Validating the Utility of Synthetic Data Generation for Clinical Research

Krikov S[1,2], Wilson A[1,2], Gregg M[1], Crockett DK[3]
[1]Parexel International, Waltham, MA, USA,
[2] University of Utah, Salt Lake City, UT, USA,
[3]Intermountain Health, Salt Lake City, UT, USA

Objectives

- **Clinical research often requires collaboration and data sharing.**
- Collaborative research can speed up research and improve findings, but using sensitive data like patient info raises privacy concerns (Figure 1). These challenges can negate any potential time savings and, in fact, be entirely prohibitive.
- One emerging solution to data sharing comes from the emerging field of synthetic data generation (SDG).
- The current study explores two promising SDG methods – an open-source method and a proprietary method - and evaluates them on a specific causal effect estimation task.

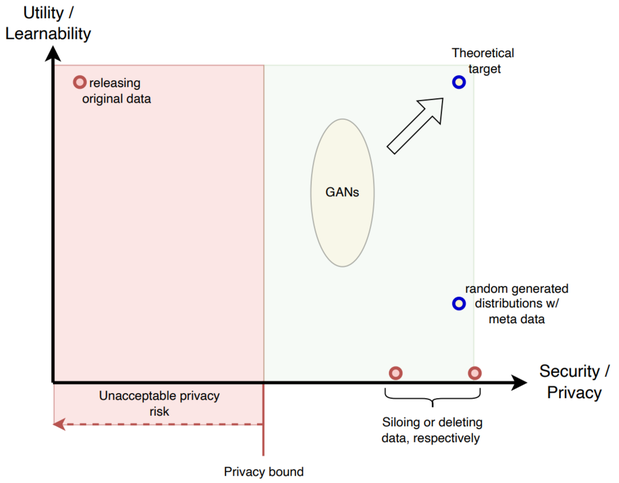

Figure 1. **Information retention – security/privacy** trade-off in synthetic data generation and sharing.

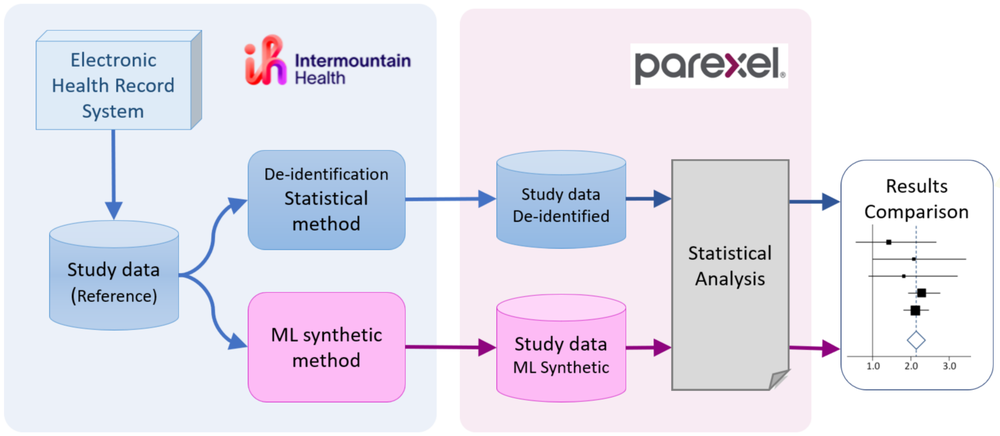

Figure 2a. Concept flow of data generation and analysis

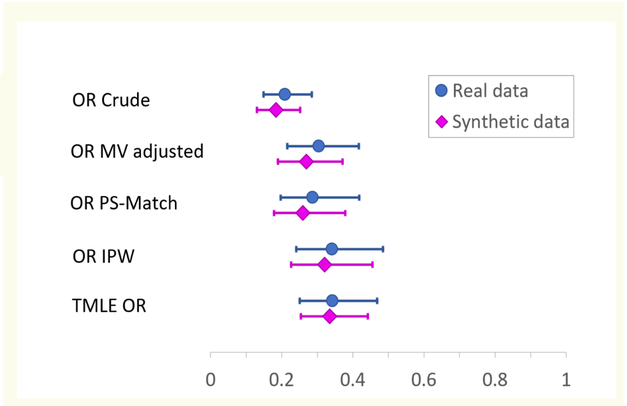

Figure 2b. Effect estimates indistinguishable using real and synthetic data

Methods

- In this study, we established an evaluation framework to assess synthetic data quality by comparing target causal effect estimates across different estimation methods (Figure 2a).
- Successful synthetic data was defined as preserving both effect relationships and confounding structures necessary for accurate causal inference.
- We estimated the target effect of medication exposure on death within 90 days using crude odds ratio, propensity score, and tmle-adjusted methods. We compared these estimates among the original and synthetic datasets.

Results

- **The results (Figure 2b) indicate that advanced SDG methods are successful in supporting accurate causal estimates by maintaining confounding structures in a kidney disease progression case study.**

Conclusions

- Synthetic data offers a pragmatic balance between data utility and privacy protection. It also enables broader data accessibility and collaboration while allowing for the inclusion of rare or underrepresented conditions in research, enhancing the scope and depth of studies.
- The effectiveness of synthetic data relies heavily on the selected generation method. Each method presents a trade-off between complexity, realism, and computational efficiency, influencing how closely the synthetic data mirrors the original dataset's information and relationships. As such, selecting the appropriate synthetic data generation technique is crucial for achieving accurate and meaningful research outcomes in clinical studies.