Custom AI Data Synthesizer for Women's Health

Tia Pope[1][0009−0009−0823−6070] and Ahmad Patooghy[1][0000−0003−2647−2797]

North Carolina A&T State University, Greensboro NC 27411, USA

Abstract. Women's health research faces significant gaps due to historical neglect, leading to underrepresentation in clinical trials and medical research. This deficiency results in undiagnosed conditions, suboptimal healthcare, and adverse outcomes, perpetuating a cycle of neglect and unmet needs. President Joe Biden's recent Executive Order highlights the urgency of prioritizing investments in women's health research, recognizing the underrepresentation of women in clinical trials and medical research. Conditions like endometriosis and postpartum hemorrhage exemplify the social, public health, and economic implications of this neglect. To address these gaps, we propose a custom data synthesizer to generate synthetic data for women's health research. This solution aims to create robust and trustworthy datasets, countering the limitations of current datasets and enabling the generation of relevant data in a timely and scalable manner. Implementing a custom data synthesizer for women's health research is expected to support innovation by providing reliable datasets. By bridging the data availability gap, this solution can accelerate research and innovation, ultimately leading to better health outcomes for women and promoting health equity and inclusivity.

Keywords: Synthetic Data Generation · Large Language Models · Healthcare · Machine Learning · Fast Healthcare Interoperability Resources

1 The Problem: Data Deficiency in Women's Health

Despite comprising over half of the global population, women's health research has been historically overlooked, leading to significant gaps in understanding and addressing their unique health needs. Recognizing this issue, President Joe Biden signed an Executive Order on March 18, 2024, to prioritize women's health research investments and integrate it across federal research portfolios [1]. Women are vastly underrepresented in clinical trials and medical research, with less than 2% of research funding allocated to their reproductive health. This neglect results in undiagnosed conditions, suboptimal healthcare, and adverse outcomes, perpetuating a cycle of neglect and unmet needs in women's health innovation [2]. Conditions such as endometriosis and postpartum hemorrhage exemplify the consequences of neglecting women's health research, leading to significant social, public health, and economic implications. Barriers persist in clinical research, hindering our understanding of women's health conditions. Addressing these issues will improve women's lives and have significant financial benefits, potentially lifting women out of poverty and boosting societal well-being [3], [4].

Improving women's health research is crucial for promoting health equity and inclusivity. By prioritizing investments, integrating women's health across research portfolios, and assessing unmet needs, we can address the underrepresentation of women in medical research and pave the way for better health outcomes and innovation [2], [3], [4]. To be clear, this lack of research has led to little to no robust or trustworthy datasets on women's health. Synthetic data generation helps close the gap and increase the speed of research and innovation.

1.1 Literature Review

We conducted an independent literature review and found that large language models (LLMs) have propelled advancements in natural language processing (NLP) but encounter obstacles in deployment due to cost, responsiveness, and privacy concerns. Our review evaluated LLMs' role in synthetic data generation across domains like healthcare, robotics, and text classification, aiming to mitigate reliance on expensive, human-labeled data. In healthcare, machine learning methods, particularly generative adversarial networks (GANs) and variational autoencoders (VAEs), generate synthetic data to preserve patient privacy and aid algorithm development. Similarly, LLMs contribute to robotics by creating synthetic environments and expert demonstrations, enhancing task-level generalization. Text classification benefits from LLMs to augment data diversity and accuracy, though effectiveness varies. Challenges include evaluating output quality, addressing bias, and ensuring privacy. Despite obstacles, LLMs offer advantages such as reduced labeling needs, enhanced privacy, and improved generalization. Future research is needed to address these challenges and fully harness synthetic data's potential. This synthesis emphasizes the transformative impact of synthetic data in healthcare, advocating for interdisciplinary efforts to ensure responsible and effective deployment. Our review provides a robust foundation for developing a custom data synthesizer tailored for women's health research. Insights from the review underscore the potential of synthetic data generation, particularly LLMs and generative AI techniques like GANs, combined with other methods. The synthesizer can facilitate women's health research advancements by addressing key challenges such as data scarcity, privacy concerns, and the need for diverse and representative datasets. The synthesizer aims to produce high-quality, realistic data suitable for various research applications by incorporating methodologies to evaluate output quality and mitigate biases. Moreover, interdisciplinary collaboration and ethical considerations highlighted in the review underscore the importance of engaging experts from diverse fields to ensure the synthesizer aligns with best practices and addresses the unique challenges and opportunities in women's health research.

2 The Goal: Custom Data

The introduction of mainstream AI tools and LLMs has opened new horizons in healthcare, offering unprecedented opportunities for data analysis, decision

support, and patient care optimization. Synthetic data emerges as a cornerstone in this evolution, providing a pathway to harness the power of data while safeguarding patient privacy. The comprehensive literature review conducted before this proposal has guided us in developing the proposed research topics for this project. This proposal focuses on leveraging LLMs to generate and enhance synthetic data and its security challenges in healthcare, aiming to establish a more trustworthy, reliable, and ethical framework for healthcare intelligence.

The need for voluminous and diverse datasets grows as the healthcare industry leans increasingly towards AI-driven solutions. However, the reliance on actual patient data raises privacy and ethical concerns. Synthetic data from LLMs could help, but questions remain about its effectiveness and ethics. Our research aims to understand synthetic data in healthcare, noting its strengths and limits in quality and scale, considering how AI systems grow larger and demand more data. Overall, AI heavily depends on data, making its quality vital. While tools like Synthea exist, more straightforward solutions are needed. This research seeks to create a framework for generating and using synthetic data effectively and ethically in healthcare.

2.1 Research Questions

This research will try to address the following research questions.

- What are the methods for generating synthetic multimodal data in healthcare, particularly focusing on LLMs and generative AI, and how can we evaluate the quality and representativeness of such data?
- How are security measures applied to protect synthetic healthcare data, considering privacy regulations and ethical concerns, and what strategies and standards exist for integrating synthetic multimodal data into healthcare systems?
- How do we validate the usefulness of synthetic healthcare data and demonstrate its impact on intelligence, decision-making, and outcomes in women's health research and real-world scenarios?

3 The Methodology: Our Planned Approach

This research project will utilize a comprehensive methodology to explore and tackle various hypotheses, emphasizing women's health. Our approach will encompass a range of techniques, including data collection, the development of customized frameworks tailored to women's unique needs, rigorous quality assessments, and sophisticated statistical analyses. By integrating these methods, we aim to delve into critical aspects of women's health and contribute to advancing knowledge in this vital area. We will prioritize diseases and medical issues predominantly affecting women, including breast, cervical, and ovarian cancers, reproductive and menopausal health, and cardiovascular conditions. This emphasis aligns with the NIH's strategic priorities to mitigate health disparities and facilitates the development of targeted interventions and treatments.

We propose a novel data synthesizer for women's health research (see Fig. 1). This innovative framework combines standardization with customizability, offering modularity to integrate seamlessly into research and clinical environments (see Fig. 2). Figure 2 considers data collection [A] data synthesis [B] and data validation [C] approaches.

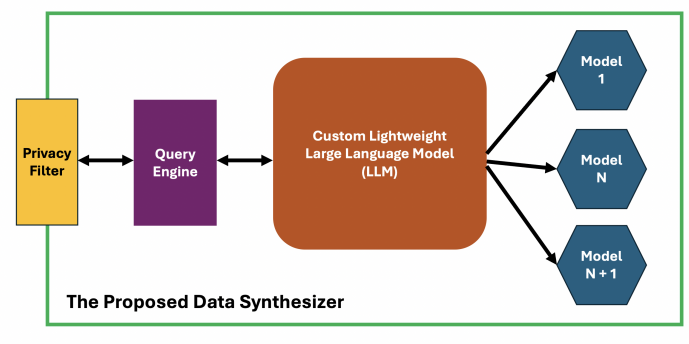

Fig. 1. Illustration of the Proposed Data Synthesizer

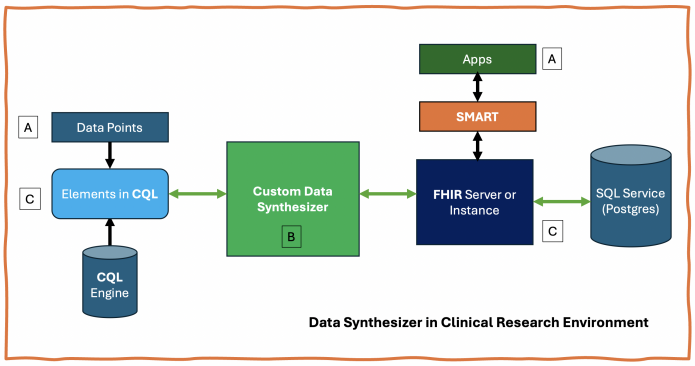

Fig. 2. Illustration of the Proposed Data Synthesizer in a Clinical Research Environment

4 The Expected Results: Our Hypotheses

We hypothesize that synthetic healthcare data can be generated through the strategic application of LLMs that maintain high utility for research and decision-making while addressing critical security, reliability, and ethics concerns. We propose two key sub-hypotheses as part of this proposal:

- **Hypothesis 1** suggests that using synthetic data in healthcare research can enhance security and privacy, lowering the risk of breaches and encouraging innovation, especially in women's health.
- **Hypothesis 2** builds on this idea, stating that improving methods for creating realistic synthetic healthcare data will make healthcare intelligence from synthetic datasets more reliable and useful, particularly in women's health research.

5 The Inquiries: My Needs

Question 1 Are my research goals too broad? Are there additional areas I should consider? Must consider?
Question 2 Are there specific methodologies or techniques I should prioritize or explore further? Is there any I should eliminate?
Question 3 What are my proposed research's potential limitations or challenges? Any particular gaps from a research or technical perspective?
Question 4 What areas of genomics could I incorporate?
Question 5 Mathematically, are there theorems I should consider?
Question 6 How should I evaluate my work?
Question 7 Are there specific resources or collaborations I should pursue to enhance my research?

Acknowledgments. Sponsored by Department of Education Title III HBGI Grant.

Disclosure of Interests. We have no competing interests to disclose.

References

1. New Actions to Advance Women's Health Research and Innovation,https://www.whitehouse.gov/briefing-room/statements-releases/2024/03/18/fact-sheet-president-biden-issues-executive-order-and-announces-new-actions-to-advance-womens-health-research-and-innovation/, last accessed 2024/4/10
2. Women are second-class citizens when it comes to health. Closing the gap could be worth $1 trillion,https://www.weforum.org, last accessed 2024/4/10
3. Women's health research lacks funding, https://www.nature.com/immersive/d41586-023-01475-2/index.html, last accessed 2024/4/10
4. Closing the women's health gap: A $1 trillion opportunity to improve lives and economies, https://www.mckinsey.com/mhi/our-insights/closing-the-womens-health-gap-a-1-trillion-dollar-opportunity-to-improve-lives-and-economies, last accessed 2024/4/10