Advancing Medical Software: The Role of Synthetic Data in Developing Predictive Models

Cianna Grama

Thesis:

Researchers need to advance and implement synthetic data generation within the field of medical software development because of the potential to protect patient privacy, address data limitations, and improve the performance and scalability of predictive models.

What is Synthetic Data?

(GenRocket, 2024)

- New, fake data produced by controlled synthetic representations
- Generated by simulated situations - AI models apply sampling techniques to real data
- Types: fully, partial, hybrid
- Example: simulated set of patient medical records

Uses/Benefits of Synthetic Data

(Hashemi-Pour, et al., 2024)

- Training software models
- Augment small datasets
- Capture rare scenarios
- Cost-effective data testing
- Protect privacy
- Contains no identifiable information

What is Predictive Modeling?

- Advanced analytics to predict future events
- Uses data mining, machine learning, artificial intelligence
- Generates recommendations based on statistical trends
- Example: medical records, demographics, socioeconomic characteristics, used to identify patient susceptibility to illnesses

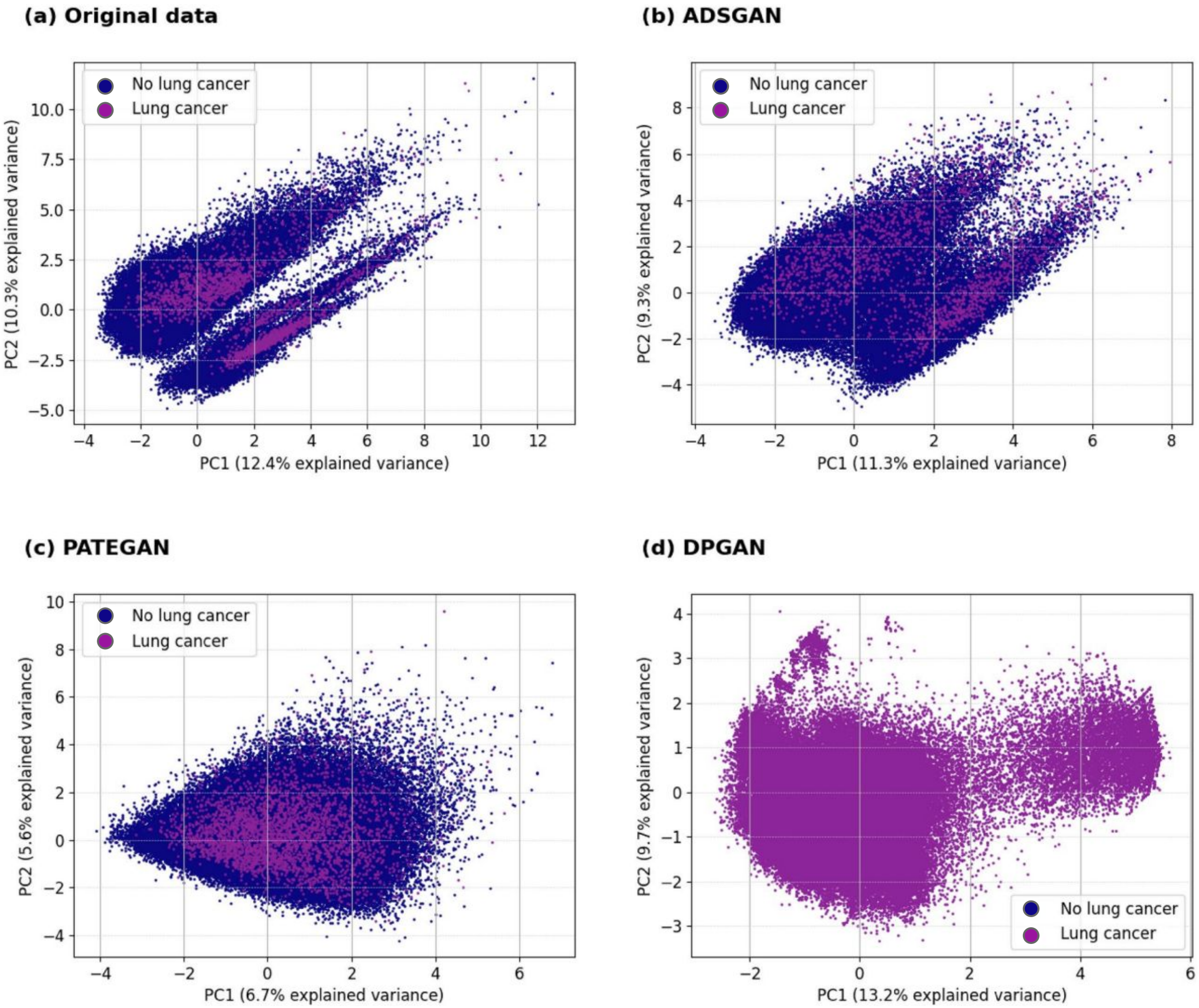
(Dunskiy, 2022)

Study: Synthetic data for privacy preserving clinical risk prediction

(Qian et. al,, 2024)

- Created 3 synthetic datasets based on UK biobank health data
- Synthetic datasets -> predictive models for lung cancer
- Synthetic data predictive models tested by accuracy compared to real data predictive models

Results:



(a) Original data

(b) ADSGAN

(c) PATEGAN

(d) DPGAN

Takeaways: Synthetic data for privacy preserving clinical risk prediction

(Qian et. al,, 2024)

- 2/3 predictive models produced reliable results
- Similar distributions, grouping, precision and recall
- Synthetic data outperformed models trained on oversampled real data
- Accuracy while maintaining privacy

Study: Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs)

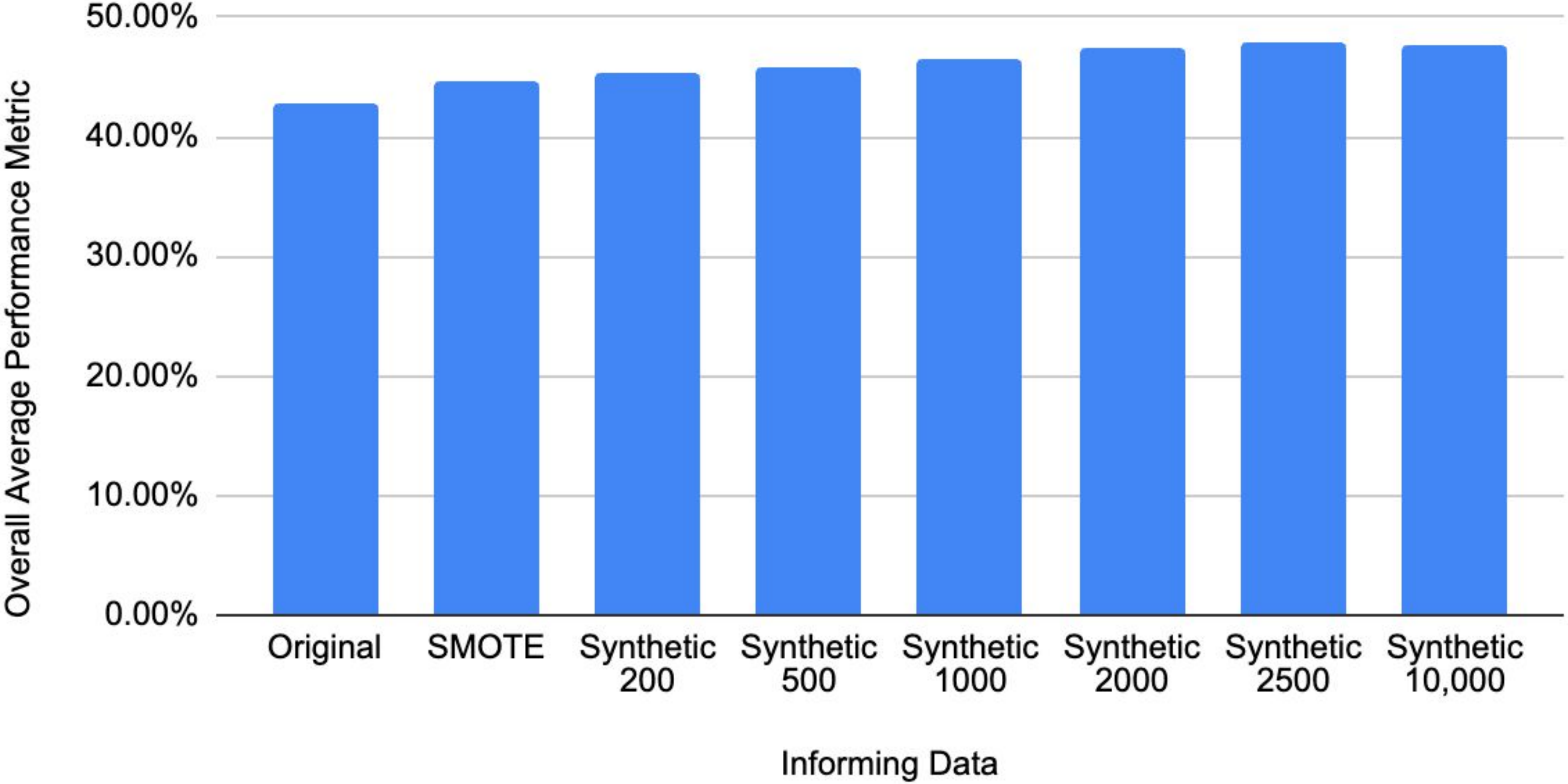
(Ghosheh et. al,, 2023)

- Generate synthetic data to fill in gaps -> train predictive models

- Data trained 3 machine learning modes on 3 different types of datasets

- Models evaluated by ability to predict if patient would develop an infection

- Models trained on original data and synthetic data compared



Average Performance of Models Using Different Synthetic Data Generations

Takeaways: Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs)

(Ghosheh et. al,, 2023)

- Models trained on most synthetic data outperformed the original dataset
- Synthetic data is viable solution to datasets with limitations caused by small, incomplete, imbalance sets
- Synthetic data enhances inclusivity of limited datasets

Counterargument

- Synthetic data may not fully capture rare, subtle, or complex patterns found in real-world healthcare data, leading to models that are less accurate or generalizable when applied in real clinical settings

- Not meant to completely replace real world data, meant to enhance data availability, improve privacy, and enable earlier-stage model development

Future Source

A large-scale comparative study that evaluates clinical outcomes of predictive models trained exclusively on synthetic data versus models trained on real patient data across multiple hospitals or healthcare systems

Discussion Questions

- Would you trust a healthcare tool or app if you knew it was trained mostly using synthetic patient data?

- Do you think using synthetic data to train healthcare models is a good idea if it protects patient privacy, even if it’s not 100% perfect?

References

GenRocket. “Adopting a New Synthetic Data Paradigm for Software Testing.” GenRocket Blog, 3 Oct. 2024, www.genrocket.com/blog/adopting-a-new-synthetic-data-paradigm-for-software-testing/.

Ghosheh, G. O., Thwaites, C. L., & Zhu, T. (2023). Synthesizing electronic health records for predictive models in low-middle-income countries (lmics). Biomedicines, 11(6), 1749. https://doi.org/10.3390/biomedicines11061749

Hashemi-Pour, C., Yasar, K., & Laskowski, N. (2024, December 26). What is synthetic data? examples, use cases and benefits: TechTarget. Search CIO. https://www.techtarget.com/searchcio/definition/synthetic-data

Predictive modeling in Healthcare: Benefits & Use Cases. Demigos. (n.d.). https://demigos.com/blog-post/predictive-modeling-in-healthcare/

Qian, Z., Callender, T., Cebere, B. et al. Synthetic data for privacy-preserving clinical risk prediction. Sci Rep 14, 25676 (2024). https://doi.org/10.1038/s41598-024-72894-y

Thank you!