Evaluating Approaches to Integrating Social Determinants in Synthetic Data Generation

Yili Zhang, PhD, Jia Li Dong, MS, Kanchi Krishnamurthy, MS, Adil Alaoui, MS, BMA, Peter McGarvey, PhD

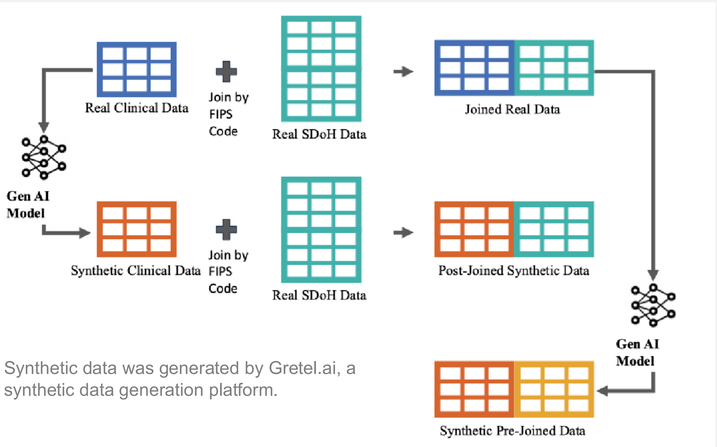
Innovation Center for Biomedical Informatics, Georgetown University, Washington, DC

Background

Synthetic medical data[1] facilitates secure research and data sharing.[2] However, effective methods for combining synthetic clinical data with Social Determinants of Health (SDoH) , which is crucial for equitable healthcare but remain underexplored.

Goal: To evaluate two approaches for integrating SDoH data with synthetic clinical data generated by an LLM-based generative AI model.

SDoH Data Integration Methods with Synthetic Data



Synthetic data was generated by Gretel.ai, a synthetic data generation platform.

Figure 1. Real, Post-Joined Synthetic, and Pre-Joined Synthetic Data Generation and Integration Pipeline.

Table 1. Integration Strategies Compared

Approach	Synthesis	SDoH Data	Key Feature
1. Baseline (Real Data)	No synthesis	Real SDoH	Clinical data joined with real SDoH by real FIPS code
2. Post-Joined	Clinical Data Alone	Real SDoH	Synthetic clinical data is separately generated, then joined with real SDoH data using synthetic geographic identifiers.
3. Pre-Joined	Clinical+SDoH	Synthetic SDoH	Clinical and SDoH variables are simultaneously generated (fully synthetic data).

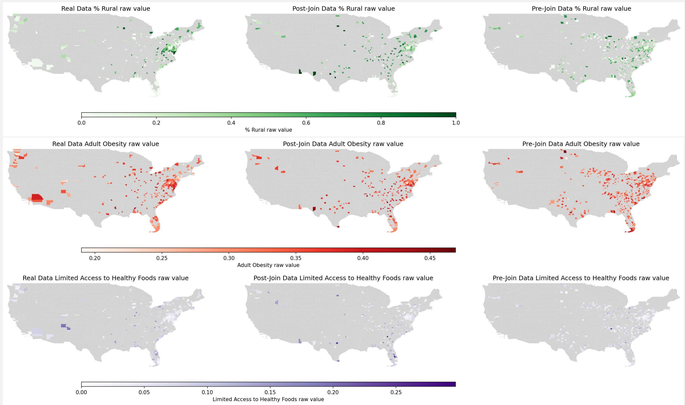
Results

The study analyzed ten variables from the clinical dataset and fifteen variables from the SDoH dataset. This poster presents only the representative variables due to space limitations.

Variables	Real	Post-Join	Pre-Join
Percentage of Female (sex_cd)	69.57%	68.48%	70.60%
Percentage of English (language)	97.68%	89.75%	85.82%
% Rural raw value	0.07 ± 0.13	0.08 ± 0.13	0.09 ± 0.15
Limited Access to Healthy Foods raw value	0.03 ± 0.02	0.03 ± 0.02	0.03 ± 0.02
Adult Obesity raw value	0.30 ± 0.06	0.29 ± 0.06	0.30 ± 0.06

Table 1. Summary Statistics for Clinical and SDoH Features Across Three Datasets. For clinical utilization variables, both synthetic datasets deviate from the real dataset. The pre-joined synthetic dataset maintains values that are closer to the real data across all clinical utilization measures. For SDoH variables, both synthetic datasets demonstrate a generally consistent alignment with the real data.

Figure 2. Heat Maps of the Three Datasets for Representative SDoH Variables. Colored areas indicate counties presented in the dataset, gray areas represent regions with no patients. The geographic distributions in the three datasets appear broadly similar but still exhibit noticeable differences. Since both the real dataset and the post-joined synthetic dataset are merged with real SDoH data, their heat maps exhibit identical density distributions for each variable. In contrast, the pre-joined synthetic dataset, which contains fully synthetic SDoH values, may display variations in density compared to the real data.



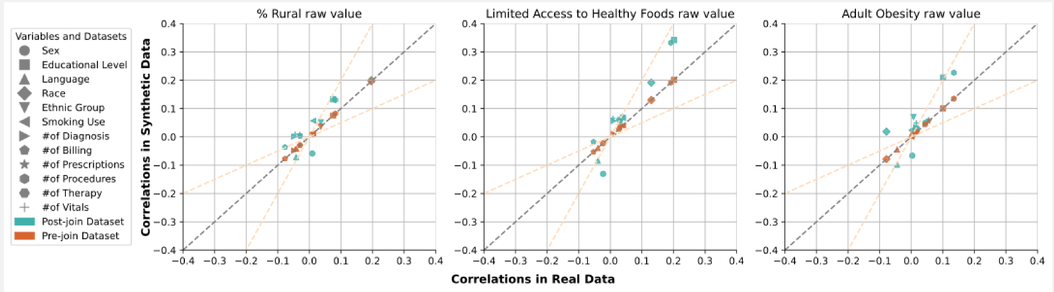


Figure 3. Correlation Comparisons Between the Real Dataset and the Post-Joined and Pre-Joined Synthetic Data

Data Source and Evaluation Matrix

Data & Variables

- Real Data: EHR dataset of 8,813 adult cancer patients from Georgetown University.
- SDoH Data: County-level features (e.g., income, food access, pollution, age) from the 2023 County Health Rankings & Roadmaps, joined via FIPS codes.
- Synthetic Model: Gretel.ai's Generative AI model.[3]

Evaluation Matrix:

- Descriptive Analysis: Comparison of summary statistics (e.g., percentages, means/SDs).
- Geospatial Analysis: Heat maps to visualize county-level SDoH distributions.
- Correlation Assessment: Scatter plots comparing clinical-SDoH correlation coefficients between real and synthetic datasets (vs. a slope-1 line).

Takeaways

- The Post-Joined approach (trained on clinical only) was better at replicating real SDoH distributions.
- The Pre-Joined approach (trained on clinical + SDoH) was better at modeling continuous feature correlations due to its richer feature context.
- Both methods successfully preserve general SDoH statistical characteristics but introduce geographic biases and struggle to maintain fidelity for near-zero correlations.

Future Opportunities:

- Enhancing generative AI models to address the biases in geographic representation.
- Evaluating performance using larger, more complex, and multi-institutional datasets to achieve broader and more balanced geographic coverage.

Funding

This research was, in part, funded by the National Institutes of Health (NIH) Agreement NO. 1OT2OD032581-01

Citations

1. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, Mooney SD, Malin BA. A multifaceted benchmarking of synthetic electronic health record generation models. Nature communications. 2022 Dec 9;13(1):7609.
2. Zhang Y, Dong JL, Xue B, Xiong Y, Gupta S, Van Segbroeck M, Shara N, McGarvey P. Exploring the Utilization of Synthetic Data in Unsupervised Clustering for Opioid Misuse Analysis. InAMIA Annual Symposium Proceedings 2025 May 22 (Vol. 2024, p. 1313).
3. Noruzman AH, Ghani NA, Zulkifli NS. Gretel. ai: Open-source artificial intelligence tool to generate new synthetic data. MALAYSIAN JOURNAL OF INNOVATION IN ENGINEERING AND APPLIED SOCIAL SCIENCES (MYJIEAS). 2021 Nov 30;1(01):15-22.