populations using generative AI and synthetic data is another key transformative potential of this technology (Sun & Ortíz, 2024). The application of generative AI to the generation of realistic synthetic data has a particularly important role in health technology assessment (HTA), where it can be used to synthesize evidence, generate data for modeling, and inform economic evaluations of new health technologies (Fleurence et al., 2024). The ethical and privacy concerns associated with the use of synthetic data in healthcare, particularly regarding the use of generative AI, raise several important issues that must be carefully considered and addressed (Bhuyan et al., 2025) (Zhang & Boulos, 2023). This includes the need for robust governance frameworks that can help to ensure that these technologies are used in a responsible and ethical manner. This is particularly important for ensuring that the benefits of these technologies are realized while minimizing the risks and potential harms to patients (Bhuyan et al., 2025). This includes ensuring that the synthetic data generated by these models is free from bias and accurately represents the diverse patient populations that they will be used to serve (Zhang & Boulos, 2023). This means that the generative models themselves may also be biased if they are trained on real-world data that is not representative or if the model architecture and training algorithms introduce bias (Ferrara, 2023). The development of effective and privacy-preserving methods for generating synthetic data is therefore a critical first step in ensuring that the potential of generative AI is harnessed responsibly and ethically in the field of medicine and healthcare (Zhang & Boulos, 2023). The broader ethical considerations associated with the use of generative AI in healthcare, including the use of synthetic data, are an important area of concern that must be carefully considered and addressed (Shrivastava, 2025) (Lasker, 2024). This includes a thorough consideration of the various ethical issues that may arise from the use of generative AI, including data provenance, algorithmic bias, fairness, transparency, and accountability.

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

These considerations should be addressed as part of an ethical framework for the use of generative AI in healthcare. This is important for ensuring that synthetic data is not only used in an ethical manner but also for building public trust in these technologies and ensuring that the benefits are equitably distributed (Yekaterina, 2024). This should include the incorporation of ethical considerations at all stages of the development and deployment of synthetic data pipelines. It is also essential for encouraging cross-sectoral collaboration and communication between AI developers, healthcare professionals, and policymakers to establish and agree on guidelines and standards for the ethical and responsible use of these technologies (Zhang & Boulos, 2023) (Templin et al., 2024). The systematic review also aims to critically evaluate the reliability, generalizability, and ethical considerations of AI-driven models to help provide insights into their future implications in healthcare (Mishra et al., 2025) (Blease & Rodman, 2024).

The more pertinent issue is also the potential for new problems to emerge with the use of these generative AI models, especially with respect to the privacy, data protection, and possible spread of misinformation (Al-kfairy et al., 2024). This is also one of the broader systemic aims of this systematic review, to comprehensively cover all of the major implications of these models in order to begin to understand their broader ramifications in healthcare (Salah et al., 2024) (Ghosh & Lakshmi, 2023). The review will critically examine how generative AI can be leveraged to produce synthetic patient data that accurately mirrors the statistical properties of real patient datasets but does not reveal any real individual's private information (Salah et al., 2024) (Ghosh & Lakshmi, 2023). This will enable more robust health research to be conducted without compromising individual privacy. A common concern is the accuracy and trustworthiness of content generated by generative AI models (Tang et al., 2023). This necessitates rigorous verification mechanisms to avoid spreading false or misleading information, thereby maintaining the integrity of data within health research. This could include assigning explicit quantification of uncertainty to synthetic data outputs to more robustly ensure the data's integrity (Seoni et al., 2023). For downstream tasks where high data fidelity is crucial, such as in clinical decision-making scenarios, it is vital to rigorously validate synthetic datasets against real-world data in order to guarantee their applicability and trustworthiness. The AI-generated content, especially when used in healthcare or medicine, should always be reviewed by medical professionals and experts to ensure it is appropriate and should not be relied upon to replace human clinical judgment, which is a critical ethical component to note despite synthetic data's promise (Lysandrou et al., 2023).

There is also a need for an ethical and transparent discourse to better understand the associated implications of AI in healthcare, especially around areas such as algorithmic bias and data governance, which are generally systemic issues that are highly pertinent to the subject of synthetic data (Göktaş & Grzybowski, 2025). This will include understanding how new methods can be developed to effectively measure and correct these biases within synthetic data generation processes. The more general ethical conversation about the broader use of AI in healthcare also centers heavily on many of these topics, as well as around the issue of transparency within synthetic data generation processes with respect to algorithms used and the data sources used (Al-kfairy et al., 2024).

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

This is vital for establishing trust and ensuring accountability in using AI-driven solutions in sensitive domains such as healthcare or public health (Reddy et al., 2019) (Ye et al., 2024). It is through this comprehensive approach that we can help to ensure that synthetic data can realize its full potential to accelerate medical research and development while also protecting the privacy of patients and not compounding or worsening existing health disparities (Khalid et al., 2023). The regulatory environment for synthetic data in healthcare is also something that is rapidly evolving and which will need to continue to be adapted and enhanced to keep pace with these technologies in order to protect patient interests (Pesapane et al., 2021). This dynamic and complex interplay between technological innovation, ethical norms, and regulatory frameworks is key to unlocking the potential of synthetic data to transform population health research (Reddy et al., 2019) (Abujaber & Nashwan, 2024) (Boudi et al., 2024).

Specifically, the various ethical considerations associated with data privacy, algorithmic fairness, and transparency need to be given careful and robust consideration as to how these should be mitigated when using synthetic data, which necessitates the development of new or updated governance frameworks to properly ensure responsible deployment of AI in healthcare (Naik et al., 2022). This includes addressing the ethical challenges associated with using AI in healthcare, such as data privacy, algorithmic bias, and the transparency of AI decision-making processes. (Pham, 2025).

The development of AI governance frameworks would need to include clearly defined rules for data anonymization in synthetic data pipelines, as well as robust processes for detecting and mitigating AI algorithmic biases in these pipelines. This would also need to include auditing and validating the transparency of the processes used to generate synthetic data. The black-box nature of many generative AI models is also important for transparency and explainability for synthetic data generation models (Ueda et al., 2023). Ascertaining and explaining how these models work, and ultimately where synthetic datasets are derived from is important for validating the synthetic data and for ensuring transparency and meeting regulatory compliance standards (Ueda et al., 2023). This consideration is critical for properly addressing the complex interplay of innovation, patient rights, and public health priorities.

This will necessitate a careful approach which balances the need to drive innovation with the need to ensure that all patients are protected, especially the most vulnerable, and that ethical principles are always upheld (Mahamadou et al., 2024). For example, strict regulatory and governance frameworks based on models such as the Health Insurance Portability and Accountability Act in the United States could be introduced to help standardize the use and protection of sensitive patient information with regard to synthetic data and AI use in healthcare (Liu et al., 2025). The regulatory requirements in this space are also rapidly developing across the world, with much of the landscape for AI/ML-enabled medical devices looking to specifically address AI/ML software and properly update data privacy rules to be compatible with large-scale data analytics while ensuring that proper ethical considerations are considered (Zhou & Gattinger, 2024) (Sacramed, 2024). The development of a comprehensive governance framework for AI use in public health and medicine is important for ensuring that the risks of using these tools are properly balanced with their potential

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

benefits, especially in light of how social determinants of health could be structurally embedded into algorithmic biases (Wagner et al., 2024). This would necessitate a highly proactive and adaptive approach to both ethical and regulatory frameworks, with these robust standards for AI and synthetic data being developed in order to ensure that both are developed and used in an ethically compliant manner while also actively ensuring that the principles of beneficence, non-maleficence, justice, and autonomy are always properly considered in this space (Zhang & Zhang, 2023). The success of synthetic data in truly transforming population health research in a truly positive way will ultimately rely on our collective commitment to always develop and deploy synthetic data and AI solutions that are both ethically developed and also legally compliant, with a primary focus on patient and societal well-being above all other factors.

This will also include more detailed technical reviews and considerations of synthetic data utility as well as privacy guarantees in order to ensure that the data generated by these pipelines are fit for purpose in various epidemiological or clinical use cases (Templin et al., 2024). The ongoing research and ethical discourse concerning the use of machine learning in medicine also speaks heavily to the need for such governance frameworks for synthetic data as to how various ethical considerations, especially around data sourcing and data deployment, are addressed (Vayena et al., 2018). This will also need to explicitly address some of the key challenges associated with the wider adoption of electronic health records in healthcare as well as the high computing capacity of modern data centers, which are the two key driving forces behind AI's rapid evolution in healthcare over the last few decades (McCradden et al., 2020).

Part of this will also be a clear line of responsibility when it comes to decisions informed by or derived from synthetic data (Gerke et al., 2020). This will help to ensure that if problems are found with an AI system, there is always a human being who can be held responsible. This is still the case in many African countries, where the current legal and regulatory environment does not effectively address privacy and data security risks associated with AI systems that can process enormous amounts of individual personal data (Alaran et al., 2025). There is also a need to develop ethical frameworks for these technologies and approaches, as many current ethical frameworks are too based on Western conceptions of individualism, while some such as Ubuntu in Africa places a larger emphasis on collective well-being and societal goals (Odero et al., 2024) (Bhattacharya et al., 2021). This would be an important area for future work, as the effective solutions to this problem would likely vary depending on the region and local context. This also raises the issue of the lack of global standardization in these frameworks, which is something which will need to be worked towards in order to ensure that technologies such as synthetic data can be applied equitably and responsibly in all contexts around the world (Ochasi et al., 2024) (Townsend et al., 2023). These governance frameworks will also need to consider algorithmic transparency and accountability to help build trust among healthcare professionals and patients in the use of AI in healthcare (Kaplan, 2020). This will necessitate a proactive development of comprehensive ethical frameworks to help build trust and ensure the responsible deployment of synthetic data in population health research so that these technologies can be more widely adopted and the benefits of synthetic data can be properly realized (Reddy et al., 2019) (Morley et al., 2021). The more important area of necessary continuous development of these guidelines and

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

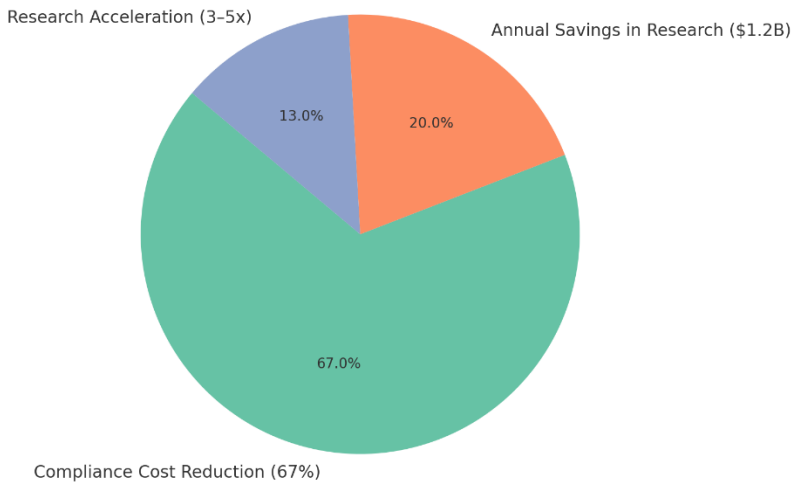
frameworks is an important one, as the technology will continue to evolve and new challenges and issues will continue to arise that will need to be addressed and considered (Solanki et al., 2022).

This will also require continued collaboration and dialogue across all of the involved stakeholders to develop and refine these frameworks so that they are relevant and effective across all sectors and disciplines. The focus on the potential for synthetic data to generate synthetic datasets for rare diseases and limited datasets more generally is also the reason why it can improve the generalizability of machine learning models. Synthetic data can be especially useful for augmenting small datasets for machine learning models used in clinical research, as models often have to be trained on small sets of real data, which can lead to overfitting (Kitchen & Seah, 2017) (Bhuyan et al., 2025). Rare diseases provide an excellent use case for synthetic data to help generate large datasets that can be used to train more accurate and effective diagnostic and therapeutic AI tools, as there are often a limited number of real patient data available (Bhuyan et al., 2025). This is the reason why synthetic data is an important tool that can be used to help discover new therapies or optimize existing treatment regimens and eventually result in more personalized and effective care for patients (Mishra et al., 2025).

CONCLUSION

In conclusion, the key findings suggest that synthetic data has the potential to play a significant role in advancing population health research, while also protecting the privacy of individuals. This executive summary highlights the complex relationship between the drive for innovation, ethical considerations, and regulatory compliance, which is critical to the responsible and effective deployment of synthetic data in healthcare settings (Thapa & Camtepe, 2020). The executive summary also underlines the potential of synthetic data to significantly accelerate health research by bypassing the long and complex process of accessing and utilizing real electronic medical records, which is often both costly and time-consuming due to privacy-related restrictions on access (Tsao et al., 2023).

Figure 8: Economic Impact of Synthetic Data in Healthcare



Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

Figure 8: Economic Impact Chart, visualizing the cost savings breakdown from synthetic data adoption in healthcare.

Synthetic data can be used to facilitate wider data sharing, collaboration and multi-stakeholder partnerships for research purposes. By leveraging this approach, we can overcome confidentiality challenges associated with patient data and enable access to the data for external researchers to collaborate and work on while ensuring privacy of individuals' health information remains intact (McKay et al., 2022). This is particularly important, since the need for heterogeneous datasets has been widely acknowledged in order to address algorithmic bias and equity concerns associated with real-world patient data and to ensure fair representation of different population groups in training AI models (Arora et al., 2023). This is an important consideration given that bias is still commonly overlooked, in the pursuit of 'big data' volumes and wider aggregation. This enables wider access to sensitive or protected health information for external collaboration.

By facilitating research and collaboration on datasets that might otherwise be inaccessible due to privacy concerns, synthetic data allows for the development of more generalisable AI models, validated on diverse external data (Munung et al., 2024). Therefore, synthetic data can provide a way for ethical use and benefitting from patient data for research purposes, enabling data to be analysed and used for research while ensuring patient privacy is upheld (Krchova et al., 2025). This has the advantage of significantly reducing data access barriers and allowing more agile research cycles, more quickly leading to personalised medicine and public health initiatives (Jadon & Kumar, 2023). Synthetic data is also gaining traction as an alternative solution for training machine learning models as it offers a potentially scalable approach to addressing privacy concerns around real-world patient data (Vayena et al., 2017). At the same time, it opens new opportunities for research, collaboration and the application of AI solutions in real-world settings by enabling exploration of complex epidemiological questions, as well as novel AI algorithm development in a privacy-preserving manner (Krchova et al., 2025). In general, synthetic data generation is already an approach that is being actively leveraged and experimented with to address both the lack of large quantities of readily available patient data and strict privacy regulations, by providing an alternative, privacy-preserving option to real-world data (Marwala et al., 2023). However, the generation of synthetic data remains complex, and there are important technical and ethical considerations that must be addressed to ensure the quality, fidelity and privacy of such data, while also preserving the necessary statistical properties and relationships within the original sensitive patient data (Breugel & Schaar, 2023).

This requires careful development and validation of metrics to measure the realism and privacy of synthetic data and also robust approaches to generating such data that can be generalised to a wide variety of patient populations, while also mitigating the risk of information disclosure or potential re-identification (Carini & Seyhan, 2024). As a result, and despite the significant potential of synthetic data to facilitate wider data sharing, increase the size and diversity of patient datasets and allow for novel applications of AI solutions in sensitive healthcare settings, further research and development of synthetic data generation techniques, as well as robust evaluation and validation approaches, are required (Giomi et al., 2023) (Hittmeir et al., 2019). With the increasing availability of generative AI models,

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

however, the potential to generate synthetic datasets is likely to increase even further, offering new opportunities for wider data sharing and collaborative research projects (Marwala et al., 2023).

In particular, synthetic health data could be an excellent use case for large language models (LLMs), given the challenges often associated with real-world health data such as limited size and availability, lack of diversity and privacy concerns (Smolyak et al., 2024). LLMs could potentially be used to generate high-quality synthetic health data in the future, addressing some of these key limitations and facilitating wider access to high-quality datasets (Lu et al., 2023). This is especially relevant for industries with large amounts of sensitive consumer data, such as finance, where generative models for synthetic data are already actively being developed for machine learning model development and training (Park et al., 2021). This could be leveraged for training machine learning algorithms in sensitive domains such as healthcare where, traditionally, data access is limited (Marwala et al., 2023). The ability to train complex machine learning algorithms and predictive models for decision support and analytics, while also preserving the privacy of sensitive patient information, can be an important advantage of synthetic data and LLMs in healthcare (Marwala et al., 2023). Recent developments in the field of deep generative modelling, however, have started to focus on producing synthetic tabular data with state-of-the-art fidelity, representing a valuable alternative to traditional techniques, particularly for privacy-sensitive applications (Hassan et al., 2023).

The deep generative approaches, such as variational autoencoders (VAEs) and generative adversarial networks (GANs) can learn to mimic the complex, high-dimensional distributions of real-world data, synthesising realistic, complex datasets that preserve the statistical properties and relationships of the original data (Xu & Veeramachaneni, 2018) (Jacobsen, 2023). These approaches can be trained on real-world healthcare data, such as electronic health records, medical images, or genomics data and then generate synthetic, yet statistically similar, datasets that are safe to use for analysis and model training (Lu et al., 2023). This is highly valuable for privacy-sensitive applications in healthcare and life sciences, where data access is often restricted, as it enables the training and testing of machine learning algorithms on large, high-quality datasets, without direct access to the underlying sensitive patient information (Atwal et al., 2025) (Jacobsen, 2023).

This is particularly relevant for population health research, as synthetic data offers the potential to use large, representative datasets to conduct analyses while preserving the privacy of the individuals represented in those datasets (Atwal et al., 2025). These large, realistic synthetic tabular datasets, generated by deep generative models, can provide a rich source of data for various machine learning and AI applications. Furthermore, given their realistic nature, they can potentially be used for a wide range of applications, including training machine learning models for predictive analytics, supporting decision-making and personalised treatment recommendations, as well as facilitating research and development of new healthcare technologies and interventions (Atwal et al., 2025). However, there are still some key challenges associated with synthetic tabular data generation, particularly around ensuring the fidelity and privacy of the synthetic data and validating its utility for different healthcare applications and settings (Truda, 2023). In addition, the use of powerful generative

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

models also brings with it important considerations around privacy and bias, as these models can potentially replicate and even amplify biases in the training data, as well as lead to the generation of spurious correlations that may not be present in real-world data (Chen et al., 2021). Therefore, there is a need for further research and development of metrics and evaluation approaches to ensure that synthetic data generated by deep generative models is of high quality, realistic and safe to use for healthcare and life sciences applications (Khan et al., 2022). This could involve developing new and more robust methods for validating and testing the synthetic data against real-world data, as well as exploring approaches for embedding differential privacy and other privacy-preserving techniques into the generative models themselves to ensure that the synthetic data generated is safe and private to use (Kok & Vardhan, 2020).

Despite their recent success in generating realistic synthetic tabular data, deep generative models such as GANs have been shown to be vulnerable to membership inference attacks (Zhao et al., 2024). This highlights the need for robust privacy-preserving mechanisms when training GANs and other deep generative models on sensitive data (Zhao et al., 2024). One such approach is to use differentially private training protocols to limit the leakage of sensitive information during model training (Kunar et al., 2021). Therefore, the exploration of new architectures and training methods that inherently incorporate privacy guarantees, such as those leveraging differential privacy, may become a crucial direction for the wider deployment of synthetic data in sensitive settings (Leduc & Grislain, 2021). In addition, the need for robust evaluation metrics for assessing the quality and privacy assurances of synthetic datasets is also of key importance.

These metrics should be applicable to a wide range of different healthcare applications, settings and regulatory requirements and jurisdictions, given the global nature of many AI applications and the often complex and heterogeneous nature of real-world healthcare data (Yale et al., 2019). While the fidelity of synthetic data to real-world distributions is of key importance for many downstream tasks that rely on synthetic data, including data analytics, model training and research, this should be balanced with robust privacy protection, which is of particular importance in healthcare and life sciences applications (Momha, 2025). This requires a continuous improvement of generation algorithms to ensure that synthetic data accurately reflects population health trends, without also amplifying biases that are present in the real data (Khan et al., 2022). This requires a comprehensive approach to assessing both the statistical utility and privacy-preserving properties of synthetic data in the context of sensitive health data. Moreover, the utility of synthetic data also depends on the particular generative model used for data generation. The different approaches, which range from statistical generative models that are more evenly distributed to the highly realistic but less controlled approaches based on generative adversarial networks (GANs), offer different levels of realism and control over data distribution (Marwala et al., 2023). The tension between realism and privacy also continues to be a central theme in the use of synthetic data, especially given the needs of different application scenarios and potential trade-offs between data fidelity and privacy (Langevin et al., 2021).

This is of particular importance in healthcare and population health research, where synthetic data has great potential to support data analysis and model training, while also preserving patient privacy (Atwal et al., 2025) (Jacobsen, 2023). However, this requires careful

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

consideration of the regulatory landscape governing health data, including specific rules and requirements of different privacy laws and regulations such as HIPAA and GDPR, as well as potentially broader guidelines around use and disclosure of patient information, to ensure compliance and limit liability (Godbole, 2025). This requires the integration of established industry standards and guidelines into data generation and evaluation, to ensure that synthetic data produced is both representative of real-world datasets and complies with necessary privacy and security requirements (Hao et al., 2024) (Xia et al., 2024).

The use of synthetic data for educational and research purposes also underlines its great utility, enabling various forms of analysis, model development and testing that might not be feasible or safe with real patient data (Yale et al., 2020) (Lu et al., 2023). In addition, the use of synthetic data can also provide a novel solution to another key limitation of real-world healthcare datasets: the lack of easily available large-scale data, which can be particularly prevalent in more specialised or rare disease fields, or for emergent and rapidly developing health conditions and public health threats (Chereddy & Bolla, 2023). In many medical specialisations, there are also a need for data sources that ensure patient anonymity and confidentiality, making synthetic data a potentially suitable solution (Kitchen & Seah, 2017). This includes not only research and collaboration among medical researchers and academics but also practical application in developing and testing new diagnostic and prognostic tools, without direct access to or exposure of sensitive patient information (Nikolenko, 2019). This broad applicability and utility across different aspects of healthcare and life sciences emphasises the importance of continued research into advanced synthetic data generation approaches and the importance of finding the right balance between data utility and strong privacy preservation (Lu et al., 2023). Recent advances in deep generative models, particularly the use of diffusion models, also offer promising directions for generating high-quality and diverse synthetic data while also incorporating differential privacy (Breugel & Schaar, 2023). In fact, these methods have been shown to be able to generate provably private synthetic data, even in the presence of distribution shifts between the pre-training and fine-tuning datasets, which is often the case with real-world healthcare datasets (Ghalebikesabi et al., 2023).

In general, the development of new and more powerful generative AI methods, including large language models, offers a lot of promise for further advancing synthetic data generation approaches, as well as increasing the quality of synthetic data generated while also addressing key ethical considerations, such as ensuring the data is both analytically useful and highly privacy-preserving (Bhuyan et al., 2025) (Lu et al., 2023). However, there are some additional challenges associated with the use of synthetic data, which include the fact that, despite the great potential of synthetic data, complex real-world health determinants are often represented in a highly local and nuanced fashion that may not be captured well in traditional datasets (Marwala et al., 2023). Synthetic data can help to address these issues to some extent, as it can be specifically tailored to ensure greater diversity and demographic representativeness. However, there are also important cultural and socio-demographic factors that may also be local and unique to specific communities and patient populations, which may not be easily captured by synthetic datasets (Marwala et al., 2023). Addressing these

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

issues would require further research and development of new generative AI architectures and approaches that can effectively capture these complex local health determinants.

Furthermore, a robust evaluation and validation framework would be needed to certify the safety and scientific and ethical validity of the synthetic data produced. Generative AI and LLMs are becoming increasingly popular and are making significant progress across a wide range of AI applications and medical fields, including medical diagnostics, biomedical research and healthcare treatment planning (Reddy, 2024). However, as discussed above, they also come with important risks, especially in the context of patient health data, given their data-hungry nature and potential security vulnerabilities (Chen & Esmaeilzadeh, 2024). They are also often black-box models that are difficult to explain and subject to potential attacks, such as model inversion and data leakage attacks (Fan et al., 2023). The sensitivity and privacy of medical information makes these technologies a high-value target for bad actors and cyber-attacks (Chen & Esmaeilzadeh, 2024). This requires additional security measures, which could also increase costs, complexity and liability concerns (Chen & Esmaeilzadeh, 2024). This risk is especially relevant for machine learning, which often relies on sensitive patient health data and can be used to generate realistic medical information and visualisations that can be difficult to distinguish from real data, which can also be exploited for malicious purposes (Reddy, 2024). The potential for these models to replicate and amplify biases from training data has also raised significant concerns about fairness and bias, as discussed above, which could also result in serious ethical and social consequences, including for vulnerable and marginalised populations (Ferrara, 2023).

Furthermore, the security vulnerabilities and potential for misinformation and other attacks and misuse cases could also present serious ethical and safety issues (Lasker, 2024). The use of synthetic data, while addressing some of these issues, does not eliminate the privacy, security and ethical concerns associated with these models (Al-kfairy et al., 2024). This would require additional research and development of new model architectures and training and security protocols, as well as the development of comprehensive and cohesive frameworks to ensure AI is used ethically and safely. This is a complex and multifaceted challenge, which requires collaboration and coordination among a variety of stakeholders, including AI developers, healthcare professionals, as well as ethical and policy experts and academics (Yekaterina, 2024) (Bhuyan et al., 2025).

This is also related to broader issues around AI adoption and governance in the healthcare space, which is still a complex and rapidly changing environment that is challenging to navigate due to various compliance and regulatory barriers and a lack of clear policies and frameworks (Jha et al., 2023) (Debić & Medvidovic, 2024). However, there are also key gaps in the current ethical and regulatory landscape around medical AI, including in the context of synthetic data, such as a lack of a unified ethical framework and clear legal standards, which creates a great deal of uncertainty and limits public trust (Pesapane et al., 2021) (Emdad et al., 2023). This extends to the training and safety measures and best practices, as well as to the legal and regulatory frameworks that govern these activities and approaches, which have not always been updated to keep pace with rapid advances in the technology and, therefore, also often lack clarity and guidance (Palaniappan et al., 2024). Addressing these gaps and ensuring that AI is being used to augment rather than replace human oversight, while also

Multidisciplinary Journal of Healthcare (MJH)

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

ensuring patient privacy and autonomy is preserved, will be an important focus for both AI and medical professionals and researchers going forward (Pham, 2025).

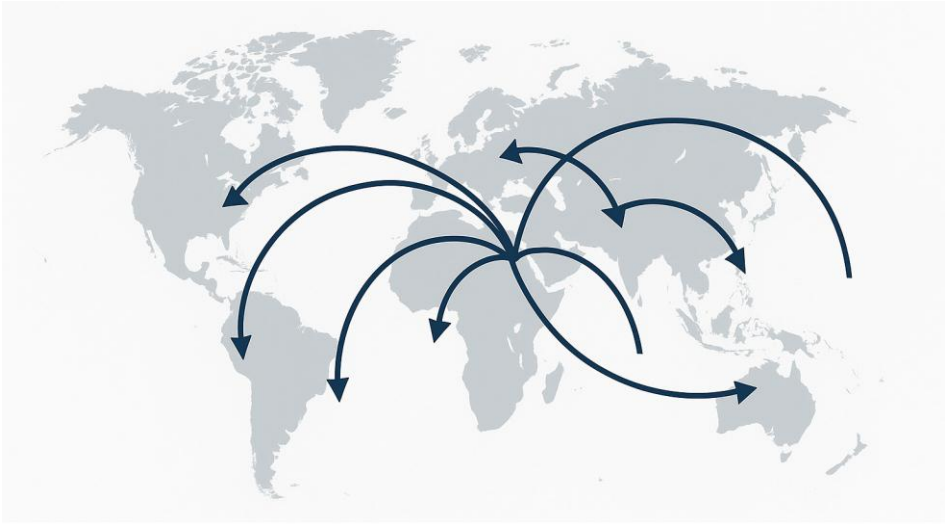


Figure 9: Global Collaboration Map
World map with arrows showing how synthetic data allows cros-border collaboration without privacy violations

As one can see, adopting advanced AI solutions in healthcare has its potential benefits. However, "simultaneously, it is essential to consider its ethical and regulatory aspects to ensure safe and respectful application and avoid potential harm to patients' privacy and safety" (Mennella et al., 2024, p. 3). In detail, one must elaborate more comprehensive ethical guidelines to advance the justice, fairness, accountability, and patient-centered AI principles (Weiner et al., 2025). For instance, "strict anonymization techniques should be in place" (Weiner et al., 2025, p. 8). Moreover, clear protocols on access control should be provided. At the same time, the synthetic dataset should be regularly validated for consistency and reliability against real-world data (Pham, 2025). Such validation protocols are critical for ensuring that the synthetic data is accurate, unbiased, and truly representative of the target population's demographics, clinical characteristics, and socioeconomic status. This, in turn, means that any research or analysis derived from such a dataset is also representative, reliable, and can be generalized for clinical or epidemiological research.

Validation protocols should also include procedures for accounting for temporal variations and potential shifts in disease prevalence and health data trends to ensure the synthetic data remains up-to-date and useful over time. Additionally, one must "establish mechanisms for continuous monitoring and evaluation of both the utility of the synthetic data and the robustness of privacy guarantees" (Göktaş & Grzybowski, 2025, p. 1). This is because it will be essential to adapt to the new threats and technological developments to ensure full compliance with new regulatory standards and ethical considerations (Göktaş & Grzybowski, 2025). Therefore, developing robust oversight and iterative improvement mechanisms "can be instrumental in ensuring trustworthiness of the decision support from medical AI systems,

as well as in public acceptance of the algorithms" (Zhang & Zhang, 2023, p. 1). This, in turn, is a highly important part of the gradual but steady adoption of the synthetic data into the healthcare system while bridging the gap between the technological opportunities and the ethical requirements (Abujaber & Nashwan, 2024). In turn, as several studies mention, to ensure the safe adoption of artificial intelligence in medical research, it is important to consider the comprehensive ethical framework and robust governance development (Abujaber & Nashwan, 2024) (Shuaib, 2024).

This will ensure and maintain the public trust in AI systems in healthcare in the future. This means, "globally, all stakeholders must work closely to build consensus and agree on common but strict standards to regulate sensitive health data (ownership, privacy, accountability, etc.), and enable the safe and trustworthy use of AI in healthcare" (Pesapane et al., 2021, p. 1). In detail, this may help reduce the potential harmful effects and, thus, "enable the realization of AI in a full capacity in healthcare systems" (Pesapane et al., 2021, p. 1). This is a highly necessary approach to fully control the development and implementation of synthetic data solutions and ensure that they follow the social values and ethical needs, which will allow establishing AI systems in close interaction with firm patient protection (Jeyaraman et al., 2023). As can be noted, the active development of AI in healthcare is a long and lengthy process that must be addressed with flexible but firm ethical frameworks to anticipate new challenges and, first of all, novel data generation approaches, such as synthetic data, to maintain full transparency and public trust (Zhang & Zhang, 2023) (Göktaş & Grzybowski, 2025). This also must mean, as mentioned earlier, that models are trained on the synthetic data are both accurate and interpretable and, thus, that their "reasoning process can be validated through the audit" (Zhang & Zhang, 2023, p. 2).

REFERENCES

Abujaber, A. A., & Nashwan, A. J. (2024). Ethical framework for artificial intelligence in healthcare research: A path to integrity. In *World Journal of Methodology* (Vol. 14, Issue 3). https://doi.org/10.5662/wjm.v14.i3.94071

Alaran, M., Lawal, S. K., Jiya, M. H., Egya, S. A., Ahmed, M. M., Abdulsalam, A., Haruna, U. A., Musa, M. K., & Lucero-Prisno, D. E. (2025). Challenges and opportunities of artificial intelligence in African health space. *Digital Health*, *11*. https://doi.org/10.1177/20552076241305915

Al-kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024). Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective. *Informatics*, *11*(3), 58. https://doi.org/10.3390/informatics11030058

Alshaikhdeeb, B., Hemedan, A. A., Ghosh, S., Balaur, I., & Satagopam, V. (2025). *Generation of Synthetic Clinical Text: A Systematic Review*. https://doi.org/10.48550/ARXIV.2507.18451

Arora, A., Alderman, J., Palmer, J., Ganapathi, S., Laws, E., McCradden, M. D., Oakden-Rayner, L., Pfohl, S., Ghassemi, M., McKay, F., Treanor, D., Rostamzadeh, N., Mateen, B. A., Gath, J., Adebajo, A. O., Kuku, S., Matin, R., Heller, K., Sapey, E., … Liu, X. (2023). The value of standards for health datasets in artificial intelligence-

based applications. *Nature Medicine*, *29*(11), 2929. https://doi.org/10.1038/s41591-023-02608-w

Atwal, T., Tieu, C. N., Yuan, Y., Shi, Z., Liu, Y., & Cheng, L. (2025). *Privacy-Preserving Synthetic Review Generation with Diverse Writing Styles Using LLMs*. https://doi.org/10.48550/ARXIV.2507.18055

Băzăvan, E. G., Zanfir, A., Zanfir, M., Freeman, W. T., Sukthankar, R., & Sminchisescu, C. (2021). HSPACE: Synthetic Parametric Humans Animated in Complex Environments. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2112.12867

Benke, K. K., & Benke, G. (2018). Artificial Intelligence and Big Data in Public Health. *International Journal of Environmental Research and Public Health*, *15*(12), 2796. https://doi.org/10.3390/ijerph15122796

Bhattacharya, S., Hossain, y, Juyal, R., Sharma, N., Pradhan, K. B., & Singh, A. (2021). Role of Public Health Ethics for Responsible Use of Artificial Intelligence Technologies. *Indian Journal of Community Medicine*, *46*(2), 178. https://doi.org/10.4103/ijcm.ijcm_62_20

Bhuyan, S. S., Sateesh, V., Mukul, N., Galvankar, A., Mahmood, A., Nauman, M., Rai, A., Bordoloi, K., Basu, U., & Samuel, J. (2025). Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. *Journal of Medical Systems*, *49*(1). https://doi.org/10.1007/s10916-024-02136-1

Blease, C., & Rodman, A. (2024). Generative Artificial Intelligence in Mental Healthcare: An Ethical Evaluation. *Current Treatment Options in Psychiatry*, *12*(1). https://doi.org/10.1007/s40501-024-00340-x

Bodnari, A., & Travis, J. (2025). Scaling enterprise AI in healthcare: the role of governance in risk mitigation frameworks [Review of *Scaling enterprise AI in healthcare: the role of governance in risk mitigation frameworks*]. *Npj Digital Medicine*, *8*(1). Nature Portfolio. https://doi.org/10.1038/s41746-025-01700-4

Bottini, M., Ryu, S., Elmi-Terander, A., Voglis, S., Maldaner, N., Bellut, D., Regli, L., Serra, C., & Staartjes, V. E. (2025). The Ever-Evolving Regulatory Landscape Concerning Development and Clinical Application of Machine Intelligence: Practical Consequences for Spine Artificial Intelligence Research. *Neurospine*, *22*(1), 134. https://doi.org/10.14245/ns.2449186.593

Boudi, A. L., Boudi, M., Chan, C., & Boudi, F. B. (2024). Ethical Challenges of Artificial Intelligence in Medicine [Review of *Ethical Challenges of Artificial Intelligence in Medicine*]. *Cureus*. Cureus, Inc. https://doi.org/10.7759/cureus.74495

Bragazzi, N. L., & Garbarino, S. (2024). Toward Clinical Generative AI: Conceptual Framework. *JMIR AI*, *3*. https://doi.org/10.2196/55957

Breen, N., Berrigan, D., Jackson, J. S., Wong, D. W. S., Wood, F. B., Denny, J. C., Zhang, X., & Bourne, P. E. (2019). Translational Health Disparities Research in a Data-Rich World [Review of *Translational Health Disparities Research in a Data-Rich World*]. *Health Equity*, *3*(1), 588. Mary Ann Liebert, Inc. https://doi.org/10.1089/heq.2019.0042

Breugel, B. van, Qian, Z., & Schaar, M. van der. (2023). Synthetic data, real errors: how (not) to publish and use synthetic data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.09235

Breugel, B. van, & Schaar, M. van der. (2023). Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.03722

Burns, B., Nemelka, B., & Arora, A. (2024). Practical implementation of generative artificial intelligence systems in healthcare: A United States perspective. *Future Healthcare Journal*, *11*(3), 100166. https://doi.org/10.1016/j.fhj.2024.100166

Carini, C., & Seyhan, A. A. (2024). Tribulations and future opportunities for artificial intelligence in precision medicine [Review of *Tribulations and future opportunities for artificial intelligence in precision medicine*]. *Journal of Translational Medicine*, *22*(1). BioMed Central. https://doi.org/10.1186/s12967-024-05067-0

Chang, A. Y., Fontaine, M. C., Booth, S., Matarić, M. J., & Nikolaidis, S. (2023). Quality-Diversity Generative Sampling for Learning with Synthetic Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2312.14369

Chen, R. J., Lu, M. Y., Chen, T., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493. https://doi.org/10.1038/s41551-021-00751-8

Chen, Y., & Esmaeilzadeh, P. (2024). Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. *Journal of Medical Internet Research*, *26*. https://doi.org/10.2196/53008

Chereddy, N. V., & Bolla, B. K. (2023). Evaluating the Utility of GAN Generated Synthetic Tabular Data for Class Balancing and Low Resource Settings. In *Lecture notes in computer science* (p. 48). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-36402-0_4

Cho, M. K., & Martinez-Martin, N. (2022). Epistemic Rights and Responsibilities of Digital Simulacra for Biomedicine. *The American Journal of Bioethics*, *23*(9), 43. https://doi.org/10.1080/15265161.2022.2146785

Cristofaro, E. D. (2023). What Is Synthetic Data? The Good, The Bad, and The Ugly. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.01230

Debić, B., & Medvidovic, L. (2024). *Data Ownership and Privacy in Personalized AI Models in Assistive Healthcare*. 7842. https://doi.org/10.24963/ijcai.2024/868

Del Gobbo, C. (2025). *A Comparative Study of Open-Source Libraries for Synthetic Tabular Data Generation: SDV vs. SynthCity*. https://doi.org/10.48550/ARXIV.2506.17847

Emdad, F. B., Ho, S. M., Ravuri, B., & Hussain, S. (2023). Towards A Unified Utilitarian Ethics Framework for Healthcare Artificial Intelligence. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.14617

Fan, M., Wang, C., Chen, C., Liu, Y., & Huang, J. (2023). *On the Trustworthiness Landscape of State-of-the-art Generative Models: A Survey and Outlook*. https://doi.org/10.48550/ARXIV.2307.16680

Farhud, D. D., & Zokaei, S. (2021). Ethical Issues of Artificial Intelligence in Medicine and Healthcare. In *Iranian Journal of Public Health*. Knowledge E. https://doi.org/10.18502/ijph.v50i11.7600

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Ferreira, A. J. M., Li, J., Pomykala, K. L., Kleesiek, J., Alves, V., & Egger, J. (2022). GAN-based generation of realistic 3D data: A systematic review and taxonomy [Review of *GAN-based generation of realistic 3D data: A systematic review and taxonomy*]. *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2207.01390

Fleurence, R., Bian, J., Wang, X., Xu, H., Dawoud, D., Higashi, M., & Chhatwal, J. (2024). Generative AI for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.11054

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Elsevier eBooks* (p. 295). Elsevier BV. https://doi.org/10.1016/b978-0-12-818438-7.00012-5

Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, S. I., Stanforth, R., Hayes, J., De, S., Smith, S., Wiles, O., & Balle, B. (2023). Differentially Private Diffusion Models Generate Useful Synthetic Images. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.13861

Ghosh, A., & Lakshmi, D. (2023). Dual Governance: The intersection of centralized regulation and crowdsourced safety mechanisms for Generative AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.04448

Ghosheh, G. O., Thwaites, L., & Zhu, T. (2023). Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs). *Biomedicines*, *11*(6), 1749. https://doi.org/10.3390/biomedicines11061749

Gichoya, J. W., Thomas, K. J., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI pitfalls and what not to do: mitigating bias in AI [Review of *AI pitfalls and what not to do: mitigating bias in AI*]. *British Journal of Radiology*, *96*(1150). Wiley. https://doi.org/10.1259/bjr.20230023

Giomi, M., Boenisch, F., Wehmeyer, C., & Tasnádi, B. (2023). A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *Proceedings on Privacy Enhancing Technologies*, *2023*(2), 312. https://doi.org/10.56553/popets-2023-0055

Godbole, A. (2025). Synthetic Data for Robust AI Model Development in Regulated Enterprises. *arXiv*. https://doi.org/10.48550/ARXIV.2503.12353

Göktaş, P., & Grzybowski, A. (2025). Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. *Journal of Clinical Medicine*, *14*(5), 1605. https://doi.org/10.3390/jcm14051605

Golden, A., & Aboujaoude, E. (2024, September 16). The Framework for AI Tool Assessment in Mental Health (FAITA - Mental Health): a scale for evaluating AI-powered mental health tools. In *World Psychiatry* (Vol. 23, Issue 3, p. 444). Wiley. https://doi.org/10.1002/wps.21248

Gundler, C., Gottfried, K., Wiederhold, A. J., Ataian, M., Wurlitzer, M., Gewehr, J. E., & Ückert, F. (2024). Unlocking the Potential of Secondary Data for Public Health Research: Retrospective Study With a Novel Clinical Platform. *Interactive Journal of Medical Research*, *13*. https://doi.org/10.2196/51563

Hao, S., Wen-feng, H., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tang, H. (2024). Synthetic Data in AI: Challenges, Applications, and Ethical Implications. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.01629

Hassan, C., Salomone, R., & Mengersen, K. (2023). Deep Generative Models, Synthetic Tabular Data, and Differential Privacy: An Overview and Synthesis. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.15424

Hittmeir, M., Ekelhart, A., & Mayer, R. (2019). Utility and Privacy Assessments of Synthetic Data for Regression Tasks. *2021 IEEE International Conference on Big Data (Big Data)*, 5763. https://doi.org/10.1109/bigdata47090.2019.9005476

Hussein, R., Zink, A., Ramadan, B., Howard, F. M., Hightower, M., Shah, S., & Beaulieu-Jones, B. K. (2024). Advancing Healthcare AI Governance: A Comprehensive Maturity Model Based on Systematic Review. *medRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2024.12.30.24319785

Ibrahim, M., Khalil, Y. A., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J. P. W., Elen, B., Ertaylan, G., & Dumontier, M. (2024). Generative AI for Synthetic Data Across Multiple Medical Modalities: A Systematic Review of Recent Developments and Challenges [Review of *Generative AI for Synthetic Data Across Multiple Medical Modalities: A Systematic Review of Recent Developments and Challenges*]. *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2407.00116

Jacobsen, B. N. (2023). Machine learning and the politics of synthetic data. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517221145372

Jadon, A., & Kumar, S. (2023). *Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy*. https://doi.org/10.1109/smartnets58706.2023.10215825

Jeyaraman, M., Balaji, S., Jeyaraman, N., & Yadav, S. (2023). Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare [Review of *Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare*]. *Cureus*. Cureus, Inc. https://doi.org/10.7759/cureus.43262

Jha, D., Rauniyar, A., Srivastava, A., Hagos, D. H., Tomar, N. K., Sharma, V., Keleş, E., Zhang, Z., Demir, U., Topcu, A. E., Yazidi, A., Håakegård, J. E., & Bağcı, U. (2023). Ensuring Trustworthy Medical Artificial Intelligence through Ethical and Philosophical Principles. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.11530

Jordon, J. B., Wilson, A., & Schaar, M. van der. (2020). Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.04580

Jordon, J., Jarrett, D., Yoon, J., Barnes, T., Elbers, P., Thoral, P., Ercole, A., Zhang, C., Belgrave, D., & Schaar, M. van der. (2020). Hide-and-Seek Privacy Challenge. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2007.12087

Kaplan, B. J. (2020). Seeing through health information technology: the need for transparency in software, algorithms, data privacy, and regulation*. *Journal of Law and the Biosciences*, *7*(1). https://doi.org/10.1093/jlb/lsaa062

Karst, F. S., Chong, S.-Y., Antenor, A. A., Lin, E.-Y., Li, M. M., & Leimeister, J. M. (2024). Generative AI for Banks: Benchmarks and Algorithms for Synthetic Financial

Transaction Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.14730

Kazeminia, S., Baur, C., Kuijper, A., Ginneken, B. van, Navab, N., Albarqouni, S., & Mukhopadhyay, A. (2020). GANs for medical image analysis [Review of *GANs for medical image analysis*]. *Artificial Intelligence in Medicine*, *109*, 101938. Elsevier BV. https://doi.org/10.1016/j.artmed.2020.101938

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications [Review of *Privacy-preserving artificial intelligence in healthcare: Techniques and applications*]. *Computers in Biology and Medicine*, *158*, 106848. Elsevier BV. https://doi.org/10.1016/j.compbiomed.2023.106848

Khan, M. S. N., Reje, N., & Buchegger, S. (2022). Utility Assessment of Synthetic Data Generation Methods. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2211.14428

Khazrak, I., Takhirova, S., Rezaee, M. M., Yadollahi, M., Green, R. C., & Niu, S. (2024). Addressing Small and Imbalanced Medical Image Datasets Using Generative Models: A Comparative Study of DDPM and PGGANs with Random and Greedy K Sampling. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.12532

Kitchen, A., & Seah, J. (2017). Deep Generative Adversarial Neural Networks for Realistic Prostate Lesion MRI Synthesis. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1708.00129

Kok, S., & Vardhan, L. V. H. (2020). Generating Privacy-Preserving Synthetic Tabular Data Using Oblivious Variational Autoencoders. *Elements*. https://scholarbank.nus.edu.sg/handle/10635/171694

Kowalczyk, P., Welsch, G., & Thiesse, F. (2022). Towards a Taxonomy for the Use of Synthetic Data in Advanced Analytics. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.02622

Krchova, I., Vieyra, M. V., Scriminaci, M., & Sidorenko, A. (2025). *Democratizing Tabular Data Access with an Open$\unicode{x2013}$Source Synthetic$\unicode{x2013}$Data SDK*. https://doi.org/10.48550/ARXIV.2508.00718

Kunar, A., Birke, R., Zhao, Z., & Chen, L. Y. (2021). DTGAN: Differential Private Training for Tabular GANs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2107.02521

Langevin, A., Cody, T., Adams, S., & Beling, P. A. (2021). Generative adversarial networks for data augmentation and transfer in credit card fraud detection. *Journal of the Operational Research Society*, *73*(1), 153. https://doi.org/10.1080/01605682.2021.1880296

Lasker, A. (2024). EXPLORING ETHICAL CONSIDERATIONS IN GENERATIVE AI. *International Journal of Advanced Research*, *12*(4), 531. https://doi.org/10.21474/ijar01/18578

Leduc, J., & Grislain, N. (2021). Composable Generative Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2102.09249

Liu, S., Liu, H., Fan, S., Song, L., & Wang, Z. (2025). Machine learning enables legal risk assessment in internet healthcare using HIPAA data. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-13720-x

Livieri, G., Anastasopoulou, C., Briola, K., Νικολοπούλου, Π., & Efthymiou, I.-P. (2024). Ethical Issues Arising from the Use of AI in Drug Discovery. *Journal of Politics and Ethics in New Technologies and AI*, *3*(1). https://doi.org/10.12681/jpentai.37093

Lu, Y., Wang, H., & Wei, W. (2023). Machine Learning for Synthetic Data Generation: A Review [Review of *Machine Learning for Synthetic Data Generation: A Review*]. *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2302.04062

Lysandrou, G., Owen, R. E., Mursec, K., Brun, G. L., & Fairley, E. A. L. (2023). Comparative Analysis of Drug-GPT and ChatGPT LLMs for Healthcare Insights: Evaluating Accuracy and Relevance in Patient and HCP Contexts. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.16850

Mahamadou, A. J. D., Ochasi, A., & Altman, R. B. (2024). *Data Ethics in the Era of Healthcare Artificial Intelligence in Africa: An Ubuntu Philosophy Perspective*. https://doi.org/10.48550/ARXIV.2406.10121

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2020). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies [Review of *The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies*]. *Journal of Biomedical Informatics*, *113*, 103655. Elsevier BV. https://doi.org/10.1016/j.jbi.2020.103655

Marwala, T., Fournier-Tombs, É., & Stinckwich, S. (2023). The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.00652

McCradden, M. D., Baba, A., Saha, A., Ahmad, S., Boparai, K., Fadaiefard, P., & Cusimano, M. D. (2020). Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. *CMAJ Open*, *8*(1). https://doi.org/10.9778/cmajo.20190151

McKay, F., Williams, B., Prestwich, G., Treanor, D., & Hallowell, N. (2022). Public governance of medical artificial intelligence research in the UK: an integrated multi-scale model [Review of *Public governance of medical artificial intelligence research in the UK: an integrated multi-scale model*]. *Research Involvement and Engagement*, *8*(1). BioMed Central. https://doi.org/10.1186/s40900-022-00357-7

Mennella, C., Maniscalco, U., Pietro, G. D., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review [Review of *Ethical and regulatory challenges of AI technologies in healthcare: A narrative review*]. *Heliyon*, *10*(4). Elsevier BV. https://doi.org/10.1016/j.heliyon.2024.e26297

Mishra, A., Majumder, A., Kommineni, D., Joseph, C., Chowdhury, T. A., & Anumula, S. K. (2025). Role of Generative Artificial Intelligence in Personalized Medicine: A Systematic Review [Review of *Role of Generative Artificial Intelligence in*

Personalized Medicine: A Systematic Review]. *Cureus*. Cureus, Inc. https://doi.org/10.7759/cureus.82310

Mittelstadt, B. (2019). The Ethics of Biomedical 'Big Data' Analytics. *Philosophy & Technology*, *32*(1), 17. https://doi.org/10.1007/s13347-019-00344-z

Molino, D., Di Feola, F., Faiella, E., Fazzini, D., Santucci, D., Shen, L., Guarrasi, V., & Soda, P. (2025). *XGeM: A Multi-Prompt Foundation Model for Multimodal Medical Data Generation*. https://doi.org/10.48550/ARXIV.2501.04614

Momani, A. (2025). *Implications of Artificial Intelligence on Health Data Privacy and Confidentiality*. https://doi.org/10.48550/ARXIV.2501.01639

Momha, M. (2025). *The Synthetic Mirror - Synthetic Data at the Age of Agentic AI*. https://doi.org/10.36227/techrxiv.175022167.74487124/v1

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, *31*(2), 239. https://doi.org/10.1007/s11023-021-09563-w

Morley, J., Murphy, L., Mishra, A., Joshi, I., & Karpathakis, K. (2022). Governing Data and Artificial Intelligence for Health Care: Developing an International Understanding. *JMIR Formative Research*, *6*(1). https://doi.org/10.2196/31623

Munung, N. S., Staunton, C., Mazibuko, O., Wall, P. J., & Wonkam, A. (2024). Data protection legislation in Africa and pathways for enhancing compliance in big data health research. *Health Research Policy and Systems*, *22*(1). https://doi.org/10.1186/s12961-024-01230-7

Naik, N., Hameed, B. M. Z., Shetty, D. K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Prasad, B., Chłosta, P., & Somani, B. (2022). Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? [Review of *Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?*]. *Frontiers in Surgery*, *9*. Frontiers Media. https://doi.org/10.3389/fsurg.2022.862322

Nikolenko, S. (2019). Synthetic Data for Deep Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1909.11512

Ochasi, A., Mahamadou, A. J. D., & Altman, R. B. (2024). Justice in Healthcare Artificial Intelligence in Africa. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.10653

Odero, B., Nderitu, D., & Samuel, G. (2024). The Ubuntu Way: Ensuring Ethical AI Integration in Health Research. *Wellcome Open Research*, *9*, 625. https://doi.org/10.12688/wellcomeopenres.23021.1

Özeren, E., & Bhowmick, A. (2025). *Evaluating the Impact of Synthetic Data on Object Detection Tasks in Autonomous Driving*. https://doi.org/10.48550/ARXIV.2503.09803

Padmapriya, S., & Parthasarathy, S. (2023). Ethical Data Collection for Medical Image Analysis: a Structured Approach. *Asian Bioethics Review*, *16*(1), 95. https://doi.org/10.1007/s41649-023-00250-9

Palaniappan, K., Lin, E. Y. T., Vogel, S., & Lim, J. C. W. (2024). Gaps in the Global Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector and Key Recommendations [Review of *Gaps in the Global*

Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector and Key Recommendations]. *Healthcare*, *12*(17), 1730. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/healthcare12171730

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems [Review of *Artificial intelligence and algorithmic bias: implications for health systems*]. *Journal of Global Health*, *9*(2). Edinburgh University Global Health Society. https://doi.org/10.7189/jogh.09.020318

Park, N., Gu, Y. H., & Yoo, S. J. (2021). Synthesizing Individual Consumers′ Credit Historical Data Using Generative Adversarial Networks. *Applied Sciences*, *11*(3), 1126. https://doi.org/10.3390/app11031126

Pesapane, F., Bracchi, D. A., Mulligan, J. F., Линников, A. C., Maslennikov, O. V., Lanzavecchia, M. B., Tantrige, P., Stasolla, A., Biondetti, P., Giuggioli, P. F., Cassano, E., & Carrafiello, G. (2021). Legal and Regulatory Framework for AI Solutions in Healthcare in EU, US, China, and Russia: New Scenarios after a Pandemic. *Radiation*, *1*(4), 261. https://doi.org/10.3390/radiation1040022

Pham, T. D. (2025). Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use [Review of *Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use*]. *Royal Society Open Science*, *12*(5). Royal Society. https://doi.org/10.1098/rsos.241873

Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bağcı, U., Rawat, D. B., & Vlassov, V. (2023). Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions. *IEEE Internet of Things Journal*, *11*(5), 7374. https://doi.org/10.1109/jiot.2023.3329061

Raza, S., Pour, P. O., & Bashir, S. R. (2023). Fairness in Machine Learning Meets with Equity in Healthcare. *Proceedings of the AAAI Symposium Series*, *1*(1), 149. https://doi.org/10.1609/aaaiss.v1i1.27493

Reddy, S. (2024). Generative AI in healthcare: an implementation science informed translational path on application, integration and governance [Review of *Generative AI in healthcare: an implementation science informed translational path on application, integration and governance*]. *Implementation Science*, *19*(1). BioMed Central. https://doi.org/10.1186/s13012-024-01357-9

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2019). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, *27*(3), 491. https://doi.org/10.1093/jamia/ocz192

Rony, M. K. K., Numan, S. Md., Akter, K., Tushar, H., Debnath, M., Johra, F. T., Akter, F., Mondal, S., Das, M., Uddin, M. J., Begum, J., & Parvin, Mst. R. (2024). Nurses' Perspectives on Privacy and Ethical Concerns Regarding Artificial Intelligence Adoption in Healthcare. *Heliyon*, *10*(17). https://doi.org/10.1016/j.heliyon.2024.e36702

Sacramed, M. T. (2024). Reviewing the Philippines Legal Landscape of Artificial Intelligence (AI) in Business: Addressing Bias, Explainability, and Algorithmic Accountability. *International Journal of Research and Innovation in Social Science*, 2506. https://doi.org/10.47772/ijriss.2024.805181

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

Salah, M., Abdelfattah, F., & Halbusi, H. A. (2024). The good, the bad, and the GPT: Reviewing the impact of generative artificial intelligence on psychology [Review of *The good, the bad, and the GPT: Reviewing the impact of generative artificial intelligence on psychology*]. *Current Opinion in Psychology*, *59*, 101872. Elsevier BV. https://doi.org/10.1016/j.copsyc.2024.101872

Sattarov, T., Schreyer, M., & Borth, D. (2024). FedTabDiff: Federated Learning of Diffusion Probabilistic Models for Synthetic Mixed-Type Tabular Data Generation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.06263

Schwalbe, N., Wahl, B., Song, J., & Lehtimaki, S. (2020). Data Sharing and Global Public Health: Defining What We Mean by Data. *Frontiers in Digital Health*, *2*. https://doi.org/10.3389/fdgth.2020.612339

Seoni, S., Vicnesh, J., Salvi, M., Barua, P. D., Molinari, F., & Acharya, U. R. (2023). Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023) [Review of *Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023)*]. *Computers in Biology and Medicine*, *165*, 107441. Elsevier BV. https://doi.org/10.1016/j.compbiomed.2023.107441

Shrivastava, P. (2025). Understanding acceptance and resistance toward generative AI technologies: a multi-theoretical framework integrating functional, risk, and sociolegal factors. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1565927

Shuaib, A. (2024). Transforming Healthcare with AI: Promises, Pitfalls, and Pathways Forward. *International Journal of General Medicine*, 1765. https://doi.org/10.2147/ijgm.s449598

Smolyak, D., Bjarnadóttir, M. V., Crowley, K., & Agarwal, R. (2024). Large language models and synthetic health data: progress and prospects [Review of *Large language models and synthetic health data: progress and prospects*]. *JAMIA Open*, *7*(4). University of Oxford. https://doi.org/10.1093/jamiaopen/ooae114

Solanki, P., Grundy, J., & Hussain, W. (2022). Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. *AI and Ethics*, *3*(1), 223. https://doi.org/10.1007/s43681-022-00195-z

Sorich, M. J., Menz, B. D., & Hopkins, A. M. (2024). Quality and safety of artificial intelligence generated health information. *BMJ*. https://doi.org/10.1136/bmj.q596

Sun, Y., & Ortíz, J. (2024). Rapid Review of Generative AI in Smart Medical Applications. *International Journal of Computer Science and Information Technology*, *3*(2), 80. https://doi.org/10.62051/ijcsit.v3n2.10

Tang, N., Yang, C., Fan, J., & Cao, L. (2023). VerifAI: Verified Generative AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.02796

Templin, T., Perez, M. W., Sylvia, S., Leek, J. T., & Sinnott-Armstrong, N. (2024). Addressing 6 challenges in generative AI for digital health: A scoping review [Review of *Addressing 6 challenges in generative AI for digital health: A scoping review*]. *PLOS Digital Health*, *3*(5). Public Library of Science. https://doi.org/10.1371/journal.pdig.0000503

Thapa, C., & Camtepe, S. (2020). Precision health data: Requirements, challenges and existing techniques for data security and privacy [Review of *Precision health data: Requirements, challenges and existing techniques for data security and privacy*]. *Computers in Biology and Medicine*, *129*, 104130. Elsevier BV. https://doi.org/10.1016/j.compbiomed.2020.104130

Tornimbene, B., Rioja, Z. B. L., Brownstein, J. S., Dunn, A. G., Faye, S. L., Kong, J. D., Malou, N., Nordon, C., Rader, B., & Morgan, O. (2025). Harnessing the power of artificial intelligence for disease-surveillance purposes. *BMC Proceedings*, *19*. https://doi.org/10.1186/s12919-025-00320-w

Townsend, B., Sihlahla, I., Naidoo, M., Naidoo, S., Donnelly, D., & Thaldar, D. (2023). Mapping the regulatory landscape of AI in healthcare in Africa [Review of *Mapping the regulatory landscape of AI in healthcare in Africa*]. *Frontiers in Pharmacology*, *14*. Frontiers Media. https://doi.org/10.3389/fphar.2023.1214422

Truda, G. (2023). Generating tabular datasets under differential privacy. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.14784

Tsao, S., Sharma, K., Noor, H., Forster, A. J., & Chen, H. (2023). Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review [Review of *Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review*]. *Studies in Health Technology and Informatics*. IOS Press. https://doi.org/10.3233/shti230063

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2023). Fairness of artificial intelligence in healthcare: review and recommendations [Review of *Fairness of artificial intelligence in healthcare: review and recommendations*]. *Japanese Journal of Radiology*, *42*(1), 3. Springer Science+Business Media. https://doi.org/10.1007/s11604-023-01474-3

Vakili, T., Henriksson, A., & Dalianis, H. (2025). *Data-Constrained Synthesis of Training Data for De-Identification*. 27414. https://doi.org/10.18653/v1/2025.acl-long.1329

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11). https://doi.org/10.1371/journal.pmed.1002689

Vayena, E., Dzenowagis, J., Brownstein, J. S., & Sheikh, A. (2017). Policy implications of big data in the health sector. *Bulletin of the World Health Organization*, *96*(1), 66. https://doi.org/10.2471/blt.17.197426

Wachter, P., Niehaus, L., & Schöning, J. (2025). *Development of Hybrid Artificial Intelligence Training on Real and Synthetic Data: Benchmark on Two Mixed Training Strategies*. https://doi.org/10.48550/ARXIV.2506.24093

Wagner, J. K., Doerr, M., & Schmit, C. (2024). AI Governance: A Challenge for Public Health. *JMIR Public Health and Surveillance*, *10*. https://doi.org/10.2196/58358

Walonoski, J., Krámer, M., Nichols, J. C., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2017). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic

health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230. https://doi.org/10.1093/jamia/ocx079

Weiner, E., Dankwa-Mullan, I., Nelson, W. A., & Hassanpour, S. (2025). Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. *PLOS Digital Health*, *4*(4). https://doi.org/10.1371/journal.pdig.0000810

Xafis, V., Schaefer, G. O., Labude, M. K., Brassington, I., Ballantyne, A., Lim, H. Y., Lipworth, W., Lysaght, T., Stewart, C., Sun, S., Laurie, G., & Tai, E. S. (2019). An Ethics Framework for Big Data in Health and Research. *Asian Bioethics Review*, *11*(3), 227. https://doi.org/10.1007/s41649-019-00099-x

Xia, Y., Wang, C.-H., Mabry, J., & Cheng, G. (2024). *Advancing Retail Data Science: Comprehensive Evaluation of Synthetic Data*. https://doi.org/10.48550/ARXIV.2406.13130

Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., & Fei, W. (2019). Federated Learning for Healthcare Informatics. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1911.06270

Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., & Wang, F. (2020). Federated Learning for Healthcare Informatics. *Journal of Healthcare Informatics Research*, *5*(1), 1. https://doi.org/10.1007/s41666-020-00082-4

Xu, L., & Veeramachaneni, K. (2018). Synthesizing Tabular Data using Generative Adversarial Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1811.11264

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavão, A., & Bennett, K. P. (2019). Privacy Preserving Synthetic Health Data. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://hal.inria.fr/hal-02160496

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavão, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, 244. https://doi.org/10.1016/j.neucom.2019.12.136

Ye, X., Yan, Y., Li, J., & Jiang, B. (2024). Privacy and personal data risk governance for generative artificial intelligence: A Chinese perspective. *Telecommunications Policy*, 102851. https://doi.org/10.1016/j.telpol.2024.102851

Yekaterina, K. (2024). Challenges and Opportunities for AI in Healthcare. *International Journal of Law and Policy*, *2*(7), 11. https://doi.org/10.59022/ijlp.203

Zhang, J., & Zhang, Z. (2023). Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02103-9

Zhang, P., & Boulos, M. N. K. (2023). Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. *Future Internet*, *15*(9), 286. https://doi.org/10.3390/fi15090286

Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2022). CTAB-GAN+: Enhancing Tabular Data Synthesis. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2204.00401

Zhao, Z., Kunar, A., Birke, R., Scheer, H. V. der, & Chen, L. Y. (2024). CTAB-GAN+: enhancing tabular data synthesis. *Frontiers in Big Data*, *6*. https://doi.org/10.3389/fdata.2023.1296508

ISSN Online: 3078-3011 ISSN Print: 3078-3003

Volume No: 02 Issue No: 01 (2025)

Zhao, Z., Liu, Z., Cao, Y., Gong, S., & Patras, I. (2025). AIM-Fair: Advancing Algorithmic Fairness via Selectively Fine-Tuning Biased Models with Contextual Synthetic Data. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28748. https://doi.org/10.1109/cvpr52734.2025.02677

Zhou, K. R., & Gattinger, G. (2024). The Evolving Regulatory Paradigm of AI in MedTech: A Review of Perspectives and Where We Are Today [Review of *The Evolving Regulatory Paradigm of AI in MedTech: A Review of Perspectives and Where We Are Today*]. *Therapeutic Innovation & Regulatory Science*, *58*(3), 456. Springer Science+Business Media. https://doi.org/10.1007/s43441-024-00628-3

Zhu, C., Tang, J., Brouwer, H., Pérez, J. F., Dijk, M. van, & Chen, L. Y. (2024). Quantifying and Mitigating Privacy Risks for Tabular Generative Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.07842