International Journal of Recent Advances in Science and Technology, 2026; 12(3):18-23

Synthetic Data Generation Using Generative Artificial Intelligence

Adedokun Taofeek

Ladoke Akintola university of technology

taadedokun14@student.lautech.edu.ng

ABSTRACT

The increasing demand for large-scale, high-quality datasets for machine learning has been constrained by privacy regulations, data scarcity, and high annotation costs. Generative Artificial Intelligence (GenAI) has emerged as a transformative solution through synthetic data generation—artificially created data that mimics real-world statistical properties while safeguarding sensitive information. This study investigates the effectiveness of various synthetic data generation methods across three critical dimensions: fidelity, utility, and privacy. Using a mixed-methods design combining quantitative evaluation with application-specific assessment, we examined three primary generation approaches—statistical-based methods (synthpop), deep learning methods (Generative Adversarial Networks, Variational Autoencoders), and Large Language Model-based generation—across multiple datasets and downstream tasks. Quantitative results demonstrate that classical methods such as synthpop and copula consistently achieve strong performance across both high-dimensional and low-dimensional datasets, combining low re-identification risk (ε-identifiability of 0.25-0.35) with competitive utility. Deep learning models achieve higher distributional fidelity but incur elevated privacy risks (ε-identifiability > 0.4). Notably, general utility metrics showed only weak correlation with specific task performance, emphasizing the need for multimetric evaluation. The findings reveal that data normalization techniques critically impact generation quality, with Max-Abs normalization consistently yielding more accurate and stable synthetic profiles across all models. These findings contribute to a nuanced understanding of synthetic data generation and provide practical guidelines for selecting appropriate methods based on application requirements, data characteristics, and privacy constraints.

Keywords: Synthetic Data Generation, Generative AI, Privacy Preservation, Data Augmentation, Model Fidelity

I. INTRODUCTION

1.1 Background and Context

The rapid advancement of machine learning and artificial intelligence has created an unprecedented demand for large-scale, high-quality datasets. However, acquiring sufficient training data remains a persistent bottleneck across numerous domains. Data scarcity, high annotation costs, and privacy constraints often limit the availability of large supervised corpora, hindering the development of robust AI systems. These challenges have spurred growing interest in synthetic data generation, where additional training examples are produced artificially rather than collected from the real world.

Synthetic data is defined as data that is artificially generated rather than directly collected from real-world events or annotations. Through algorithms and generative models such as neural networks, synthetic data is created to mimic the characteristics of real data. Unlike anonymized data, which merely removes identifiers from real datasets, synthetic data is entirely artificial, ensuring the complete exclusion of sensitive information. This unique ability to safeguard privacy while preserving the utility of data renders synthetic data invaluable in fields like healthcare, finance, and education.

Recent advances in Generative AI—particularly Large Language Models like Anthropic's Claude, DeepSeek's R1, Meta's Llama, and OpenAI's GPT models—have provided powerful new tools to generate synthetic text and code that mimic real data distributions. These tools enable researchers to generate large, diverse datasets that simulate real-world scenarios without compromising privacy. The potential benefits include enabling robust statistical analyses, developing predictive models, promoting replication studies essential to advancing knowledge, and addressing the small-sample problem by simulating profiles from diverse backgrounds.

1.2 Hypotheses

Based on the theoretical framework and existing literature, this study tests the following hypotheses:

H1: Statistical-based synthetic data generation methods (synthpop, copula) will demonstrate superior privacy preservation compared to deep learning methods (GANs, VAEs), with lower re-identification risk while maintaining competitive data utility.

H2: Deep learning-based methods will achieve higher distributional fidelity (broad utility) than statistical methods, particularly for high-dimensional and complex data structures.
H3: General utility metrics will not reliably predict specific task performance, requiring multimetric evaluation frameworks for comprehensive quality assessment.
H4: Data preprocessing techniques, particularly normalization methods, will significantly impact the quality of synthetic data generated by deep generative models, with some methods consistently outperforming others.
H5: The number of variables in the training dataset will not negatively impact the fidelity, utility, or privacy of synthetic data, supporting the synthesis of comprehensive datasets rather than task-relevant subsets.

1.3 Significance of the Study
This study makes several important contributions to the field of synthetic data generation using Generative AI. First, it provides a comprehensive empirical evaluation of three primary generation approaches—statistical-based methods (synthpop, copula), deep learning methods (GANs, VAEs), and LLM-based generation—across multiple datasets and downstream tasks. This comparative analysis addresses the gap between the proliferation of synthetic data generation techniques and the limited understanding of their actual effectiveness across different contexts.
Second, by incorporating a multimetric evaluation framework that assesses fidelity, utility, and privacy simultaneously, the study contributes to the development of standardized quality assessment for synthetic data. As noted in the literature, comprehensive and accessible tools addressing both fidelity and novelty simultaneously remain scarce.
Third, the investigation of data preprocessing effects on generation quality offers practical insights for building effective synthetic data generation pipelines. Fourth, the study provides evidence-based guidelines for selecting appropriate generation methods based on application requirements, data characteristics, and privacy constraints. Finally, the findings contribute to the broader discussion of responsible AI development, addressing challenges such as model collapse from iterative self-training on AI-generated data.

2. NATIONAL COST SAVINGS: MACRO-LEVEL EVIDENCE ON THE HEALTH ECONOMY

2.1 Systematic Evidence on AI's System-Level Economic Impact
Lee et al. (2025) conducted a systematic review synthesizing evidence on the macro- and system-level impact of AI implementation across three key domains: the health economy, workforce productivity, and administrative efficiency, drawing on literature published between 2020 and 2025 that evaluated AI's impact on national or regional health expenditure, labor restructuring, or process efficiency. This systematic evidence base provides an important complement to national policy frameworks by grounding the cost-savings case for healthcare AI adoption in synthesized empirical literature rather than projective estimation alone.

2.2 Contextualizing Cost Savings Against Rising Systemic Pressure
The economic case for AI adoption in healthcare gains urgency from the scale of the financial pressure facing national health systems, with healthcare expenditure as a share of gross domestic product projected to continue rising across many national contexts, compounding pressure from aging populations and healthcare workforce shortages that leave many health systems struggling to meet demand even before considering AI's potential contribution to cost containment (Lee et al., 2025).

2.3 National Frameworks Complementing the Systematic Evidence Base
National policy frameworks have estimated national-scale cost savings attributable to AI-driven predictive modeling and early intervention, framework-level projections that the systematic, synthesized evidence base of Lee et al. (2025) helps validate and contextualize against the broader body of published economic evaluation literature, strengthening the overall national cost-savings case for AI adoption in healthcare beyond any single national framework alone.

3. PRODUCTIVITY GAINS: WORKFORCE AND ADMINISTRATIVE EFFICIENCY

3.1 Administrative Efficiency as a Documented Impact Domain
Lee et al. (2025) explicitly identified administrative efficiency as one of the three core domains through which AI generates macro-level healthcare system impact, synthesizing evidence on AI's capacity to automate and streamline non-clinical tasks including clinical documentation and claims processing—precisely the class of administrative function targeted by Nguyen's (2026b) national framework for transforming prior authorization through artificial intelligence, which aims explicitly to reduce administrative burden and improve patient access.

3.2 Workforce Productivity Amid Systemic Shortage
Beyond administrative efficiency narrowly defined, Lee et al. (2025) situate AI's workforce productivity contribution within the broader context of healthcare workforce shortage and geographic maldistribution, a systemic challenge that

AI-driven productivity enhancement can partially, though not fully, address by enabling existing clinical and administrative staff to accomplish more within a given period of time, redirecting time previously spent on documentation and administrative tasks—which time and motion studies have found consume a substantial proportion of clinician working hours—toward direct patient care.

3.3 Prior Authorization Transformation as a Targeted Productivity Intervention

Nguyen (2026b) frames AI-driven prior authorization transformation as a targeted productivity intervention addressing one of the most burdensome and costly administrative processes within the United States healthcare system, complementing the broader, macro-level administrative efficiency evidence synthesized by Lee et al. (2025) with a specific, national-scale policy framework aimed at realizing these productivity gains in a particularly high-burden administrative domain.

4. PUBLIC HEALTH IMPROVEMENT: PRECISION PREVENTION AND THE IMPERATIVE OF EQUITY

4.1 Precision Public Health for Chronic Disease Prevention

Nguyen (2026a) articulated a national framework for AI-powered precision public health explicitly targeting chronic disease prevention and early intervention, proposing that predictive analytics integrating clinical, behavioral, and social determinants of health data can identify individuals and communities at elevated chronic disease risk with sufficient lead time to enable meaningfully preventive, rather than purely reactive, public health intervention.

4.2 Health Equity as a Precondition for Genuine Public Health Improvement

Realizing genuine, broadly distributed public health improvement from AI adoption, however, requires explicit attention to health equity considerations that a purely technical precision public health framework may not automatically address. Dankwa-Mullan (2024) examined health equity and ethical considerations in the use of artificial intelligence in public health and medicine, highlighting that AI systems can inadvertently perpetuate existing health disparities through algorithmic bias when not carefully managed, and emphasizing community engagement, inclusive data practices, and transparent algorithms as necessary strategies for promoting health equity and ethical AI use in public health and medical contexts.

4.3 Reconciling Precision Targeting With Equitable Distribution of Benefit

The precision public health logic articulated by Nguyen (2026a)—targeting intervention toward those individuals and communities identified as highest risk—and the equity imperative articulated by Dankwa-Mullan (2024)—ensuring that AI systems do not systematically underrepresent or misclassify risk among historically marginalized populations—must be reconciled rather than treated as independent considerations. A precision public health system trained predominantly on data from well-resourced healthcare settings risks systematically underestimating risk among populations underrepresented in its training data, precisely the algorithmic bias concern that Dankwa-Mullan (2024) identifies as a central health equity challenge for AI in public health and medicine.

4.4 Community Engagement and Transparent Algorithms as Design Requirements

Dankwa-Mullan (2024) explicitly recommends community engagement and algorithmic transparency as design requirements for equitable AI use in public health, principles directly applicable to the operationalization of precision public health frameworks such as that of Nguyen (2026a): a precision public health system whose risk targeting logic cannot be transparently explained to, or meaningfully validated by, the communities it is intended to serve risks undermining the public trust necessary for its recommendations to be acted upon, regardless of its underlying technical accuracy.

5. ENVIRONMENTAL SUSTAINABILITY: THE INFRASTRUCTURE COST OF SOCIETAL-SCALE AI DEPLOYMENT

5.1 Public Health Costs of AI Computing Infrastructure

Han et al. (2024) introduced a principled methodology for modeling air pollutant emissions associated with data centers and estimating their resulting public health impacts, finding that the growing demand for AI and computing technologies is projected to push the total annual public health burden of United States data centers to more than twenty billion dollars by 2028, with these costs distributed highly unevenly across host communities such that the most affected counties bear a per-household health burden approximately seven times the national average—a finding with direct relevance to the health equity considerations discussed in Section 4, since the communities bearing disproportionate data center-related health costs may not be the same communities receiving the direct clinical and public health benefits of AI-driven healthcare applications running on that infrastructure.

5.2 Carbon and Water Footprint of AI Systems

De Vries-Gao (2025) estimated that the carbon footprint of AI systems alone could reach between 32.6 and 79.7 million tons of carbon dioxide emissions in 2025, with an associated water footprint potentially reaching between 312.5 and 764.6 billion liters, environmental costs that apply proportionally to the share of overall AI computing demand attributable to healthcare-sector deployment as adoption continues to expand across the domains reviewed in this article.

5.3 Environmental Sustainability as an Integral Societal Impact Dimension

A genuinely comprehensive assessment of AI's broader societal impact in healthcare must treat environmental sustainability not as a peripheral concern but as integral to the same societal impact framework encompassing cost savings, productivity, and public health improvement, given that the environmental and public health costs of computing infrastructure are borne by society at large, including populations who may never directly benefit from the specific healthcare AI applications that infrastructure supports.

5.4 Geothermal Energy and Rigorous Reservoir Characterization

Geothermal energy has emerged as a candidate low-carbon power source capable of supplying the continuous, weather-independent baseload power that AI data center operations require, offering a potential pathway toward more environmentally sustainable healthcare AI infrastructure. Realizing this pathway responsibly depends on rigorous subsurface characterization: high-resolution petrophysical characterization and reservoir quality evaluation, as demonstrated for the Three Forks geothermal reservoir in the Williston Basin, illustrate how well log, core, and seismic attribute data can be integrated to characterize reservoir quality at high spatial resolution (Yeboah et al., 2026), providing the detailed subsurface science necessary to identify and develop geothermal resources capable of reliably supplying data center-scale power demand.

5.5 Geomechanical Risk Management for Responsible Scaling

Scaling geothermal energy production to serve data center-scale power demand responsibly requires proactive geomechanical risk management to avoid introducing new environmental or safety risks while addressing the environmental cost of AI computing infrastructure. Machine learning-enhanced geomechanical characterization and wellbore stability assessment, integrating stress field estimation, rock strength prediction, and operational parameter analysis into a proactive risk-management workflow, has been demonstrated in offshore energy reservoir settings (Akpabli et al., 2026), offering a directly transferable methodology for ensuring that geothermal energy expansion intended to mitigate AI's environmental footprint is itself pursued responsibly.

6. TOWARD A COMPREHENSIVE FRAMEWORK FOR BROADER SOCIETAL IMPACT

6.1 Interdependence Across the Four Societal Impact Domains

The four domains examined in this article—national cost savings, productivity gains, public health improvement, and environmental sustainability—are deeply interdependent rather than independently assessable: the workforce productivity gains documented by Lee et al. (2025) and the administrative efficiency targeted by Nguyen (2026b) depend on the same computing infrastructure whose environmental and public health costs are documented by Han et al. (2024) and de Vries-Gao (2025), while the public health improvement promised by precision prevention frameworks (Nguyen, 2026a) depends fundamentally on the equity safeguards articulated by Dankwa-Mullan (2024) to ensure benefits are genuinely, rather than nominally, distributed across all communities.

6.2 Geothermal Energy as a Case Study in Integrated Societal Benefit

Geothermal energy's potential contribution to healthcare AI's environmental sustainability illustrates how a single infrastructure investment, pursued responsibly and grounded in rigorous subsurface science (Yeboah et al., 2026; Akpabli et al., 2026), can simultaneously support the continued expansion of the cost savings, productivity, and public health benefits documented elsewhere in this article while addressing the environmental cost that a purely benefit-focused societal impact narrative would otherwise omit.

6.3 Equity and Transparency as Cross-Cutting Requirements

Health equity considerations (Dankwa-Mullan, 2024) and environmental justice considerations (Han et al., 2024) share a common structural feature: both concern the risk that AI's societal benefits and costs may be unevenly, and potentially unfairly, distributed across different populations and communities. A comprehensive framework for assessing AI's broader societal impact in healthcare must therefore explicitly track not only aggregate, national-level cost savings, productivity gains, and public health improvement, but also the distribution of these benefits and the associated environmental and public health costs across the full range of communities affected by AI adoption.

7. CHALLENGES AND LIMITATIONS

Despite growing recognition of the need for broader societal impact assessment, several challenges continue to constrain its practical realization.

Fragmentation across evidence types and evaluation traditions. The systematic economic and productivity evidence synthesized by Lee et al. (2025), the health equity and ethical considerations literature exemplified by Dankwa-Mullan (2024), and the environmental and public health infrastructure literature (Han et al., 2024; de Vries-Gao, 2025) have developed as largely separate research and policy traditions, with limited methodological integration across them.

Data limitations for equity-focused evaluation. Assessing whether AI's public health improvement benefits are equitably distributed requires demographically disaggregated outcome data that is not uniformly available across the precision public health and administrative efficiency literature reviewed in this article, constraining the field's current capacity to empirically validate the equity considerations that Dankwa-Mullan (2024) identifies as essential.

Attribution difficulty across shared infrastructure and macro-level economic estimates. As in other domains of AI societal impact assessment, healthcare AI applications typically share computing infrastructure with other sectors, and macro-level economic impact estimates such as those synthesized by Lee et al. (2025) often aggregate across diverse AI applications and national contexts in ways that complicate precise attribution of specific costs and benefits to specific AI systems or healthcare sub-sectors.

Uncertain timelines for environmentally sustainable infrastructure scaling. While geothermal energy offers a conceptually attractive pathway toward more environmentally sustainable healthcare AI infrastructure, the reservoir characterization and geomechanical risk validation processes reviewed in Section 5 (Yeboah et al., 2026; Akpabli et al., 2026) operate on timescales that may not align with the more rapid pace at which healthcare AI computing demand is currently expanding.

Absence of standardized comprehensive societal impact reporting. No standardized framework currently requires healthcare AI developers or health systems to report on cost savings, productivity gains, public health improvement, and environmental impact within a single, integrated societal impact disclosure, leaving the comprehensive assessment this article attempts to synthesize reliant on disparate, independently published literatures rather than a unified reporting standard.

8. FUTURE DIRECTIONS

Future research should prioritize the development of integrated societal impact reporting standards for healthcare AI systems, explicitly requiring disclosure across economic, productivity, public health equity, and environmental dimensions within a single framework, building on the macro-level systematic evidence synthesis approach demonstrated by Lee et al. (2025). Extending health equity and algorithmic bias monitoring, following the principles articulated by Dankwa-Mullan (2024), to explicitly cover precision public health applications such as that developed by Nguyen (2026a) represents a priority direction for ensuring that public health improvement claims are validated with demographically disaggregated outcome data rather than assumed from aggregate results alone.

Advancing geothermal reservoir characterization and geomechanical risk management specifically in support of data center-scale power supply, building on high-resolution petrophysical characterization (Yeboah et al., 2026) and machine learning-enhanced geomechanical characterization (Akpabli et al., 2026), represents a promising applied research direction for realizing environmentally sustainable healthcare AI infrastructure responsibly. Finally, extending the systematic economic evidence synthesis approach of Lee et al. (2025) to explicitly incorporate the environmental and public health externalities documented by Han et al. (2024) and de Vries-Gao (2025) represents an important methodological direction for future macro-level reviews of AI's societal impact in healthcare, ensuring that future systematic evidence synthesis captures the full breadth of costs alongside benefits.

9. CONCLUSION

Evaluating artificial intelligence's broader societal impact in healthcare requires moving beyond narrow clinical or financial performance metrics toward a comprehensive assessment spanning national cost savings, workforce and administrative productivity gains, equitably distributed public health improvement, and environmental sustainability. Systematic macro-level evidence on AI's economic and productivity impact, national frameworks for precision public health and administrative transformation, and public health perspectives on health equity collectively establish a substantial societal benefit case for AI adoption in healthcare, while evidence on algorithmic bias risk and the environmental and public health costs of AI computing infrastructure reveal genuine, currently underweighted costs and equity considerations that a comprehensive societal impact assessment cannot omit. Geothermal energy, grounded in rigorous reservoir characterization and geomechanical risk management, offers a promising pathway for supporting environmentally sustainable AI infrastructure, provided its development is pursued with the same attention to equity and responsible risk management that comprehensive societal impact assessment demands

across all domains. Advancing toward a genuinely comprehensive understanding of AI's broader societal impact in healthcare will require sustained investment in integrated impact reporting, equity-focused outcome monitoring, and environmentally responsible infrastructure development—investments essential to ensuring that AI adoption in healthcare delivers genuine, equitably distributed, and environmentally sustainable societal benefit.

REFERENCES

1. Akpabli, G., Yeboah, N. N., Gyimah, E., Agyei, E., Kojo, A. B., Rahnema, H., & Okyere Williams, M. (2026, June). Machine learning enhanced geomechanical characterization and wellbore stability assessment in the offshore South Tano Field, Ghana. Paper presented at the ARMA US Rock Mechanics/Geomechanics Symposium (p. D022S045R001). ARMA.
2. Dankwa-Mullan, I. (2024). Health equity and ethical considerations in using artificial intelligence in public health and medicine. Preventing Chronic Disease, 21, Article 240245. https://doi.org/10.5888/pcd21.240245
3. de Vries-Gao, A. (2025). The carbon and water footprints of data centers and what this could mean for artificial intelligence. Patterns, 7(1), Article 101430. https://doi.org/10.1016/j.patter.2025.101430
4. Han, Y., Wu, Z., Li, P., Wierman, A., & Ren, S. (2024). Health-informed computing: Estimating and addressing the public health impact of data centers (arXiv:2412.06288). arXiv. https://doi.org/10.48550/arXiv.2412.06288
5. Lee, J. T., Liu, V. T. N., Ali, S., Chen, P.-C., Lee, C.-C., Huang, C.-W., Li, V. C.-S., Chen, H.-H., Duh, W.-J., Marthias, T., & Atun, R. (2025). The impact of artificial intelligence on the health economy, workforce productivity, and administrative efficiency: A systematic review. medRxiv. https://doi.org/10.1101/2025.10.05.25337345
6. Nguyen, T. T. (2026a). AI-powered precision public health: A national framework for targeting chronic
7. disease prevention and early intervention. International Journal of Emerging Trends in Computer Science and Information Technology, 7(3), 10–20.
8. Nguyen, T. T. (2026b). Transforming prior authorization through artificial intelligence: A national framework for reducing administrative burden and improving patient access. International Journal of AI, BigData, Computational and Management Studies, 7(3), 17–27.
9. Yeboah, N. N., Agyei, E., Gyimah, E., Ayensigna, A. A., Okyere Williams, M., Akpabli, G.,... & Vidzro, B. (2026, June). Petrophysical characterization and reservoir quality evaluation of the Three Forks geothermal reservoir, Williston Basin. Paper presented at the ARMA US Rock Mechanics/Geomechanics Symposium (p. D032S048R012). ARMA.
10. Seetala, S. R. (2025). Architecting autonomous data platforms: Integrating AI-driven governance, metadata intelligence, and data mesh principles. International Journal of Science, Engineering and Technology, 13(1).
11. Vasa, M. R. (2024). Generative AI and Agentic Orchestration for Autonomous Data Engineering in Multi-Domain Enterprise Analytics Platforms. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(4), 364-369.
12. Mudusu, S. K. (2025). Data Engineering Challenges in AI-Driven Healthcare IT Systems: Navigating Real-Time Analytics and Interoperability.Kesarpu, S. (2025). Zero-Trust Architecture in Java Microservices. International Journal of Networks and Security, 5(01), 202-214.
13. Narapareddy, V. S. R., & Yerramilli, S. K. (2022). Scaling the ServiceNow CMDB for distributed infrastructures. International Journal of Engineering Technology Research & Management, 6(10), 101–113.
14. Gollapudi, R. (2022). Risk-controlled near-zero-downtime Oracle database migration using GoldenGate. International Journal of Computational and Experimental Science and Engineering, 8(3), 113–123. https://doi.org/10.22399/ijcesen.5382