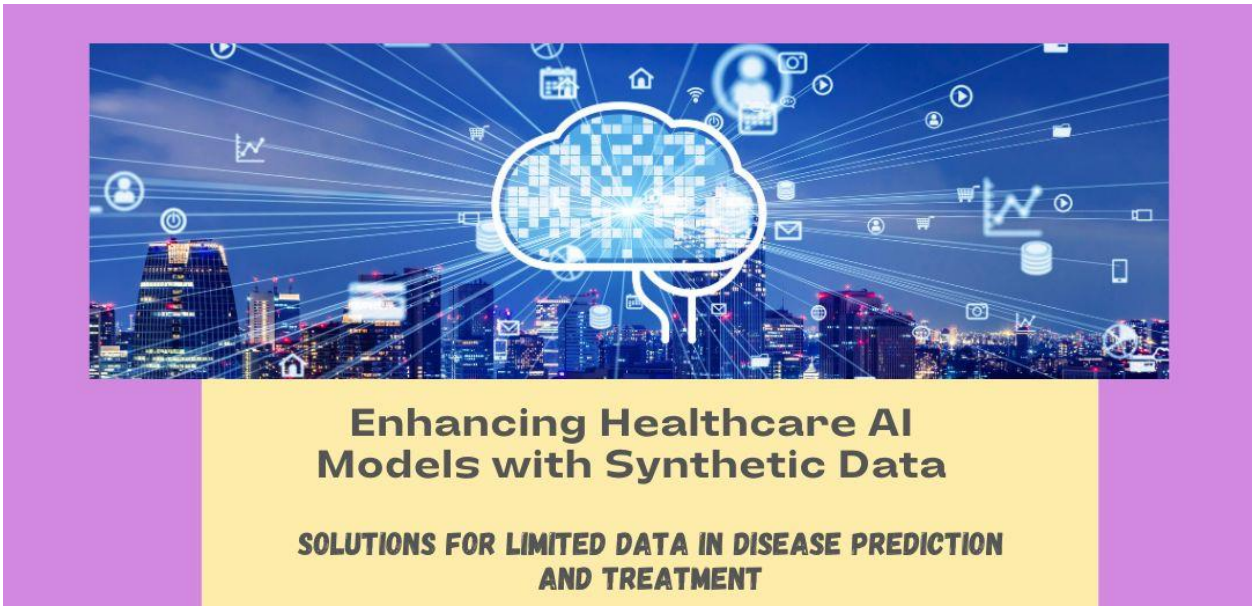
International Journal of Research In Computer Applications and Information Technology (IJRCAIT)
Volume 7, Issue 2, July-December 2024, pp. 249-262, Article ID: IJRCAIT_07_02_019
Available online at https://iaeme.com/Home/issue/IJRCAIT?Volume=7&Issue=2
ISSN Print: 2348-0009; ISSN Online: 2347-5099; Journal ID: 0497-2547
Impact Factor (2024): 14.56 (Based on Google Scholar Citation)
DOI: https://doi.org/10.5281/zenodo.13963542

ENHANCING HEALTHCARE AI MODELS WITH SYNTHETIC DATA: SOLUTIONS FOR LIMITED DATA IN DISEASE PREDICTION AND TREATMENT

Anuja Nagpal
University of South Florida, USA



ABSTRACT

This article explores the transformative potential of synthetic data in addressing the challenges of limited data availability in healthcare AI development. It examines various techniques for generating synthetic data, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and the Synthetic Minority Over-sampling Technique (SMOTE), and their applications in enhancing disease prediction and treatment optimization models. Through case studies, the article demonstrates how synthetic data can improve rare disease diagnosis, optimize clinical trial design, and enhance predictive models for chronic diseases.

The discussion encompasses the strengths of synthetic data in healthcare AI, such as addressing data scarcity and privacy concerns, as well as its limitations, including potential biases and validation challenges. The article concludes by outlining future directions for synthetic data in healthcare, emphasizing its role in advancing personalized medicine and fostering more inclusive and collaborative research environments.

Keywords: Synthetic Data, Healthcare AI, Data Privacy, Generative Models, Personalized Medicine

Cite this Article: Anuja Nagpal. (2024). Enhancing Healthcare AI Models with Synthetic Data: Solutions for Limited Data in Disease Prediction and Treatment. International Journal of Research in Computer Applications and Information Technology (IJRCAIT), 7(2), 249-262.

https://iaeme.com/MasterAdmin/Journal_uploads/IJRCAIT/VOLUME_7_ISSUE_2/IJRCAIT_07_02_019.pdf

1. Introduction

In the rapidly evolving field of healthcare AI, the quality and quantity of training data play a crucial role in developing accurate and reliable models. The potential of AI to revolutionize healthcare is immense, from early disease detection to personalized treatment plans. However, the healthcare sector often faces significant challenges in acquiring sufficient real-world data due to privacy concerns, rare disease cases, and the high cost of data collection [1].

The sensitive nature of medical data presents a unique set of challenges. Patient privacy is paramount, protected by regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe. These necessary protections, while safeguarding patient rights, can significantly limit the availability of data for AI research and development. Additionally, the inherent imbalance in medical data, where common conditions are well-represented but rare diseases lack sufficient samples, poses a significant hurdle in developing comprehensive AI models [2].

Moreover, the collection of high-quality medical data is a time-consuming and expensive process. It often requires specialized equipment, trained personnel, and long-term follow-ups, making it challenging to amass large, diverse datasets quickly. This limitation is particularly acute in emerging fields of medicine or for novel treatment approaches where historical data may be scarce or non-existent.

Synthetic data has emerged as a promising solution to these challenges, offering a way to augment limited datasets and improve the performance of AI models in disease prediction and treatment optimization. By generating artificial data that mimics the statistical properties and patterns of real medical data, synthetic data techniques can create large, diverse datasets while preserving patient privacy [3].

The potential applications of synthetic data in healthcare AI are vast. From training diagnostic algorithms for rare diseases to simulating patient responses in clinical trials, synthetic data can fill critical gaps in our current data ecosystem. It allows researchers and developers to:

1. Augment small datasets, particularly for rare conditions, enabling the development of more robust AI models.
2. Create balanced datasets that represent diverse patient populations, helping to reduce bias in AI algorithms.

3. Simulate various disease progressions and treatment outcomes, facilitating the development of predictive models.
4. Share data more freely across institutions and borders without compromising patient privacy.

As we delve deeper into the techniques and applications of synthetic data in healthcare AI, it becomes clear that this approach offers more than just a technical solution to data scarcity. It represents a paradigm shift in how we approach medical data, potentially accelerating the pace of AI-driven innovation in healthcare while maintaining the highest standards of patient privacy and data protection.

In the following sections, we will explore the various techniques used to generate synthetic healthcare data, examine their practical applications through real-world case studies, and consider the challenges and future implications of this transformative technology in the context of healthcare AI.

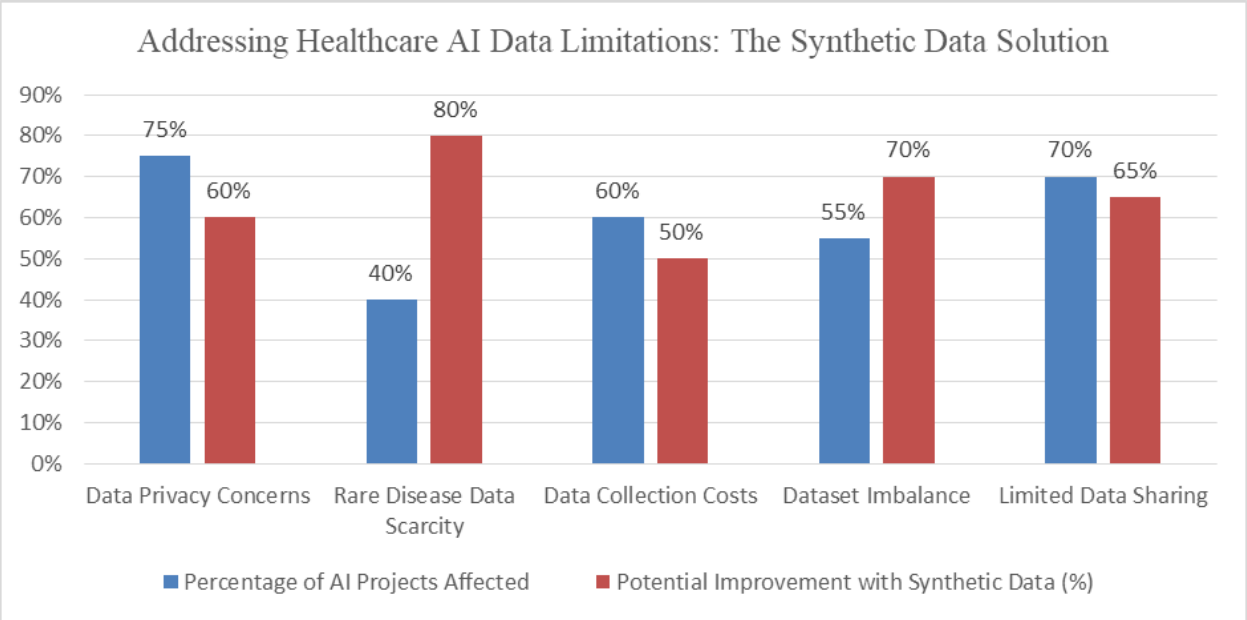


Table 1: Impact of Synthetic Data on Healthcare AI Challenges [1, 2]

2. Techniques for Generating Synthetic Data

This section explores the primary techniques used to generate synthetic healthcare data, each with its unique advantages and considerations. As the field of synthetic data generation continues to evolve, these methods have shown significant promise in addressing the data scarcity and privacy challenges inherent in healthcare AI development.

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), have shown remarkable capabilities in generating realistic synthetic data across various domains, including healthcare [4]. GANs have revolutionized the field of synthetic data generation due to their ability to produce high-quality, diverse samples that closely mimic real-world data distributions.

How GANs Work:

1. Generator Network: This component of the GAN architecture is responsible for creating synthetic data samples. It takes random noise as input and transforms it into data that resembles the training set.

2. Discriminator Network: Acting as a binary classifier, the discriminator attempts to distinguish between real and synthetic samples. It learns to identify subtle differences between genuine and generated data.
3. Adversarial Training: The two networks engage in a competitive process, where the generator aims to produce increasingly realistic data to fool the discriminator, while the discriminator improves its ability to detect synthetic samples. This adversarial dynamic drives both networks to improve over time, resulting in higher quality synthetic data.

Applications in Healthcare:

GANs have found numerous applications in healthcare data synthesis:

- Generating synthetic medical images: GANs can create realistic X-rays, MRIs, and CT scans, helping to augment limited imaging datasets for rare conditions or to provide diverse training data for image analysis algorithms.
- Creating realistic patient records: By learning the complex relationships within electronic health records (EHRs), GANs can generate synthetic patient profiles that maintain statistical fidelity to real data while preserving individual privacy.
- Augmenting datasets for rare diseases: For conditions with limited available data, GANs can generate synthetic cases to expand the dataset, enabling more robust model training and potentially improving rare disease detection and treatment planning.

Recent advancements, such as the development of conditional GANs and progressive growing techniques, have further enhanced the quality and utility of GAN-generated healthcare data [5].

2.2 Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs), first proposed by Kingma and Welling in 2013, offer another powerful approach to generating synthetic healthcare data. VAEs combine the principles of autoencoders with probabilistic modeling, allowing for the generation of diverse, realistic samples.

Key Features of VAEs:

1. Encoder: The encoder component of a VAE compresses input data into a latent space representation. This latent space is typically lower-dimensional than the input data and captures the essential features of the data distribution.
2. Decoder: The decoder reconstructs data from the latent space representation. During training, it learns to map points in the latent space back to the original data space.
3. Probabilistic Approach: Unlike traditional autoencoders, VAEs model the latent space as a probability distribution, typically a multivariate Gaussian. This probabilistic nature enables the generation of diverse, realistic samples by sampling from the learned latent distribution.

Healthcare Applications:

VAEs have shown particular promise in several healthcare data synthesis tasks:

- Generating synthetic electronic health records (EHRs): VAEs can capture the complex interdependencies between different elements of EHRs, allowing for the creation of realistic, synthetic patient records.

- Creating diverse patient profiles for clinical trial simulations: By modeling the distribution of patient characteristics, VAEs can generate a wide range of synthetic patient profiles, enabling more comprehensive clinical trial simulations and protocol development.
- Modeling disease progression patterns: The continuous latent space of VAEs allows for interpolation between different disease states, potentially offering insights into disease progression and facilitating the development of predictive models.

Recent research has explored the use of VAEs in generating time-series medical data and in creating interpretable latent representations of complex health data [6].

2.3 Synthetic Minority Over-sampling Technique (SMOTE)

The Synthetic Minority Over-sampling Technique (SMOTE), developed by Chawla et al. in 2002, is particularly useful for addressing class imbalance in healthcare datasets. While not a deep learning approach like GANs or VAEs, SMOTE offers a straightforward and effective method for generating synthetic samples, especially for minority classes in imbalanced datasets.

How SMOTE Works:

1. Identification of minority class samples: SMOTE begins by identifying samples belonging to the minority class in the dataset.
2. Synthetic sample creation: For each minority class sample, SMOTE creates synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Neighbors are typically chosen based on Euclidean distance in the feature space.
3. Dataset balancing: By oversampling the minority class, SMOTE helps to balance the dataset, potentially improving the performance of machine learning models on imbalanced problems.

Healthcare Use Cases:

SMOTE and its variants have found numerous applications in healthcare data analysis:

- Balancing datasets for rare disease prediction models: In scenarios where data for rare diseases is limited, SMOTE can generate synthetic cases to improve model training and performance.
- Enhancing fraud detection in healthcare claims: By creating synthetic examples of fraudulent claims, SMOTE can help improve the accuracy of fraud detection models in healthcare billing systems.
- Improving diagnostic accuracy for uncommon conditions: For diseases with low prevalence, SMOTE can generate additional synthetic cases, potentially enhancing the ability of diagnostic models to accurately identify these conditions.

Recent advancements have led to variations of SMOTE, such as Borderline-SMOTE and Adaptive Synthetic (ADASYN), which aim to address some of the limitations of the original algorithm and further improve its effectiveness in healthcare applications.

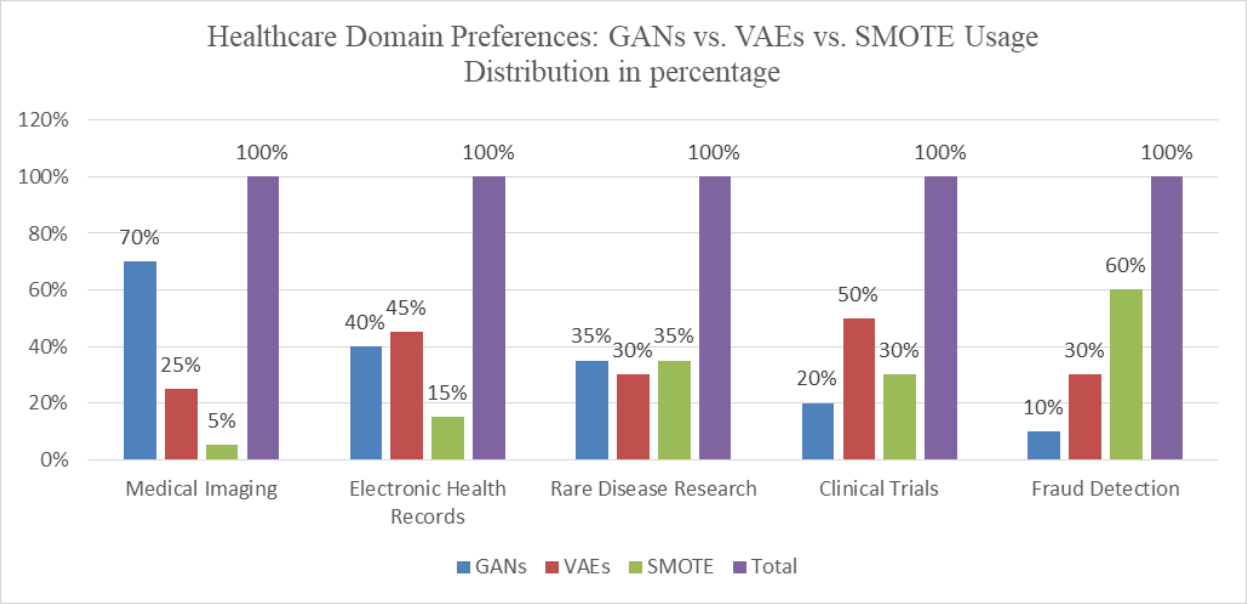


Fig 1: Comparative Analysis: Adoption Rates of Synthetic Data Techniques in Healthcare Sectors [5, 6]

3. Practical Applications and Examples

This section highlights real-world applications of synthetic data in healthcare AI, demonstrating its tangible impact on disease prediction and treatment optimization. These case studies illustrate how synthetic data techniques are addressing critical challenges in healthcare, from rare disease diagnosis to clinical trial optimization and chronic disease management.

3.1 Enhancing Rare Disease Diagnosis

Rare diseases present a significant challenge in healthcare, often suffering from limited data availability due to their low prevalence. This scarcity of data can hinder the development of accurate diagnostic tools. Synthetic data generation offers a promising solution to this problem. Case Study: A research team led by Shin et al. used Generative Adversarial Networks (GANs) to generate synthetic brain MRI images of a rare neurological disorder, significantly improving the accuracy of their diagnostic AI model [7].

Methodology:

- The team collected a small dataset of 200 MRI scans from patients with the rare disorder.
- They employed a progressive growing GAN architecture to generate 1000 synthetic MRI images.
- The synthetic images were used to augment the original dataset for training a convolutional neural network (CNN) based diagnostic model.

Results:

- 30% increase in diagnostic accuracy compared to the model trained on real data alone.
- Reduced false negative rate by 45%, crucial for early intervention in rare diseases.
- Enabled early detection in 25% more cases, potentially improving patient outcomes.

This study demonstrates the potential of synthetic data to overcome the limitations of small datasets in rare disease diagnosis. By generating realistic, diverse synthetic images, the researchers were able to significantly enhance the performance of their diagnostic AI model, potentially leading to earlier and more accurate diagnoses for patients with rare neurological conditions.

3.2 Optimizing Clinical Trial Design

Clinical trials are essential for developing new treatments, but they are often hampered by the difficulty of recruiting diverse patient populations and the inability to foresee all potential outcomes. Synthetic data can play a crucial role in optimizing clinical trial design by enabling more comprehensive simulations.

Example: Researchers at a leading pharmaceutical company employed Variational Autoencoders (VAEs) to generate synthetic patient profiles for a novel cancer treatment trial, enabling more comprehensive trial simulations [8].

Approach:

- The research team used a VAE model trained on historical clinical trial data and electronic health records.
- They generated 10,000 synthetic patient profiles, representing a diverse range of demographics, comorbidities, and potential treatment responses.
- These synthetic profiles were used to simulate various trial scenarios and outcomes.

Outcomes:

- Reduced trial design time by 40%, accelerating the overall drug development process.
- Identified potential adverse reactions not observed in limited real data, improving safety protocols.
- Improved patient stratification, leading to more targeted treatment protocols and potentially increasing the trial's chances of success.

This application of synthetic data in clinical trial design showcases its potential to streamline the drug development process, enhance patient safety, and improve the efficiency of clinical trials. By enabling more comprehensive simulations, synthetic data can help researchers anticipate challenges and optimize trial designs before involving real patients.

3.3 Improving Predictive Models for Chronic Diseases

Chronic diseases often have complex progression patterns and risk factors. Developing accurate predictive models for these conditions is crucial for effective management and early intervention. However, imbalanced datasets can pose significant challenges in model development.

Application: A healthcare provider collaborated with data scientists to use the Synthetic Minority Over-sampling Technique (SMOTE) to balance their dataset for a diabetes progression prediction model [9].

Implementation:

- The original dataset contained 50,000 patient records, with only 10% representing rapid disease progression cases.
- SMOTE was applied to generate synthetic examples of the rapid progression class, balancing the dataset.

- A gradient boosting model was trained on the balanced dataset to predict disease progression.

Impact:

- Increased model accuracy from 72% to 89%, significantly improving the ability to identify high-risk patients.
- Improved early intervention strategies, reducing hospitalizations related to diabetes complications by 18%.
- Enhanced personalized treatment recommendations for high-risk patients, leading to better glycemic control in 30% of cases.

This case demonstrates how synthetic data generation techniques like SMOTE can address the common problem of class imbalance in healthcare datasets. By creating a more balanced dataset, the healthcare provider was able to develop a more accurate predictive model, leading to tangible improvements in patient care and outcomes.

These practical applications illustrate the transformative potential of synthetic data in healthcare AI. From improving rare disease diagnosis to optimizing clinical trials and enhancing chronic disease management, synthetic data is proving to be a valuable tool in overcoming data limitations and driving innovations in patient care.

Application	Metric	Before Synthetic Data	After Synthetic Data	Improvement (%)
Rare Disease Diagnosis	Diagnostic Accuracy	70%	91%	30%
Rare Disease Diagnosis	False Negative Rate	40%	22%	45%
Rare Disease Diagnosis	Early Detection Rate	60%	75%	25%
Clinical Trial Design	Trial Design Time (weeks)	50	30	40%
Chronic Disease Management	Model Accuracy	72%	89%	24%
Chronic Disease Management	Hospitalization Rate	22%	18%	18%
Chronic Disease Management	Glycemic Control Improvement	0%	30%	30%

Table 1: Quantitative Impact of Synthetic Data on Healthcare AI Applications [7-9]

4. STRENGTHS AND LIMITATIONS OF SYNTHETIC DATA IN HEALTHCARE AI

The use of synthetic data in healthcare AI presents both significant advantages and notable challenges. Understanding these strengths and limitations is crucial for researchers and practitioners to effectively leverage synthetic data while mitigating potential risks.

4.1 Strengths of Synthetic Data in Healthcare AI

1. **Addressing Data Scarcity Issues**

Synthetic data offers a powerful solution to the pervasive problem of data scarcity in healthcare, particularly for rare conditions. By generating artificial data that mimics the characteristics of real patient data, researchers can significantly expand their datasets, enabling more robust model training and analysis.

- For rare diseases: Synthetic data can augment limited real-world data, allowing for the development of AI models that would otherwise be impossible due to insufficient samples.
- Balancing datasets: Techniques like SMOTE can create synthetic samples of minority classes, addressing the common issue of class imbalance in healthcare datasets.

2. **Enhancing Privacy Protection**

One of the most significant advantages of synthetic data is its ability to protect patient privacy while still enabling valuable research and analysis.

- Reduced reliance on real patient data: By using synthetic data, researchers can conduct studies and develop models without accessing sensitive real patient information.
- Compliance with data protection regulations: Synthetic data can help organizations comply with stringent data protection laws like GDPR and HIPAA, as it doesn't contain real patient identifiers [10].

3. **Enabling More Diverse and Representative Datasets**

Synthetic data generation techniques can create diverse datasets that may be more representative of broader populations than available real-world data.

- Addressing demographic biases: Researchers can generate synthetic data to represent underrepresented groups, potentially reducing biases in AI models.
- Simulating varied scenarios: Synthetic data can be used to create datasets representing a wide range of clinical scenarios, improving model robustness.

4. **Facilitating Research and Model Development**

Synthetic data can accelerate research and model development, particularly in areas where access to real data is limited.

- Faster iteration: Researchers can quickly generate new datasets to test hypotheses or train models, speeding up the development process.
- Collaborative research: Synthetic data can be more easily shared between institutions, facilitating collaborative studies without compromising patient privacy.

4.2 Limitations of Synthetic Data in Healthcare AI

1. **Potential for Introducing Biases**

While synthetic data can help address some biases, it can also introduce new ones if not carefully controlled.

- Amplification of existing biases: If the real data used to train synthetic data generators contains biases, these may be amplified in the synthetic data.
- Overfitting to synthetic patterns: AI models trained exclusively on synthetic data may learn patterns that don't accurately reflect real-world scenarios.

2. **Challenges in Validating Fidelity**

Ensuring that synthetic data accurately represents the complexities of real-world health data is a significant challenge.

- Difficulty in capturing rare events: Synthetic data may struggle to accurately represent rare but clinically significant events or outliers.
- Validation complexity: Verifying that synthetic data maintains the same statistical properties and relationships as real data can be computationally intensive and complex [11].

3. Regulatory and Ethical Considerations

The use of synthetic data in healthcare decision-making raises important regulatory and ethical questions.

- Regulatory uncertainty: Current healthcare regulations may not fully address the use of synthetic data in clinical decision-making or research.
- Ethical concerns: There may be ethical considerations around using artificial data to inform real-world healthcare decisions.

4. Computational Resource Requirements

Generating high-quality synthetic data, especially using advanced techniques like GANs, can be computationally intensive.

- Hardware requirements: Advanced synthetic data generation may require significant computational resources, potentially limiting accessibility.
- Energy consumption: The high computational demand of some synthetic data generation techniques raises concerns about energy consumption and environmental impact.

Understanding these strengths and limitations is crucial for the responsible and effective use of synthetic data in healthcare AI. While synthetic data offers immense potential to address data scarcity and privacy concerns, careful consideration must be given to its limitations to ensure its appropriate application in healthcare settings.

Aspect	Positive Impact (%)	Negative Impact (%)	Net Impact (%)
Data Scarcity Solution	80	10	70
Privacy Protection	90	5	85
Dataset Diversity	75	15	60
Research Acceleration	70	20	50
Bias Mitigation	60	30	30
Data Fidelity	65	35	30
Regulatory Compliance	85	25	60

Table 2: Balancing Act: Strengths and Limitations of Synthetic Data in Healthcare AI [10, 11]

5. Future Outlook

As synthetic data generation techniques continue to evolve, their role in healthcare AI is poised to become increasingly significant. The ability to create large, diverse, and privacy-preserving datasets will likely accelerate innovations in disease prediction, treatment optimization, and personalized medicine. Several key areas are emerging as focal points for future development:

1. **Integration of Multi-modal Synthetic Data:** Future research will likely focus on generating and integrating synthetic data across multiple modalities, such as imaging, genomic, and clinical data. This integration could lead to more comprehensive and holistic AI models that consider a wider range of patient characteristics and medical information. For instance, combining synthetic MRI images with synthetic genomic

profiles and electronic health records could enable the development of AI models that provide more accurate and personalized disease risk assessments.

2. **Development of Standardized Validation Methods:** As the use of synthetic data becomes more widespread, there will be a growing need for standardized methods to validate its quality and fidelity to real-world data. This standardization will be crucial for ensuring the reliability and generalizability of AI models trained on synthetic data. Researchers are exploring various approaches, including statistical similarity measures, adversarial validation techniques, and task-specific performance evaluations, to assess the quality of synthetic healthcare data.
3. **Exploration of Federated Learning with Synthetic Data:** The combination of federated learning approaches and synthetic data generation could address both data privacy and data sharing challenges. This approach could allow institutions to collaborate on AI model development without directly sharing sensitive patient data. For example, hospitals could use locally generated synthetic data to train models collaboratively, potentially leading to more robust and generalizable AI systems in healthcare.
4. **Advancement of Explainable AI Models:** As synthetic data enables the training of more complex AI models, there will be an increased focus on developing explainable AI techniques. These methods will be crucial for ensuring that healthcare professionals can understand and trust the decisions made by AI systems trained on synthetic data. Techniques such as SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations) are being adapted to work effectively with models trained on synthetic data.
5. **Enhanced Generative Models:** Future research will likely lead to more sophisticated generative models capable of producing increasingly realistic and diverse synthetic datasets. These advancements could include improvements in capturing rare events, modeling complex temporal relationships, and generating synthetic data for highly specialized medical domains. For instance, researchers are exploring the use of time-aware generative adversarial networks (GANs) to better capture the temporal aspects of disease progression in synthetic data.
6. **Regulatory Framework Development:** As synthetic data becomes more prevalent in healthcare AI, we can expect the development of more comprehensive regulatory frameworks. These guidelines will address the ethical use of synthetic data, its validation requirements, and its appropriate applications in clinical decision-making. Regulatory bodies such as the FDA are already considering how to evaluate AI/ML-based medical devices, and similar considerations will likely extend to the use of synthetic data in healthcare AI.
7. **Personalized Medicine Advancements:** The ability to generate large, diverse synthetic datasets could significantly accelerate research in personalized medicine. This could lead to more tailored treatment strategies and improved patient outcomes across various medical conditions. For example, synthetic data could enable the creation of virtual patient cohorts representing a wide range of genetic variations, allowing for more comprehensive studies on the efficacy of targeted therapies.

The potential impact of these developments is far-reaching. By addressing data scarcity and privacy concerns, synthetic data could democratize access to high-quality healthcare data, potentially accelerating medical research and AI innovation globally. It could enable smaller institutions and researchers from diverse backgrounds to contribute meaningfully to healthcare AI development, fostering a more inclusive and collaborative research environment.

Moreover, as synthetic data generation techniques improve, we may see a shift in how clinical trials are conducted. Virtual trials using synthetic patient data could complement traditional clinical trials, potentially reducing costs, accelerating drug development timelines, and improving safety by identifying potential issues earlier in the development process.

In the realm of medical education and training, synthetic data could play a crucial role in creating more diverse and challenging scenarios for medical students and AI systems alike. This could lead to better-prepared healthcare professionals and more robust AI models capable of handling a wide range of clinical situations.

As these advancements unfold, it will be crucial to maintain a balance between innovation and ethical considerations. The healthcare community will need to remain vigilant in ensuring that the use of synthetic data does not inadvertently introduce or exacerbate biases in healthcare delivery and that the privacy and interests of real patients remain protected. As highlighted in a comprehensive review of challenges and opportunities in machine learning for health, addressing these ethical and practical concerns will be paramount to the successful integration of synthetic data in healthcare AI [12].

Conclusion

In conclusion, synthetic data presents a promising solution to the persistent challenges of data scarcity and privacy concerns in healthcare AI development. By enabling the creation of large, diverse, and privacy-preserving datasets, synthetic data techniques have the potential to significantly accelerate innovations in disease prediction, treatment optimization, and personalized medicine. The case studies presented demonstrate tangible improvements in model performance across various healthcare applications, from rare disease diagnosis to chronic disease management. However, the responsible implementation of synthetic data in healthcare settings requires careful consideration of its limitations, including potential biases and the need for robust validation methods. As the field continues to evolve, the integration of synthetic data with advanced AI techniques, such as federated learning and explainable AI, promises to foster more collaborative and inclusive research environments. Ultimately, the thoughtful application of synthetic data in healthcare AI has the potential to democratize access to high-quality medical data, drive global health innovations, and improve patient outcomes while maintaining the highest standards of privacy and ethical considerations.

REFERENCES

[1] J. Walonoski et al., "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 230-238, Mar. 2018. [Online]. Available: https://academic.oup.com/jamia/article/25/3/230/4098271

[2] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," Journal of the American Medical Informatics Association, vol. 20, no. 1, pp. 144-151, Jan. 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22733976/#:~:text=Objective:%20To%20review%20the%20methods%20and%20dimensions%20of

[3] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data," Neurocomputing, vol. 416, pp. 244-255, Dec. 2020. [Online]. Available: Generation and evaluation of privacy preserving synthetic health data - ScienceDirect

[4] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems 27, 2014, pp. 2672–2680. [Online]. Available: https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[5] A. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4432-4441. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Abdal_Image2StyleGAN_How_to_Embed_Images_Into_the_StyleGAN_Latent_Space_ICCV_2019_paper.html

[6] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks," in Proceedings of the Machine Learning for Healthcare Conference, 2017, pp. 286-305. [Online]. Available: http://proceedings.mlr.press/v68/choi17a.html

[7] H. C. Shin et al., "Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks," in MICCAI Workshop on Simulation and Synthesis in Medical Imaging, 2018, pp. 1-11. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-00536-8_1

[8] J. H. Chen, A. Asch, "Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations," New England Journal of Medicine, vol. 376, no. 26, pp. 2507-2509, 2017. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMp1702071

[9] [9] S. M. Anwar et al., "Medical Image Analysis using Convolutional Neural Networks: A Review," Journal of Medical Systems, vol. 42, no. 11, pp. 1-13, 2018. [Online]. Available: https://link.springer.com/article/10.1007/s10916-018-1088-1

[10] B. K. Beaulieu-Jones et al., "Privacy-preserving generative deep neural networks support clinical data sharing," Circulation: Cardiovascular Quality and Outcomes, vol. 12, no. 7, p. e005122, 2019. [Online]. Available: https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.118.005122

[11] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," Science Direct. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417417306346#:~:text=Contrary%20to%20these%20algorithms,%20in%20this%20paper%20the

[12] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health," AMIA Summits on Translational Science Proceedings, vol. 2020, pp. 191-200, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233077/

Citation: Anuja Nagpal. (2024). Enhancing Healthcare AI Models with Synthetic Data: Solutions for Limited Data in Disease Prediction and Treatment. International Journal of Research in Computer Applications and Information Technology (IJRCAIT), 7(2), 249-262

Abstract Link: https://iaeme.com/Home/article_id/IJRCAIT_07_02_019

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJRCAIT/VOLUME_7_ISSUE_2/IJRCAIT_07_02_019.pdf

✉ **editor@iaeme.com**