Review

Exploring the Role of Synthetic Data in the Future of AI in Healthcare: A Scoping Review of Frameworks, Challenges, and Implications

Mohammad Ishtiaque Rahman * , Razuan Hossain , S.M. Sayem , Forhan Bin Emdad

Posted Date: 5 August 2025

doi: 10.20944/preprints202507.2567.v2

Keywords: synthetic data; healthcare AI; challenges; privacy; bias

Review

Exploring the Role of Synthetic Data in the Future of AI in Healthcare: A Scoping Review of Frameworks, Challenges, and Implications

Mohammad Ishtiaque Rahman [1,*], Razuan Hossain [2], S.M. Sayem [3] and Forhan Bin Emdad [4]

[1] Thomas More University
[2] Utah Valley University
[3] Bangladesh University of Professionals
[4] Governors State University
* Correspondence: rahmanm@thoasmore.edu

Abstract

Synthetic data has emerged as a transformative tool in healthcare, particularly in areas such as medical imaging, electronic health records (EHRs), and clinical trial simulation, where data privacy, diversity, and accessibility are critical. This scoping review examines current approaches to synthetic data generation in healthcare, with a focus on AI model training, privacy preservation, and bias mitigation. A comprehensive search of PubMed, IEEE Xplore, and ACM Digital Library yielded 2,906 studies, of which 42 met the inclusion criteria. Key data generation techniques included generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, Bayesian networks, federated learning, recurrent neural networks (RNNs), large language models (LLMs), agent-based models, graph-based generators, and SMOTE-based oversampling. Applications ranged from diagnostic model development to privacy-preserving data sharing and educational simulation. However, the field faces persistent challenges, including inconsistent validation practices, the absence of standard benchmarks, high computational demands, and ethical concerns related to consent and bias. This review underscores the need for standardized evaluation protocols, clearer regulatory guidance, and multidisciplinary collaboration to ensure the safe, equitable, and effective use of synthetic data in healthcare AI.

Keywords: synthetic data; healthcare ai; challenges; privacy; bias

Introduction

Synthetic data is artificially created data that mimics real-world data. Unlike real data collected from actual events or users, synthetic data is generated using algorithms, simulations, or statistical methods.[1] It is designed to resemble real data in structure, patterns, and properties while not containing any actual, identifiable personal or sensitive information.[2] This makes it especially useful in the medical field, where the protection of patient privacy is paramount. Synthetic data is employed in various healthcare domains, including oncology, neurology, and cardiology.[3] It addresses challenges such as data collection issues and regulatory constraints, facilitating the development of AI technologies in the sector.

The generation of synthetic data ranges from simple rule-based systems to machine-learning algorithms and complex Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Synthetic data can take many forms, such as patient records, medical images, or genetic information. However, the generation of synthetic data that retains the original properties while keeping the sensitive information private is very challenging. Diligent effort is necessary to ensure the data remains useful for research purposes but reduces the risk of exposing sensitive patterns or reinforcing biases.

The use of synthetic data is growing quickly. According to Gartner, by 2025, it's expected that about 60% of all data used to train artificial intelligence models will be synthetic.[4] In healthcare, synthetic data already supports more than 40% of AI-powered clinical trials and diagnostic tools. This shows the increasing trust and dependence on synthetic data in creating new solutions. Experts predict the market for synthetic data will grow to over $2 billion by 2026, highlighting its importance for both research and businesses. The global market for synthetic data is expanding rapidly, expected to grow from $351.2 million in 2023 to $2.34 billion by 2030, with an annual growth rate of 31.1%.[5] Healthcare is a big part of this growth, using synthetic data for research, clinical trials, and medical imaging. Gartner also predicts that by 2024, 60% of the data used in AI and analytics will be synthetic, helping solve problems like protecting patient privacy and dealing with limited real data, especially in healthcare.[6]

Synthetic data offers substantial benefits for healthcare AI. Primarily, it eliminates privacy concerns by excluding real patient information, enabling seamless data sharing across hospitals, research institutions, and borders while complying with regulations like GDPR and HIPAA.[1] Additionally, it addresses real-world data limitations, such as underrepresentation of specific demographics or rare diseases, by generating diverse, balanced datasets that enhance AI model fairness and accuracy.[7,8] Furthermore, its scalability allows for large-volume production, providing reusable resources to overcome data shortages in research and training.[2]

Synthetic data has rapidly gained traction in healthcare research as a solution to data privacy, data scarcity, and fairness challenges, particularly in areas such as medical imaging, electronic health records, and clinical simulation. It holds significant potential to improve data accessibility, safeguard patient privacy, and foster innovation in artificial intelligence applications. Although many studies have investigated specific generative techniques such as generative adversarial networks, variational autoencoders, and diffusion models, there remains limited consolidated understanding of how these methods are applied across different healthcare settings.[9] Critical concerns such as inconsistent validation practices, ethical issues related to consent and data ownership, risks of bias amplification, and the potential for re-identification from highly realistic synthetic datasets remain insufficiently addressed.[10] Existing reviews often concentrate on isolated technical approaches or narrowly defined applications, without incorporating broader ethical, regulatory, and practical considerations. Given the variety of methods and the complexity of their applications, a scoping review is well suited to map the current landscape, identify key research trends, and highlight areas requiring further investigation. This approach allows for a comprehensive synthesis of the literature and can inform future efforts to develop standardized evaluation protocols, ethical guidelines, and collaborative frameworks for the responsible use of synthetic data in healthcare artificial intelligence. The objective of this scoping review is to map the current landscape of synthetic data generation and use in healthcare artificial intelligence. Specifically, the review aims to:

(1) Examine the methods and models used to generate synthetic data in healthcare contexts.
(2) Explore the range of applications across domains such as medical imaging, clinical records, and research simulation.
(3) Identify key limitations, ethical concerns, and risks including privacy, bias, and validation challenges.
(4) Assess the broader implications of synthetic data for the future development and deployment of AI in healthcare.

Methods

This study employed a scoping review[11] to investigate AI-driven synthetic data generation in healthcare, focusing on its methodologies, applications, benefits, challenges, and future implications for AI development. A scoping review was chosen for its structured approach to synthesizing evidence from diverse, credible sources, ensuring a comprehensive and unbiased analysis of this rapidly evolving field. PRISMA ScR was followed for data collection and synthesis process.[12] No review protocol was developed or registered for this scoping review, as the review was intended as an exploratory synthesis of a rapidly evolving and heterogeneous body of literature. Given the breadth of technologies and applications under investigation, a flexible and iterative review design was considered more appropriate than a fixed protocol.

Search Strategy

A comprehensive literature search was conducted across three academic databases: PubMed, IEEE Xplore, and the ACM Digital Library. The search aimed to identify relevant studies published between January 2010 and December 2024. These databases were selected for their broad and complementary coverage of medical, engineering, and computational research relevant to healthcare artificial intelligence. The search was performed using a combination of controlled vocabulary, such as Medical Subject Headings in PubMed, and free-text keywords, joined by Boolean operators to enhance retrieval. The final database search was executed on January 2, 2025. No additional sources such as grey literature, preprints, or manual reference checks were included, and no contact was made with study authors to obtain further information. Search terms included:

PubMed: (("Artificial Intelligence"[MeSH] OR "Machine Learning"[MeSH] OR "Generative Adversarial Networks" OR "Variational Autoencoders" OR "Federated Learning") AND ("Synthetic Data"[All Fields] OR "Synthetic Data Generation"[All Fields] OR "Privacy-Preserving Synthetic Data"[All Fields] OR "Synthetic Health Records"[All Fields]) AND ("Healthcare"[MeSH] OR "Electronic Health Records"[MeSH] OR "Medical Informatics"[MeSH] OR "Health Information Systems"[MeSH] OR "Clinical Data"[All Fields]))

IEEE Explore: ((Artificial Intelligence OR Machine Learning OR Generative Adversarial Networks OR Variational Autoencoders OR Federated Learning) AND (Synthetic Data OR Synthetic Data Generation OR Privacy-Preserving Synthetic Data OR Synthetic Health Records) AND (Healthcare OR Electronic Health Records OR Medical Informatics OR Health Information Systems OR Clinical Data))

ACM Digital Library: ("Artificial Intelligence" OR "Machine Learning" OR "Generative Adversarial Networks" OR "Variational Autoencoders" OR "Federated Learning") AND ("Synthetic Data" OR "Synthetic Data Generation" OR "Privacy-Preserving Synthetic Data" OR "Synthetic Health Records") AND ("Healthcare" OR "Electronic Health Records" OR "Medical Informatics" OR "Health Information Systems" OR "Clinical Data"))

Inclusion and Exclusion Criteria

Studies were included if they (1) applied artificial intelligence techniques such as generative adversarial networks, variational autoencoders, or federated learning to generate synthetic data; (2) addressed healthcare contexts including clinical records, imaging, or population-level simulation; and (3) provided methodological or evaluative insights related to data quality, utility, privacy, or ethical concerns. Only peer-reviewed articles published in English between January 2010 and December 2024 were considered. The time frame was selected to capture the evolution of modern AI-based data generation techniques.

Exclusion Criteria: (1) Non-English publications; (2) Studies unrelated to healthcare applications of synthetic data; (3) Articles lacking specific methodological or application details.

Study Selection Process

A total of 2,906 records were retrieved from the database searches. After removing 947 duplicates and 1,064 ineligible records through automation tools and manual filtering, 895 unique studies were screened by title and abstract. Of these, 145 full-text articles were assessed for eligibility. Sixty-three studies were excluded at this stage for reasons including language (n = 15), lack of synthetic data focus (n = 20), or insufficient methodological detail (n = 18). Ultimately, 42 studies met all inclusion criteria and were included in the final synthesis. The selection process is illustrated in the PRISMA-ScR flow diagram.[12]

Data Extraction

A standardized form was used to extract key data from each study, including: (1) Author(s), publication year, and journal; (2) Synthetic data generation techniques (e.g., GANs, VAEs); (3) Healthcare applications (e.g., EHRs, imaging); (4) Evaluation metrics (e.g., fidelity, utility, privacy); (5) Challenges, limitations, and solutions; (6) Ethical considerations. Data extraction was performed by one reviewer and verified by a second to ensure accuracy.

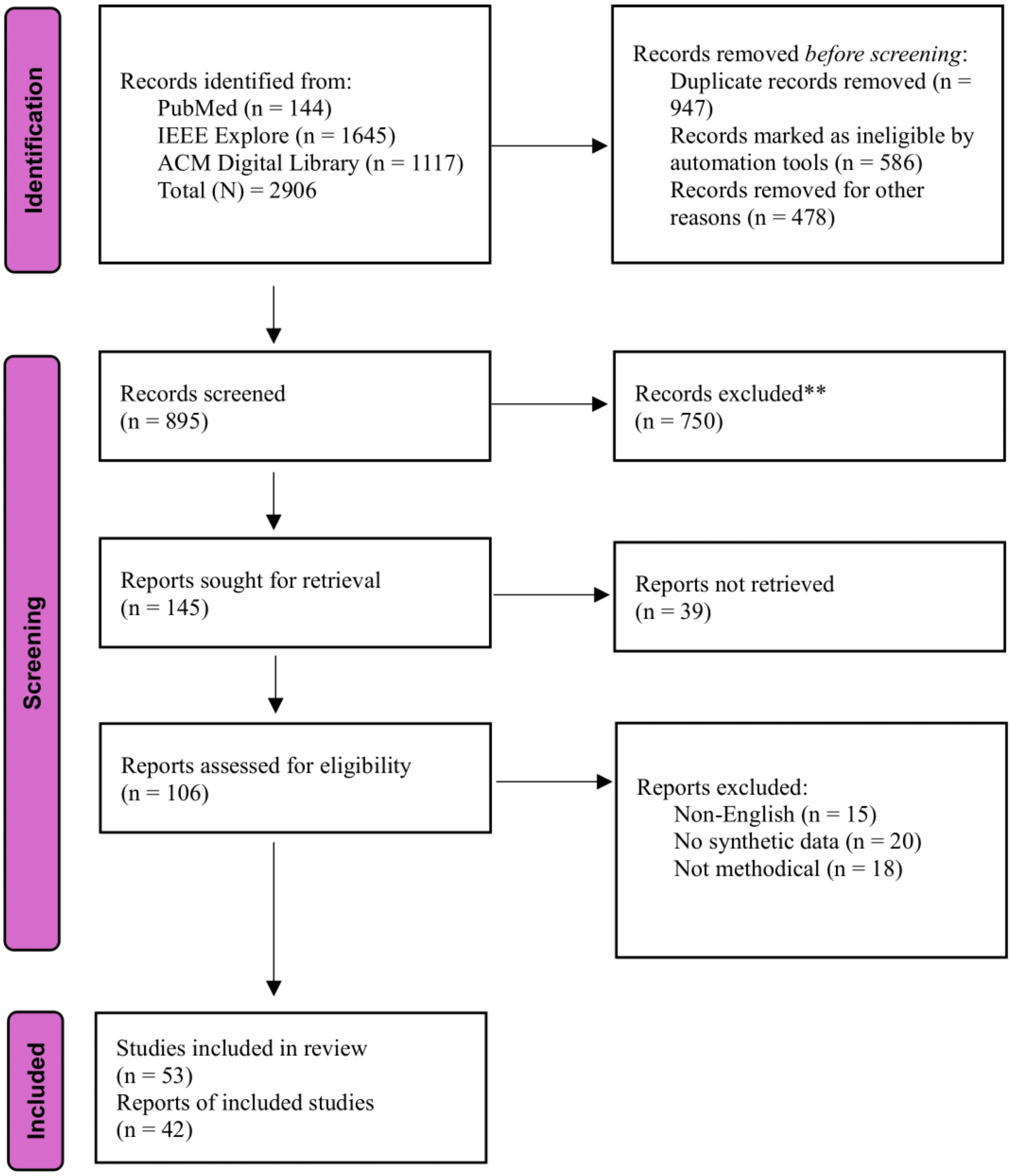
Identification
Records identified from:
PubMed (n = 144)
IEEE Explore (n = 1645)
ACM Digital Library (n = 1117)
Total (N) = 2906
Records removed *before screening*:
Duplicate records removed (n = 947)
Records marked as ineligible by automation tools (n = 586)
Records removed for other reasons (n = 478)
Screening
Records screened
(n = 895)
Records excluded**
(n = 750)
Reports sought for retrieval
(n = 145)
Reports not retrieved
(n = 39)
Reports assessed for eligibility
(n = 106)
Reports excluded:
Non-English (n = 15)
No synthetic data (n = 20)
Not methodical (n = 18)
Included
Studies included in review
(n = 53)
Reports of included studies
(n = 42)

Figure 1. PRISMA flowchart of study identification and selection.

Synthesis of Results

Data from the included studies were charted using a standardized extraction form developed by the review team. One reviewer completed the initial extraction, which was then verified by a second reviewer for accuracy. Findings were synthesized narratively to identify key themes: (1) AI-driven generation methods, (2) evaluation metrics for quality, utility, and privacy, (3) challenges such as bias and scalability, and (4) implications for future research and practice. Quantitative data, where available, were tabulated to compare reported metrics across studies, providing a structured overview of progress and limitations. This approach helped identify evidence gaps and informed recommendations for advancing the responsible use of synthetic data in healthcare AI.

Results

Characteristics of Included Studies

The 42 included studies were published between 2017 and 2024, with a noticeable increase in publications in the last three years, reflecting the growing interest in synthetic data for healthcare. Most studies employed machine learning–based techniques, including generative adversarial networks, variational autoencoders, diffusion models, and federated learning. Applications varied across domains such as electronic health records, medical imaging, clinical simulations, and privacy-preserving analytics. Several studies focused on bias mitigation and fairness, while others explored ethical concerns, regulatory implications, or educational uses. As this scoping review aimed to map the range and scope of existing research, no critical appraisal of individual sources of evidence was performed. A summary of key characteristics and citations is presented in Table 1.

Table 1. Studies Included in the Systematic Review.

No formal critical appraisal of individual sources of evidence was conducted, as the aim of this scoping review was to map the breadth and nature of existing research rather than to assess the methodological quality or risk of bias. This approach is consistent with the objectives and methodological guidance for scoping reviews, which prioritize comprehensive coverage over detailed quality evaluation. **Study**	**Methodology**	**Application Focus**	**Challenges Identified**	**Implications**
D'Amico et al. (2023)[13]	AI-based synthetic data generation	Precision medicine in hematology	Validation of synthetic data and integration with real datasets	Accelerates research and improves personalized treatment strategies
Akpinar et al. (2024)[14]	Systematic review of GAN-based techniques	Healthcare image and signal data generation	GAN stability and performance evaluation	Provides insights into potential applications and gaps in healthcare data synthesis
Aravinth et al. (2023)[15]	Comparative analysis of generative AI techniques	Tabular medical record data generation	Scalability and preservation of data utility	Facilitates better understanding of data generation approaches for EHRs
Ferreira et al. (2024)[16]	GAN-based systematic review	3D volumetric data generation	Computational complexity and realism of generated data	Advances 3D data synthesis for medical imaging and diagnostics
Rashidian et al. (2020)[17]	SMOOTH-GAN architecture	Synthetic EHR data generation	Maintaining longitudinal consistency in synthetic data	Improves quality and usability of synthetic EHR datasets for research
Nikolentzos et al. (2023)[18]	Variational graph autoencoders	Synthetic electronic health records	Complexity in representing relational structures	Enhances data synthesis with relational and temporal context
Dos Santos et al. (2024)[19]	VAE and linked data paradigm	Synthetic data generation for medical research	Integration with linked datasets and scalability	Promotes interoperability and broader applications in health research
Lenatti et al. (2023)[20]	Rule-based AI models	Characterization of synthetic health data	Incorporating domain-specific rules effectively	Improves reliability and acceptance of synthetic datasets
Arora & Arora (2022)[21]	Generative adversarial	Synthetic patient	Ethical concerns and	Guides to the ethical

	networks (GANs)	data generation	biases in GAN outputs	development of AI in healthcare
Little et al. (2023)[7]	Federated learning	Synthetic data generation for privacy-preservation	Balancing privacy and data utility	Enhance secure data sharing and collaborative research
Mosquera et al. (2023)[22]	Methodology for longitudinal data synthesis	Synthetic longitudinal health data	Maintaining temporal trends and relationships	Enables research requiring time-series health data
Sun et al. (2021)[23]	Recurrent autoencoders and GANs	Longitudinal synthetic EHR data generation	Complexity of modeling temporal dependencies	Advances the synthesis of realistic time-series health records
Kosolwattana et al. (2023)[24]	Self-inspected adaptive SMOTE (SASMOTE)	Imbalanced healthcare data classification	Over-sampling without overfitting minority classes	Improves model performance on rare medical conditions
Nicolaie et al. (2023)[25]	Synthetic population construction	Big data applications in public health	Balancing population diversity and representativeness	Supports public health simulations and policy planning
Kumichev et al. (2024)[26]	LLM-based synthetic text generation	Medical text generation for research	Preservation of medical context and coherence	Facilitates NLP research and healthcare applications
Miletic & Sariyar (2024)[27]	Benchmark study	Tabular health data generation	Performance and accuracy trade-offs	Guides to the selection of appropriate synthetic data models
Juwara et al. (2024)[28]	Synthetic data augmentation	Mitigation of covariate bias in health data	Overcoming model biases and variability	Improve equity in health data analyses
Lomotey et al. (2024)[29]	Digital twins and data trusts	Privacy in health data sharing	Balancing privacy with data usability	Facilitates secure and ethical health data sharing
Osorio-Marulanda et al. (2024)[30]	Systematic review	Privacy and evaluation metrics for synthetic data	Standardization of metrics and methods	Improves trust in synthetic data for sensitive domains
Nicholas et al. (2024)[31]	Health Gym project	Synthetic datasets in education	Engaging learners without overwhelming complexity	Enriches data science and healthcare education
Patil et al. (2024)[32]	Transformer-based DGA integration	Improved ML-based fault identification	Complexity in data integration and scalability	Enhance fault detection in healthcare systems using synthetic data
Gonzales et al. (2023)[10]	Narrative review	Synthetic data in healthcare applications	Lack of standardization and ethical considerations	Encourages unified guidelines for healthcare data synthesis
GiuffrÃ & Shung (2023)[1]	Review on synthetic data innovation	Healthcare privacy and innovation	Balancing innovation with ethical responsibilities	Guides responsible for the use of synthetic data in health technologies

Qian et al. (2024)[8]	Privacy-preserving clinical risk prediction	Synthetic data for clinical applications	Data fidelity and privacy trade-offs	Facilitates secure predictive modeling in clinical research
Burgon et al. (2024)[33]	Bias amplification evaluation framework	Bias mitigation in healthcare ML models	Challenges in systematic bias evaluation	Improves fairness and accountability in AI healthcare tools
Koetzier et al. (2024)[34]	Medical imaging synthetic data generation	Enhancing medical imaging datasets	Quality and utility of synthetic images	Advances imaging tools for better diagnostic accuracy
Rodriguez-Almeida et al. (2022)[35]	Disease prediction on small datasets	Synthetic patient data for imbalanced datasets	Balancing small sample sizes with realistic data generation	Improves disease prediction accuracy in rare conditions
Shanley et al. (2024)[9]	Ethics-focused review	Synthetic data ethics in healthcare	Adoption of AI ethics principles	Strengthens ethical frameworks for synthetic data usage
Chen et al. (2021)[36]	ML applications in synthetic data	Medicine and healthcare	Reproducibility and validation of synthetic data models	Encourages robustness in AI model development for healthcare
Goyal & Mahmoud (2024)[37]	Systematic review of generative AI	Synthetic data generation techniques	Scalability and generalizability	Broadens understanding of generative methods
Tucker et al. (2020)[38]	High-fidelity synthetic patient data	Machine learning in healthcare software testing	Achieving realism in synthetic datasets	Enhances model validation and reliability in healthcare AI
Hairani et al. (2024)[39]	Review of modified SMOTE strategies	Addressing class imbalance in health data	Adapting SMOTE for healthcare-specific needs	Improves handling of imbalanced datasets
Bigi et al. (2024)[40]	Agent-based modeling	Synthetic population for mobility analysis	Accurate parameterization and assumptions	Supports public health and operational planning
Guo & Zhao (2023)[41]	Survey of deep generative models for graph generation	Graph learning and representation	Scalability, permutation invariance, evaluation standards	Guides future research on graph-based data generation
Iannucci et al. (2017)[42]	Benchmarking of graph-based synthetic data generators	Intrusion detection system benchmarking	Model robustness and generalizability across attacks	Informs IDS design with realistic benchmark datasets
Haleem et al. (2023)[43]	Deep learning for multimodal health data synthesis	Real-time multimodal health data generation	Cross-modal consistency and real-time generation	Enables richer datasets for health monitoring systems
PawÅ,owski et al. (2023)[44]	Comparative analysis of multimodal data fusion methods	Sensor fusion and integration strategies	Selecting appropriate fusion strategy for task needs	Supports development of task-specific fusion pipelines
Gogoshin et al. (2021)[45]	Bayesian networks for probabilistic data generation	Biological simulation and data reconstruction	High-dimensional sampling and structural accuracy	Validates BNs as interpretable simulation frameworks

Kaur et al. (2020)[46]	Bayesian network application to synthetic health data	Synthetic health data generation and evaluation	Preserving associations and rare events	Shows BNs outperform deep models in certain tasks
Hosseini & Serag (2025)[47]	Diffusion models for synthetic medical image generation	Synthetic chest X-ray generation and model pretraining	Maintaining clinical fidelity and training stability	Demonstrates strong performance without real data
Naseer et al. (2023)[48]	Continuous-time diffusion model using stochastic differential equations	Electronic health record synthesis	Modeling long-term temporal dependencies and clinical coherence	Advances EHR generation with high realism and improved utility for downstream tasks

How Do Synthetic Data Generation in Healthcare Works?

Synthetic data generation in healthcare involves creating artificial datasets that replicate the characteristics of real-world medical data using computational techniques. This process enables researchers and developers to work with realistic data while safeguarding patient privacy. The generation process typically follows several key steps (Figure 2):

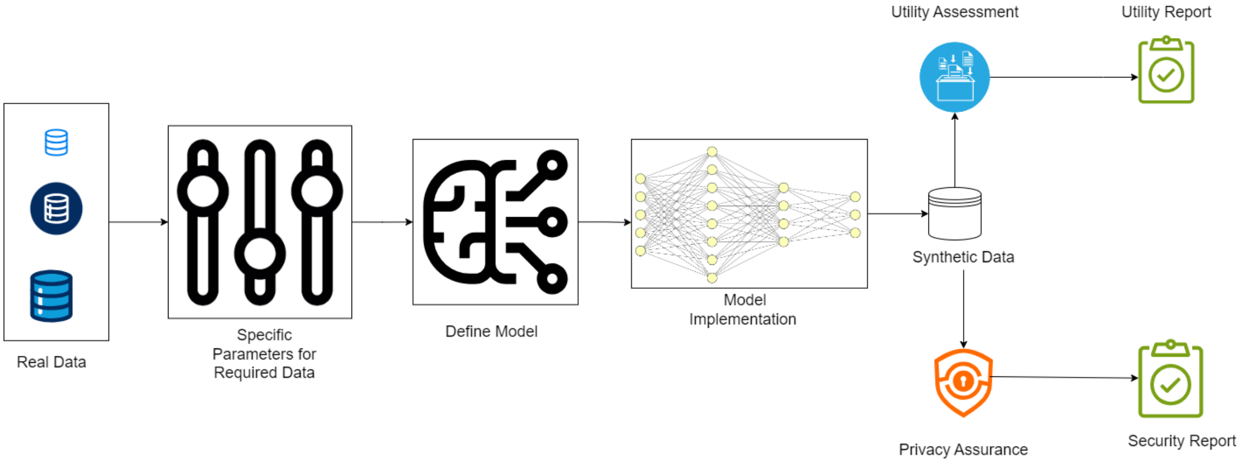

Figure 2. Synthetic Data Generation Process.

1. Data Collection and Preprocessing

The process begins with the collection of real-world healthcare data, such as electronic health records (EHRs), medical images, or genomic sequences. This data is cleaned and preprocessed to eliminate errors, inconsistencies, or irrelevant details. To comply with privacy regulations like HIPAA and GDPR, sensitive patient identifiers are often removed during this stage.[1] Preprocessing ensures the data is in a suitable format for training models while reducing risks to patient confidentiality.

2. Model Training

Next, algorithms are trained on the preprocessed real data to identify and replicate their patterns, relationships, and statistical properties. The complexity of the models varies depending on the application. Simple approaches, such as rule-based systems, rely on predefined rules or statistical distributions.[20] More advanced techniques, like machine learning algorithms, analyze deeper patterns within the data.[36] Two widely used methods are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). GANs employ a dual-model system, a generator creates synthetic data, and a discriminator evaluates its realism, iteratively refining the output until it closely resembles real data.[17] VAEs, conversely, compress data into a latent space and then reconstruct it, producing synthetic versions that maintain statistical integrity.[18] These models aim to balance realism with privacy protection.

3. Data Generation

Once trained, the model generates synthetic datasets that mirror the structure and statistical distributions of the original data without copying individual records. For example, synthetic EHRs might include fabricated patient profiles with realistic age, diagnosis, and treatment histories, while synthetic medical images could simulate X-rays or MRIs.[16] The goal is to produce data that reflects real-world trends and correlations, ensuring it is suitable for downstream applications like AI training or research.[38]

4. Validation and Evaluation

The final step involves assessing the synthetic data against specific quality metrics to ensure it meets its intended purpose. These metrics include:

- **Fidelity**: How well the synthetic data matches the statistical properties of real data.[35]
- **Utility**: Whether the data performs effectively in tasks like training machine learning models.[36]
- **Privacy**: Confirmation that no sensitive information from the original dataset can be inferred, often tested using privacy-preserving techniques like differential privacy.[30]

Validation ensures the synthetic data is both practical and compliant with ethical and legal standards. Techniques such as statistical comparisons or privacy audits are commonly employed to verify these qualities.

Current Synthetic Healthcare Data Generation Techniques

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are one of the most common approaches to generating synthetic data, particularly in areas requiring high realism, such as medical imaging and time-series analysis.[16] These networks consist of two neural architectures: a generator and a discriminator that are simultaneously trained under a competitive setup. The generator generates synthetic data, while the discriminator tries to determine its authenticity by comparing it to real data. Through this iterative process, the generator learns to generate data that the discriminator cannot classify as being either real or synthetic, thus creating realistic synthetic datasets. In medicine, GANs have been used to generate synthetic X-rays, MRIs, and CT scans for the purpose of increasing the size of datasets for the training of diagnostic AI systems.[17] Variants of GANs, such as TimeGAN, are tailored for time-series data, capturing temporal dependencies crucial for patient monitoring or disease progression analysis. Another approach, PATE-GAN, introduced differential privacy mechanisms in the generation process of synthetic datasets by adding statistical noise during training. However, GANs require a lot of computational power and high-quality data to be able to produce good results.[35] In addition, these models are subject to problems such as mode collapse, a case where the generator always generates a small number of variations of synthetic data, thereby decreasing the diversity of the generated outputs.

Variational Autoencoders (VAEs)

Another popular approach to the generation of synthetic data, especially for structured datasets such as EHRs and genomic data, is through Variational Autoencoders (VAEs).[23] VAEs compress real-world data into a lower-dimensional latent space and then decompress it to create synthetic versions. This latent space representation allows VAEs to capture the statistical properties and relationships in the original data, ensuring that the generated synthetic data is realistic and diverse. VAEs may become especially useful in healthcare to generate patient profiles that can be safely used for research without exposing sensitive information.[18] Compared to GANs, VAEs are much easier to train and robustly offer diversity in data. However, synthetic data generated by VAEs typically lack the high degree of realism displayed by GANs and are therefore less usable in applications requiring photorealistic output, such as medical imaging. VAEs also provide a probabilistic approach to data generation, which can be especially useful in tasks that require uncertainty quantification, for example, predictive modeling in healthcare.

Differential Privacy-Based Methods

Differential privacy-based methods are designed to generate synthetic data that protects individual-level information by introducing carefully calibrated statistical noise during the generation process. Rather than modifying original records, these methods create entirely new datasets that reflect the overall patterns of the real data without allowing the identification of any specific individual. This is especially important in healthcare, where strict regulations such as HIPAA and GDPR govern the use of sensitive data. One widely used approach is PrivBayes, which first learns a Bayesian network to approximate the original data distribution using low-dimensional marginals, then injects differential privacy noise into these marginals before sampling synthetic records from the network.[8] Another example is PATE-GAN, which combines generative adversarial training with the Private Aggregation of Teacher Ensembles framework to ensure that the generator learns from aggregated outputs that maintain differential privacy.[30] These methods allow researchers to share and analyze synthetic data with strong privacy guarantees. However, achieving a balance between privacy and utility is a key challenge. Adding too much noise can lower the accuracy and realism of the synthetic data, limiting its usefulness in high-stakes applications such as diagnostic model training.[8]

Graph-Based Synthetic Data Generation: GraphGAN and NetGAN

GraphGAN is an early method for generating synthetic graph data using two competing models: a generator that tries to create realistic connections between nodes, and a discriminator that tries to tell if those connections are real or fake.[49] This helps the generator learn how real graphs are structured. NetGAN improved on this by generating random walks through the graph instead of direct connections. It uses an LSTM to produce sequences of steps that look like real paths in the graph, which are then used to build a new synthetic graph.[41] Later models like MMGAN and SHADOWCAST added more control by focusing on patterns or labeled walks.[42] These approaches are useful in healthcare for creating synthetic patient networks or referral patterns when real data is limited. However, challenges remain, like making sure the graphs are realistic and scalable. Some newer systems generate very large graphs while keeping features like how connected the nodes are or how important each node is. Others combine these models with tools used in software engineering to produce clean, usable graph data even when only small real examples exist.

Multimodal Synthetic Data Generation

Multimodal synthetic data generation focuses on creating data that combines different sources, such as clinical notes, sensor signals, and wearable data while preserving the relationships between them. In healthcare, one approach called TC MultiGAN extends existing time-series generators to capture the timing and interaction of different physiological signals. Another method treats wearable and clinical events as text entries, allowing the model to handle irregular timing and complex dependencies between variables.[43] Combining multiple data types can improve both the quality and usefulness of synthetic data, as each type adds different information. Studies emphasize that the way data is fused plays a major role. Some models combine inputs at the feature stage, others make decisions separately and then merge results, while some map all inputs into a shared space. Choosing the right method depends on the task and available resources like memory and speed.[44] Overall, successful multimodal synthesis requires understanding of what each type of data contributes, keeping their timing and structure consistent, and evaluating results using the actual goals of the task rather than just visual comparisons.

Bayesian Networks for Synthetic Data Generation

Bayesian Networks (BNs) offer a powerful and transparent method for generating synthetic healthcare data by modeling the probabilistic relationships between variables. Defined as directed acyclic graphs with conditional probability tables, BNs capture dependencies and uncertainties in complex clinical settings, making them well-suited for simulating realistic patient records.[45] In practice, a BN learned from real data can be used to generate new synthetic datasets that reflect the original distribution while preserving patient privacy. When carefully constructed, BNs have been shown to match or outperform deep learning models in maintaining association patterns and handling rare clinical events.[46] These advantages, combined with interpretability and built-in mechanisms for incorporating domain knowledge, make BNs a strong choice for privacy-preserving synthetic data pipelines.

Diffusion Models for Synthetic Data Generation

Diffusion models generate synthetic data by gradually adding Gaussian noise to real samples and then learning a reverse denoising process that reconstructs high-fidelity outputs without adversarial training. In medical imaging, Denoising Diffusion Probabilistic Models have produced chest X-ray and lung segmentation samples that support downstream classifiers at or above real-data baselines, achieving AUC near 0 point 99 and Dice scores around 0 point 85 while preserving clinically useful biomarkers.[47] Their training stability and self-supervised objective make them attractive when labeled data are scarce. Moving beyond images, continuous-time diffusion frameworks such as ScoEHR couple autoencoders with stochastic differential equations to synthesize electronic health records. ScoEHR outperforms medGAN variants on joint distribution fidelity and downstream predictive utility, and clinicians in a blinded test judged its records indistinguishable from real ones.[48] Despite these successes, diffusion models demand extensive computation and careful tuning of noise schedules, and there is still no consensus on standard benchmarks for evaluating privacy leakage and task relevance. Even so, their ability to capture complex structures in both image and tabular domains positions diffusion modeling as a promising direction for privacy-preserving synthetic data in healthcare research.

Federated Learning for Synthetic Data Generation

Federated learning is increasingly explored as a privacy-preserving framework for synthetic data generation, particularly in healthcare, where sensitive patient data are distributed across multiple institutions. Instead of sharing raw data, each institution trains a local model on its own dataset. Only the resulting model parameters or gradients are shared and aggregated to produce a global model that can be used to generate synthetic data reflecting the collective knowledge of all participating sites.[7] This approach addresses critical privacy concerns governed by regulations such as HIPAA, making it possible to generate synthetic datasets without compromising control over patient records. Federated learning supports broader collaborations across hospitals and research centers by enabling synthetic data generation from otherwise siloed data.[29] However, its effectiveness is not without limitations. The success of the global model heavily depends on the quality and consistency of local data, which may vary significantly. Furthermore, federated learning requires significant computational resources and reliable communication infrastructures. These constraints can limit scalability, especially when participating institutions have unequal technological capabilities. While federated frameworks hold promises for creating representative and privacy-preserving synthetic datasets, their implementation must be carefully managed to ensure equity, consistency, and model reliability.

Recurrent Neural Networks (RNNs)

Recurrent Neural Networks, in their various forms and more developed versions such as Long Short-Term Memory, have proved highly successful for sequential or time series data generation.[23]

RNNs have been used in the health context to model monitoring data on patients, including those with heart rate, glucose level, and medication adherence over time. These models are good at finding temporal relationships and trends in sequential data, making them well-suited for applications like disease progression modeling or treatment outcome forecasting.[50] However, recurrent neural networks (RNNs) require large datasets to be trained effectively and have high computational requirements. They are also prone to issues like vanishing or exploding gradients, which can hurt their performance when dealing with long sequences.

Synthetic Minority Over-Sampling Technique (SMOTE)

The Synthetic Minority Over-Sampling Technique, or SMOTE, is a technique designed for class distribution balancing in datasets.[39] By interpolating existing instances in the dataset, SMOTE creates synthetic samples for the minority class; it generates new instances that are similar to the original examples but not identical.[24] The health industry greatly uses this method for balancing datasets in machine learning algorithms, especially those used in predicting rare diseases or complications from pharmaceuticals. SMOTE, by improving class balance, improves the performance of models and reduces bias, which in certain diagnostic models is quite significant.[51] The limitation of SMOTE to structured data may not properly be able to capture the complex interrelationship inherent in real-world data.

Agent-Based Modeling (ABM)

Agent-based modeling (ABM) is a simulation-based technique in which discrete units, named agents, interact with each other and an environment following defined rules. In healthcare, ABM is being used in generating synthetic population-level data, modeling disease spread, or simulating hospital workflows.[40] For instance, ABM can be applied to simulate how infectious diseases spread via a community, considering individual behaviors, mobility, and vaccination status. Such a technique has been rather instrumental in public health research and policy planning.[36] However, ABM is critically dependent on accurate parameterization and assumptions, which might limit its applicability in cases where the underlying assumptions do not capture the complexities of real-world dynamics.

Large Language Models for Synthetic Text Generation

Large Language Models (LLMs), such as GPT-4, have become powerful tools for generating synthetic textual data, especially in healthcare applications where structured and unstructured narratives play a key role.[26] These models are trained in vast corpora to learn linguistic patterns and clinical context, enabling them to generate realistic outputs such as synthetic clinical notes, patient narratives, and medical documentation for training and validating natural language processing models.[52] Their ability to produce contextually appropriate and grammatically coherent text makes them useful for simulating telemedicine dialogues or constructing synthetic patient histories. However, this capability comes with notable challenges. The performance of LLMs depends heavily on the quality and scale of the training data, which raises privacy concerns when medical text is involved. Even when generating synthetic data, LLMs can inadvertently reproduce sensitive patterns, or biased associations present in their training sets. Moreover, without domain-specific fine-tuning, the generated content may include clinically incorrect or misleading information. These limitations underscore the need for cautious design, careful curation of training data, and robust evaluation protocols when using LLMs to generate synthetic healthcare text. While promising, LLM-based generation must be approached with a strong emphasis on both ethical safeguards and domain validation.

Current Applications for Synthetic Data

AI Training and Model Development

Synthetic data is a critical enabler for training and developing artificial intelligence (AI) models in healthcare.[1,38] AI systems, particularly those used for diagnosis, treatment recommendations, and predictive analytics, require large, diverse, and high-quality datasets for effective training.[6,18] However, real-world healthcare data is often limited by privacy restrictions, data scarcity, and demographic imbalances.[3,10] Synthetic data addresses these limitations by providing AI developers with an abundant and diverse dataset that mimics the statistical properties of real-world data without compromising patient privacy.[13,21]

Synthetic data generation techniques, such as GANs and VAEs, are used to create realistic datasets for AI training.[14,17] These include synthetic X-rays, MRIs, and CT scans, which are instrumental in training machine learning models for medical imaging diagnostics.[16,34] Synthetic imaging datasets are particularly valuable in simulating rare medical conditions, allowing researchers to develop and test diagnostic tools for diseases that are infrequently observed in clinical settings.[23,36] The diversity and scalability of synthetic data also allow researchers to include underrepresented populations or simulate rare conditions, thereby improving the generalizability and fairness of AI systems.[25,35] By using synthetic data, researchers can perform rigorous testing and validation of AI models in a controlled environment, reducing the risk of deploying untested algorithms in real-world clinical settings.[9,31] This application is essential in fostering the development of robust, equitable, and effective AI solutions that can address complex healthcare challenges.[7,27]

Privacy-Preserving Data Sharing

Synthetic data plays a pivotal role in facilitating privacy-preserving data sharing among healthcare institutions.[1,10] Real-world healthcare data is often restricted due to the risk of exposing sensitive patient information.[3,30] Synthetic datasets, which mimic the statistical properties and patterns of real-world data without replicating individual records, provide a secure alternative.[6,36] These datasets enable collaboration between hospitals, research organizations, and private entities, fostering innovation and discovery.[17,38] For example, synthetic data allows multi-institutional studies to proceed without breaching patient confidentiality, making it particularly valuable in cross-border research initiatives where privacy laws vary significantly.[7,25] By ensuring compliance with strict regulations like GDPR and HIPAA, synthetic data enhances the scope and efficiency of collaborative healthcare research while maintaining the highest standards of privacy.[9,31]

Bias Mitigation

Synthetic data generation techniques are increasingly employed to address demographic imbalances in real-world datasets, a critical issue in healthcare research and AI model development.[3] Real-world datasets often underrepresent certain demographic groups, leading to biased AI models that perform poorly for these populations.[6] Synthetic data helps mitigate this issue by generating additional data points for underrepresented groups, ensuring that models are trained on more balanced datasets.[36] For instance, in clinical trials, synthetic data can enhance the representation of minority demographics, improving the fairness and accuracy of predictive models.[13] This application is particularly important in ensuring that healthcare AI solutions are equitable and do not perpetuate existing disparities in patient outcomes.[31]

Education and Training

Synthetic data has become a foundation for education and training in healthcare. By generating realistic datasets, educators can create simulated scenarios for medical students and healthcare professionals.[1] These synthetic datasets provide a risk-free environment for learners to practice diagnosing conditions, interpreting medical images, and performing surgical procedures.[36] For example, synthetic patient records and imaging data can be used to simulate rare medical cases, giving trainees exposure to a broader range of conditions than they would typically encounter in clinical practice.[38] This application is especially beneficial in specialties where access to real-world training data is limited, such as pediatrics or rare genetic disorders.[31] By bridging the gap between theoretical learning and practical experience, synthetic data enhances the quality of medical education and professional training.[27]

Operational Optimization

Synthetic data is also transforming healthcare operations by enabling better resource management and decision-making. Healthcare administrators use synthetic datasets to model patient flows, optimize resource allocation, and predict the outcomes of policy changes.[6] For instance, synthetic data can simulate the impact of introducing new treatment protocols or reallocating staff in emergency departments.[3] By modeling various scenarios, administrators can make data-driven decisions to improve efficiency and patient care outcomes.[38] Synthetic data is particularly useful in operational optimization because it allows healthcare systems to test and refine strategies without disrupting real-world workflows. This application has become increasingly important as healthcare systems face growing demands and limited resources, emphasizing the need for innovative solutions to enhance operational performance.[25]

Challenges and Potential Risks

No Established Data Standards

One of the key challenges in synthetic data (SD) generation is the absence of universally accepted standards. Unlike traditional datasets that follow established protocols for formatting, labeling, and structure, synthetic data lacks a cohesive framework, hindering its sharing, validation, and integration across institutions. Literature reviews highlight critical concerns in healthcare SD, including the lack of realistic data generation tools, inadequate testing, and insufficient validation to ensure clinical relevance. Although frameworks like ATEN and tools such as MDClone aim to produce data resembling real-world clinical records, they still fall short in terms of accuracy and clinical quality.[10] Moreover, there is currently no universal framework to evaluate the quality and utility of synthetic data. Existing metrics such as fidelity, utility, and privacy are inconsistently defined and applied, making it difficult to benchmark or compare SD across contexts. Establishing global standards is therefore essential to ensure synthetic data is reliable, interoperable, and suitable for use in real-world healthcare applications.[27]

Realism vs. Privacy Trade-Off

A major challenge in synthetic data (SD) generation is balancing data realism with privacy protection. While highly realistic synthetic datasets enhance analytical value, they also increase the risk of re-identification, especially when patterns or correlations mirror those in real-world data.[16] This risk is particularly critical in healthcare, where safeguarding patient confidentiality is essential. To address these concerns, techniques such as Federated Learning (FL) have emerged as promising solutions.[7] FL enables the generation of synthetic data that preserves statistical properties while minimizing exposure of original records, thereby reducing privacy and security risks. For instance, platforms like MedSyn integrate large language models (LLMs) with federated learning to generate SD in a privacy-preserving manner, enhancing data utility without compromising confidentiality.[26]

Bias Amplification

The generation of synthetic data heavily depends on the quality and composition of the original datasets. When the source data contains biases, such as unequal representation of certain demographic groups or structural disparities, these issues can be reproduced or even intensified in the synthetic output.[33] This becomes especially problematic in healthcare, where synthetic data are frequently used to train artificial intelligence models. If these models are trained on biased synthetic datasets, they may produce outcomes that are inaccurate or unfair, potentially reinforcing existing health inequities.[53] For example, if certain populations are underrepresented in clinical records, the resulting synthetic data may overlook their specific health needs, leading to models that perform poorly for those groups and contributing to unequal healthcare delivery.

Computational Complexity

The use of advanced generative models for synthetic data generation often involves significant computational demands. Techniques such as generative adversarial networks and variational autoencoders require large volumes of data and intensive processing power to train effectively.[16,23,31] These requirements can pose challenges for smaller institutions or research teams that may lack access to high-performance computing infrastructure. As a result, the benefits of synthetic data generation may remain concentrated within well-resourced organizations.[53] Furthermore, the high computational cost often translates into increased financial burden, which can hinder the scalability and broader adoption of these methods across diverse healthcare settings.

Overconfidence in Data Utility

There is a growing risk that stakeholders may overestimate the value and reliability of synthetic data, believing it to be an adequate substitute for real-world datasets in all contexts.[38] This overconfidence can lead to flawed interpretations or decisions, particularly when synthetic data lacks the nuanced complexity, variability, or rare edge cases present in actual clinical environments. Relying too heavily on such data may undermine the validity of findings, especially in high-stakes applications like diagnostics or treatment planning. When synthetic data is used without sufficient validation or understanding of its limitations, the potential for unintended harm or misinformed policy increases significantly.

Security Vulnerabilities

Although synthetic data is intended to enhance privacy by minimizing direct links to real individuals, it remains susceptible to re-identification risks.[30] Sophisticated analytical techniques can sometimes detect subtle statistical patterns within synthetic datasets that, when combined with external information, may reveal sensitive details.[29] This vulnerability is particularly concerning in healthcare, where even partial data disclosures can lead to significant privacy breaches. To address these risks, it is essential to implement strong privacy-preserving methods and to subject synthetic datasets to rigorous validation and risk assessment before use or release. Without such safeguards, the assumption of enhanced security may offer a false sense of protection.

Ethical Concerns

The process of generating synthetic data often relies on real-world datasets, raising important ethical questions about consent, ownership, and accountability. Patients and data contributors may not be fully aware of how their data is being used to create synthetic datasets.[9] Moreover, the use of synthetic data in clinical and public health applications requires strong governance frameworks to ensure transparency, accountability, and ethical integrity.[54] The absence of clearly defined ethical principles in the use of synthetic data (SD) for artificial intelligence (AI) can lead to serious ethical challenges in healthcare. These challenges hinder the ability of synthetic data to accurately reflect real-world scenarios. Key ethical concerns include responsibility, non-maleficence, privacy, transparency, justice, fairness, and equity. Addressing these issues is essential to ensure that synthetic data is used responsibly and ethically in healthcare applications.[9]

Heterogeneity Problems and Lack of Clinical Quality

While synthetic data (SD) can replicate statistical patterns and improve the accuracy of predictive models, it often lacks the heterogeneity found in real-world clinical data. This lack of diversity in outcomes and patient characteristics can limit the generalizability and usability of SD in healthcare settings.[10] Additionally, many synthetic datasets fall short in clinical quality because they are not developed following established healthcare frameworks or standards. As a result, these datasets may bypass proper validation processes, leading to unreliable or inaccurate outcomes when applied in real-world healthcare applications.[34]

The Future of Healthcare AI and the Impact of Synthetic Data

Risk of Model Overfitting to Synthetic Patterns

Synthetic data, while addressing data scarcity, risks leading AI models to overfit to its specific patterns, compromising generalization to real-world data. Overfitting occurs when models learn noise or artifacts from the synthetic generation process, such as those introduced by generative models like GANs, rather than generalizable patterns. This can result in poor performance when applied to actual patient data, undermining clinical reliability.

Research from Evaluate synthetic data quality using downstream ML introduces the Train-Synthetic-Test-Real (TSTR) method, validating synthetic data by training models on it and testing on real data, comparing performance to models trained on original data.[55] This approach ensures synthetic data captures essential statistical properties, mitigating overfitting risks. Additionally, in machine learning, synthetic data can offer real performance improvements found that with small real datasets, synthetic-trained models can outperform real-data-trained ones, but validation on real data is crucial to confirm generalizability.[56] The challenge lies in ensuring synthetic data quality, as poor fidelity can exacerbate overfitting. For instance, if synthetic data lacks the variability of real healthcare data, models may fail in clinical settings, highlighting the need for rigorous quality assessment frameworks like those proposed in recent studies.

Amplification of Bias and Inequity

Synthetic data can inherit biases from the real data used for generation, potentially amplifying inequities in healthcare AI. If the original dataset underrepresents certain demographics, such as racial or socioeconomic groups, the synthetic data may perpetuate these biases, leading to AI models that perform poorly for marginalized populations. Shahul Hameed et al., discusses methods like Generative Adversarial Networks (GANs) and Bayesian networks to reduce bias, achieving up to 92% accuracy in biomedical signals with SynSigGAN.[53] Identifying and handling data bias within primary healthcare data using synthetic data generators explores probabilistic approaches to detect and boost underrepresented data samples, improving model fairness.[57] These techniques aim to balance datasets, but their effectiveness depends on the quality of initial data, with risks of bias propagation if not addressed. This amplification can exacerbate health disparities, particularly in underserved communities, necessitating continuous monitoring and fairness metrics to ensure equitable AI outcomes.

Erosion of Trust in AI Systems

The use of synthetic data in training AI models can lead to skepticism among healthcare professionals and patients regarding the reliability and trustworthiness of these models. The perceived artificiality of the data might make it difficult for clinicians to have confidence in the decisions made by such models.[9] Building trust requires transparency about the use of synthetic data, along with clear demonstrations of the model's performance on real data. Ethical guidelines and regulatory oversight can also play a significant role in ensuring that synthetic data is used responsibly and that the models are validated appropriately.[58] For instance, validation against real data using

TSTR evaluations can demonstrate comparable performance to real-data-trained models, potentially alleviating trust concerns, but initial skepticism remains, particularly in high-stakes clinical settings.

Challenges in Validation and Benchmarking

Validating AI models trained on synthetic data presents unique challenges. Traditional validation methods rely on real-world data, but assessing how well a model trained on synthetic data generalizes to real data requires careful design.[30] The lack of standardized benchmarks for synthetic data quality and utility complicates the evaluation process. Innovative approaches, such as the TSTR method, are being developed to address these challenges. Additionally, creating hybrid datasets that combine synthetic and real data can provide a more robust testing ground for AI models.[59] However, the development of universally accepted validation metrics remains a critical area of research, with future work focusing on federated benchmarking platforms like MedPerf to ensure synthetic-trained models are rigorously assessed for clinical applicability.

Ethical Concerns About Data Ownership and Consent

Generating synthetic data from real patient data raises ethical questions regarding data ownership and consent. Patients may not have explicitly consented to their data being used to create synthetic datasets, which could be used in various applications, some of which they might not approve of.[58] To navigate these ethical dilemmas, it's important to establish clear policies and obtain informed consent from patients regarding the use of their data for synthetic generation. Moreover, ensuring that synthetic data does not contain identifiable information and adheres to privacy regulations like GDPR is paramount, with ongoing debates shaping regulatory frameworks to protect patient rights.

Security Risks and Privacy Paradox

While synthetic data is designed to protect privacy, there is a risk of re-identification, especially if the synthetic data retains certain characteristics that can be linked back to individuals.[60] This privacy paradox underscores the need for robust privacy-preserving techniques during the generation process. Techniques such as differential privacy can be employed to add noise to the data, making it difficult to trace back to original records.[8] Regular auditing and assessment of re-identification risks are also necessary to maintain the privacy benefits of synthetic data, with frameworks like the Identifiability Score measuring overlap with real data to assess privacy risks.

Overreliance and Misplaced Confidence

There is a tendency to over-rely on synthetic data, if models trained on it will perform as well as those trained on real data. This misplaced confidence can lead to the deployment of AI systems that are inadequately tested or validated, potentially resulting in clinical errors.[61] To prevent this, it's essential to use synthetic data as a supplement rather than a replacement for real data. Continuous validation and updating of models with real-world data are necessary to ensure their accuracy and reliability over time.[37] Educating stakeholders about these risks is crucial, ensuring synthetic data is used as a complement, not a replacement, to maintain model reliability.

Discussion

This scoping review aimed to examine how synthetic data is being generated, validated, and applied in healthcare AI contexts. The included studies revealed a growing and diverse body of research focused on domains such as electronic health records, medical imaging, and clinical simulations. The findings confirm that synthetic data generation in healthcare is no longer confined to a narrow set of methods but spans a broad and rapidly evolving range of techniques.[61] Most studies used deep learning–based approaches such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Recurrent Neural Networks (RNNs), diffusion models, and large language models (LLMs); probabilistic and rule-based techniques such as Bayesian networks and agent-based models; and privacy-enhanced models incorporating federated learning and differential privacy. Each of these methods offers distinct advantages depending on the data type and intended application. For instance, GANs and diffusion models are well-suited for medical image synthesis, while Bayesian networks and RNNs are more commonly applied to structured data such as electronic health records or longitudinal monitoring.[62] Multimodal and graph-based techniques further expand the potential of synthetic data by capturing complex interactions across data types and entities.[42] Despite this progress, evaluation practices remain inconsistent, with varying definitions and benchmarks for utility, privacy, and bias. Ethical concerns around consent, ownership, and fairness are frequently acknowledged but often insufficiently addressed. These findings underscore the need for methodological standardization, ethical guidelines, and multidisciplinary collaboration to responsibly advance the field.

These techniques have far-reaching implications for both research and clinical practice. By enabling access to realistic yet non-identifiable data, they allow researchers to train and evaluate machine learning models without compromising patient privacy.[60] This is particularly important in rare disease research or across underrepresented populations where real data is often scarce.[57] Diffusion models show strong promise in producing high-fidelity clinical data, while federated learning offers a framework for cross-institutional collaboration without raw data exchange.[7,47] At the same time, multimodal and graph-based synthesis methods are opening new opportunities for system-level modeling in healthcare.[44] However, while the technical capabilities of these models are growing, challenges remain in ensuring that they produce data that are not only realistic, but also useful, fair, and interpretable in clinical contexts.

Looking forward, several areas require concentrated research attention. First, there is a need for standard benchmarks and validation protocols that go beyond visual inspection or distributional similarity.[59,63] Evaluation frameworks should consider how well synthetic data supports downstream tasks such as diagnosis, prediction, or treatment planning. Second, more work is needed on fairness and representation. Many synthetic data pipelines currently replicate or even worsen the biases present in the original datasets.[64] Techniques that incorporate fairness-aware training objectives, class rebalancing, or demographic constraints should be further developed and adopted.[65] Third, the field must address gaps in domain-specific synthesis. While synthetic imaging and structured health records are well studied, areas such as behavioral health, speech data, and real-world clinical decision-making remain relatively underserved.[66]

In addition to technical challenges, the growth of synthetic data generation raises important questions around policy and governance. Current privacy regulations such as HIPAA and GDPR do not fully anticipate the complexities introduced by synthetic datasets.[30,67] Clearer legal definitions are needed to determine when synthetic data is truly de-identified and what responsibilities developers and institutions hold when sharing or using such data. Additionally, institutional review boards and data-sharing agreements should begin to incorporate synthetic data provisions, including transparency in model design and documentation of data generation practices.[8] In practice, clinical adoption of synthetic data will also require greater education and awareness among healthcare providers, who may be asked to rely on models trained in part or in full synthetic sources.

This review shows that synthetic data is poised to play a central role in shaping the future of healthcare research and artificial intelligence. The diversity of available methods provides flexibility, but also demands greater coordination in terms of evaluation, governance, and equitable design. Addressing these issues will require interdisciplinary collaboration across computer science, biomedicine, ethics, and policy. If approached thoughtfully, synthetic data can not only solve immediate privacy and access barriers but also contribute to more transparent, inclusive, and reproducible health systems.

Limitations

This review has several limitations. First, the search was limited to three academic databases (PubMed, IEEE Xplore, and ACM Digital Library), which may have excluded relevant studies published in other sources, including grey literature or preprint repositories. Second, only English-language studies were included, potentially introducing language bias. Third, no formal critical appraisal of individual sources was conducted, as the purpose of this scoping review was to map existing research rather than assess study quality. Finally, while data extraction was verified by a second reviewer, the initial charting process was completed by a single reviewer, which may introduce a risk of bias or oversight.

Conclusion

This scoping review highlights the rapid expansion and methodological diversity of synthetic data generation in healthcare artificial intelligence. The findings emphasize both the opportunities and challenges in the field, including advancements in generative modeling, gaps in evaluation practices, ethical concerns, and the need for stronger governance. To ensure safe and equitable implementation, future research should focus on developing standardized evaluation protocols, improving transparency in model development, and providing clearer regulatory guidance. Collaboration across technical, clinical, ethical, and policy domains will be essential to support the responsible and effective integration of synthetic data into healthcare systems.

Data Availability: No new data were created or analyzed in this study. Data sharing is not applicable to this article as it is based solely on a review of existing literature. The sources of data for the reviewed studies are provided in the references section.

Funding: This scoping review did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors. The authors conducted this review independently.

Use of AI: During the preparation of this work the authors used ChatGPT to improve the quality of writing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

1. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *Npj Digit Med*. 2023;6(1):1-8. doi:10.1038/s41746-023-00927-3
2. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *Npj Digit Med*. 2020;3(1):1-13. doi:10.1038/s41746-020-00353-9
3. Rujas M, Herranz RMG del M, Fico G, Merino-Barbancho B. Synthetic Data Generation in Healthcare: A Scoping Review of reviews on domains, motivations, and future applications. Published online August 9, 2024:2024.08.09.24311338. doi:10.1101/2024.08.09.24311338
4. Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning. Gartner. Accessed January 2, 2025. https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning
5. Fortune Business. Synthetic Data Generation Market | Forecast Analysis [2030]. Accessed January 2, 2025. https://www.fortunebusinessinsights.com/synthetic-data-generation-market-108433
6. Madden B. Synthetic Data in Healthcare: the Great Data Unlock. Hospitalogy. November 2, 2023. Accessed January 2, 2025. https://hospitalogy.com/articles/2023-11-02/synthetic-data-in-healthcare-great-data-unlock/
7. Little C, Elliot M, Allmendinger R. Federated learning for generating synthetic data: a scoping review. *Int J Popul Data Sci*. 2023;8(1):2158. doi:10.23889/ijpds.v8i1.2158

8. Qian Z, Callender T, Cebere B, Janes SM, Navani N, van der Schaar M. Synthetic data for privacy-preserving clinical risk prediction. *Sci Rep*. 2024;14(1):25676. doi:10.1038/s41598-024-72894-y
9. Shanley D, Hogenboom J, Lysen F, et al. Getting real about synthetic data ethics. *EMBO Rep*. 2024;25(5):2152-2155. doi:10.1038/s44319-024-00101-0
10. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. *PLOS Digit Health*. 2023;2(1):e0000082. doi:10.1371/journal.pdig.0000082
11. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*. 2018;18(1):143. doi:10.1186/s12874-018-0611-x
12. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169(7):467-473. doi:10.7326/M18-0850
13. D'Amico S, Dall'Olio D, Sala C, et al. Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. *JCO Clin Cancer Inform*. 2023;7:e2300021. doi:10.1200/CCI.23.00021
14. Akpinar MH, Sengur A, Salvi M, et al. Synthetic Data Generation via Generative Adversarial Networks in Healthcare: A Systematic Review of Image- and Signal-Based Studies. *IEEE Open J Eng Med Biol*. 2025;6:183-192. doi:10.1109/OJEMB.2024.3508472
15. Aravinth SS, Srithar S, Joseph KP, Gopala Anil Varma U, Kiran GM, Jonna V. Comparative Analysis of Generative AI Techniques for Addressing the Tabular Data Generation Problem in Medical Records. In: *2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*. ; 2023:1-5. doi:10.1109/ICRASET59632.2023.10419886
16. Ferreira A, Li J, Pomykala KL, Kleesiek J, Alves V, Egger J. GAN-based generation of realistic 3D volumetric data: A systematic review and taxonomy. *Med Image Anal*. 2024;93:103100. doi:10.1016/j.media.2024.103100
17. Rashidian S, Wang F, Moffitt R, et al. SMOOTH-GAN: Towards Sharp and Smooth Synthetic EHR Data Generation. In: *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings*. Springer-Verlag; 2020:37-48. doi:10.1007/978-3-030-59137-3_4
18. Nikolentzos G, Vazirgiannis M, Xypolopoulos C, Lingman M, Brandt EG. Synthetic electronic health records generated with variational graph autoencoders. *Npj Digit Med*. 2023;6(1):1-12. doi:10.1038/s41746-023-00822-x
19. Dos Santos R, Aguilar J. A synthetic data generation system based on the variational-autoencoder technique and the linked data paradigm | Progress in Artificial Intelligence. Accessed January 2, 2025. https://link.springer.com/article/10.1007/s13748-024-00328-x
20. Lenatti M, Paglialonga A, Orani V, Ferretti M, Mongelli M. Characterization of Synthetic Health Data Using Rule-Based Artificial Intelligence Models. *IEEE J Biomed Health Inform*. 2023;27(8):3760-3769. doi:10.1109/JBHI.2023.3236722
21. Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthc J*. 2022;9(2):190-193. doi:10.7861/fhj.2022-0013
22. Mosquera L, El Emam K, Ding L, et al. A method for generating synthetic longitudinal health data. *BMC Med Res Methodol*. 2023;23(1):67. doi:10.1186/s12874-023-01869-w
23. Sun S, Wang F, Rashidian S, et al. Generating Longitudinal Synthetic EHR Data with Recurrent Autoencoders and Generative Adversarial Networks. In: *Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB Workshops, Poly 2021 and DMAH 2021, Virtual Event, August 20, 2021, Revised Selected Papers*. Springer-Verlag; 2021:153-165. doi:10.1007/978-3-030-93663-1_12
24. Kosolwattana T, Liu C, Hu R, Han S, Chen H, Lin Y. A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *BioData Min*. 2023;16(1):15. doi:10.1186/s13040-023-00330-4
25. Nicolaie MA, Füssenich K, Ameling C, Boshuizen HC. Constructing synthetic populations in the age of big data. *Popul Health Metr*. 2023;21(1):19. doi:10.1186/s12963-023-00319-5
26. Kumichev G, Blinov P, Kuzkina Y, et al. MedSyn: LLM-Based Synthetic Medical Text Generation Framework. In: Bifet A, Krilavičius T, Miliou I, Nowaczyk S, eds. *Machine Learning and Knowledge Discovery*

in Databases. Applied Data Science Track. Springer Nature Switzerland; 2024:215-230. doi:10.1007/978-3-031-70381-2_14

27. Miletic M, Sariyar M. Large Language Models for Synthetic Tabular Health Data: A Benchmark Study. In: *Digital Health and Informatics Innovations for Sustainable Health Care Systems*. IOS Press; 2024:963-967. doi:10.3233/SHTI240571
28. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns*. 2024;5(4):100946. doi:10.1016/j.patter.2024.100946
29. Lomotey RK, Kumi S, Ray M, Deters R. Synthetic Data Digital Twins and Data Trusts Control for Privacy in Health Data Sharing. In: *Proceedings of the 2024 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems*. SaT-CPS '24. Association for Computing Machinery; 2024:1-10. doi:10.1145/3643650.3658605
30. Osorio-Marulanda PA, Epelde G, Hernandez M, Isasa I, Reyes NM, Iraola AB. Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review. *IEEE Access*. 2024;12:88048-88074. doi:10.1109/ACCESS.2024.3417608
31. Nicholas KIH, Perez-Concha O, Hanly M, et al. Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project. *JMIR Med Educ*. 2024;10(1):e51388. doi:10.2196/51388
32. Patil AJ, Naresh R, Jarial RK, Malik H. Optimized Synthetic Data Integration with Transformer's DGA Data for Improved ML-based Fault Identification. *IEEE Trans Dielectr Electr Insul*. Published online 2024:1-1. doi:10.1109/TDEI.2024.3421915
33. Burgon A, Zhang Y, Petrick N, Sahiner B, Cha KH, Samala RK. Bias amplification to facilitate the systematic evaluation of bias mitigation methods. *IEEE J Biomed Health Inform*. Published online 2024:1-12. doi:10.1109/JBHI.2024.3491946
34. Koetzier LR, Wu J, Mastrodicasa D, et al. Generating Synthetic Data for Medical Imaging. *Radiology*. Published online September 10, 2024. doi:10.1148/radiol.232471
35. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*. 2022;493:28-45. doi:10.1016/j.neucom.2022.04.053
36. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng*. 2021;5(6):493-497. doi:10.1038/s41551-021-00751-8
37. Goyal M, Mahmoud QH. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*. 2024;13(17):3509. doi:10.3390/electronics13173509
38. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digit Med*. 2020;3:147. doi:10.1038/s41746-020-00353-9
39. Hairani H, Widiyaningtyas T, Prasetya DD. Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies. *JOIV Int J Inform Vis*. 2024;8(3):1310-1318. doi:10.62527/joiv.8.3.2283
40. Bigi F, Rashidi TH, Viti F. Synthetic Population: A Reliable Framework for Analysis for Agent-Based Modeling in Mobility. *Transp Res Rec*. Published online April 15, 2024:03611981241239656. doi:10.1177/03611981241239656
41. Guo X, Zhao L. A Systematic Survey on Deep Generative Models for Graph Generation. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(5):5370-5390. doi:10.1109/TPAMI.2022.3214832
42. Iannucci S, Kholidy HA, Ghimire AD, Jia R, Abdelwahed S, Banicescu I. A Comparison of Graph-Based Synthetic Data Generators for Benchmarking Next-Generation Intrusion Detection Systems. In: *2017 IEEE International Conference on Cluster Computing (CLUSTER)*. ; 2017:278-289. doi:10.1109/CLUSTER.2017.54
43. Haleem MS, Ekuban A, Antonini A, Pagliara S, Pecchia L, Allocca C. Deep-Learning-Driven Techniques for Real-Time Multimodal Health and Physical Data Synthesis. *Electronics*. 2023;12(9):1989. doi:10.3390/electronics12091989
44. Pawłowski M, Wróblewska A, Sysko-Romańczuk S. Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. *Sensors*. 2023;23(5):2381. doi:10.3390/s23052381
45. Gogoshin G, Branciamore S, Rodin AS. Synthetic data generation with probabilistic Bayesian Networks. *Math Biosci Eng MBE*. 2021;18(6):8603-8621. doi:10.3934/mbe.2021426

46. Kaur D, Sobiesk M, Patil S, et al. Application of Bayesian networks to generate synthetic health data. *J Am Med Inform Assoc JAMIA*. 2020;28(4):801-811. doi:10.1093/jamia/ocaa303
47. Hosseini A, Serag A. Self-Supervised Learning Powered by Synthetic Data From Diffusion Models: Application to X-Ray Images. *IEEE Access*. 2025;13:59074-59084. doi:10.1109/ACCESS.2025.3555619
48. Naseer AA, Walker B, Landon C, et al. ScoEHR: Generating Synthetic Electronic Health Records using Continuous-time Diffusion Models. In: *Proceedings of the 8th Machine Learning for Healthcare Conference*. PMLR; 2023:489-508. Accessed July 29, 2025. https://proceedings.mlr.press/v219/naseer23a.html
49. Wang H, Wang J, Wang J, et al. GraphGAN: Graph Representation Learning With Generative Adversarial Nets. *Proc AAAI Conf Artif Intell*. 2018;32(1). doi:10.1609/aaai.v32i1.11872
50. Generating Longitudinal Synthetic EHR Data with Recurrent Autoencoders and Generative Adversarial Networks | Heterogeneous Data Management, Polystores, and Analytics for Healthcare. Accessed December 28, 2024. https://dl.acm.org/doi/10.1007/978-3-030-93663-1_12
51. Hussain L, Lone KJ, Awan IA, Abbasi AA, Pirzada J ur R. Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. *Waves Random Complex Media*. 2022;32(3):1079-1102. doi:10.1080/17455030.2020.1810364
52. Emdad FB, Ravuri B, Ayinde L, Rahman MI. "ChatGPT, a Friend or Foe for Education?" Analyzing the User's Perspectives on The Latest AI Chatbot Via Reddit. In: *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. Vol 2. ; 2024:1-5. doi:10.1109/IATMSI60426.2024.10502836
53. Shahul Hameed MA, Qureshi AM, Kaushik A. Bias Mitigation via Synthetic Data Generation: A Review. *Electronics*. 2024;13(19):3909. doi:10.3390/electronics13193909
54. Olson LK. *Ethically Challenged: Private Equity Storms US Health Care*. JHU Press; 2022.
55. Aysha A. Evaluate synthetic data quality using downstream ML - MOSTLY AI. September 20, 2023. Accessed February 27, 2025. https://mostly.ai/blog/synthetic-data-quality-evaluation
56. Zewe A. In machine learning, synthetic data can offer real performance improvements. MIT News | Massachusetts Institute of Technology. November 3, 2022. Accessed February 27, 2025. https://news.mit.edu/2022/synthetic-data-ai-improvements-1103
57. Draghi B, Wang Z, Myles P, Tucker A. Identifying and handling data bias within primary healthcare data using synthetic data generators. *Heliyon*. 2024;10(2):e24164. doi:10.1016/j.heliyon.2024.e24164
58. Holmes M, Kaufman BG, Pink GH. *Average Beneficiary CMS Hierarchical Condition Category (HCC) Risk Scores for Rural and Urban Providers*. Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill; 2018. https://tinyurl.com/3crc98ey
59. Pachouly J, Ahirrao S, Kotecha K, Selvachandran G, Abraham A. A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools. *Eng Appl Artif Intell*. 2022;111:104773. doi:10.1016/j.engappai.2022.104773
60. Ali M. Synthetic data is the future of Artificial Intelligence. Medium. January 18, 2023. Accessed February 27, 2025. https://moez-62905.medium.com/synthetic-data-is-the-future-of-artificial-intelligence-6fcfd2ce1a14
61. Shanley D, Hogenboom J, Lysen F, et al. Getting real about synthetic data ethics. *EMBO Rep*. 2024;25(5):2152-2155. doi:10.1038/s44319-024-00101-0
62. Kocoń J, Cichecki I, Kaszyca O, et al. ChatGPT: Jack of all trades, master of none. *Inf Fusion*. 2023;99:101861. doi:10.1016/j.inffus.2023.101861
63. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. *J Biomed Inform*. 2018;83:112-134. doi:10.1016/j.jbi.2018.04.007
64. González-Sendino R, Serrano E, Bajo J. Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Gener Comput Syst*. 2024;155:384-401. doi:10.1016/j.future.2024.02.023
65. Raji ID, Smart A, White RN, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Association for Computing Machinery; 2020:33-44. doi:10.1145/3351095.3372873

66. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? Published online December 28, 2017. doi:10.48550/arXiv.1712.09923
67. GDPR. What is GDPR, the EU's new data protection law? GDPR.eu. November 7, 2018. Accessed July 23, 2025. https://gdpr.eu/what-is-gdpr/

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.