RCTs

ISPOR 30 YEARS
Thirty years of HEOR excellence.
A future of healthcare transformation.

MSR115

GENERATION OF SYNTHETIC PATIENT DATA TO OVERCOME MACHINE LEARNING LIMITATIONS IN HEALTHCARE RESEARCH: A SYSTEMATIC REVIEW OF METHODS, PERFORMANCE, AND USE CASES

AUTHORS

M Switat[1], T Porte[1], C Bouvard[1], N Sedmak[1], A Gougeon[1,2], F Roux[1], F Mistretta[1], A Lajoinie[1]

AFFILIATIONS

[1] **RCTs**, 38 rue du Plat, 69002 Lyon, France
[2] **LBBE** (Laboratoire de Biométrie et Biologie Evolutive) UMR 5558, CNRS, Université Lyon 1, Université de Lyon, Villeurbanne, France

CONTEXT AND OBJECTIVE

The generation of synthetic patient data is an emerging strategy to support the development of machine learning (ML) models in healthcare, particularly in settings with small sample sizes or imbalanced outcomes. By artificially augmenting datasets, it allows improving model robustness and performance. This systematic literature review provides an overview of current methods, identifies common challenges, and explores their real-world applications in healthcare ML.

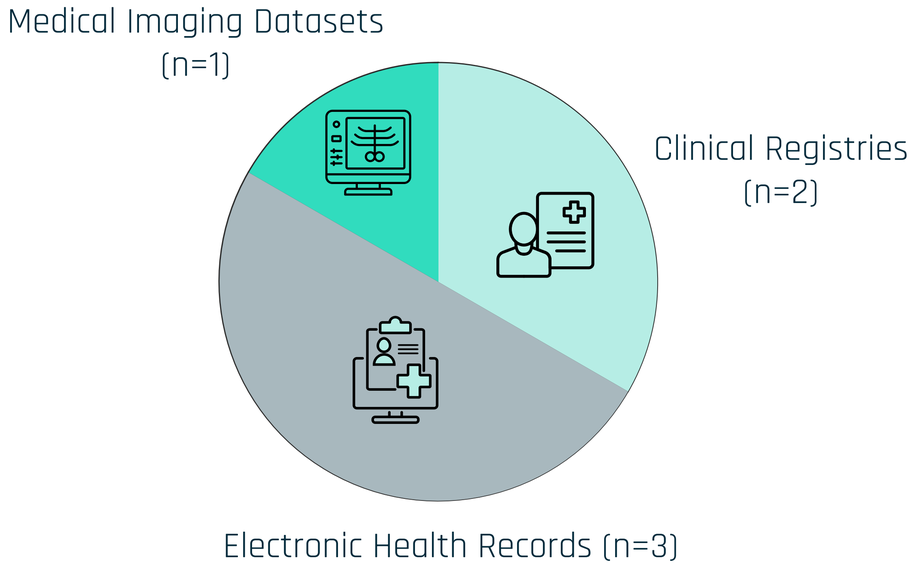
METHODOLOGY

A literature review was conducted on MEDLINE to identify studies published since 2020 on the generation of synthetic patient data for ML applications. Titles and abstracts [Ti/Abs] were screened, followed by full-text review for inclusion.

RESULTS/FINDINGS

A total of 176 studies were initially selected through title and abstract screening. After full-text review, 6 studies were included.

Distribution of Data Sources



Which models to generate synthetic data?

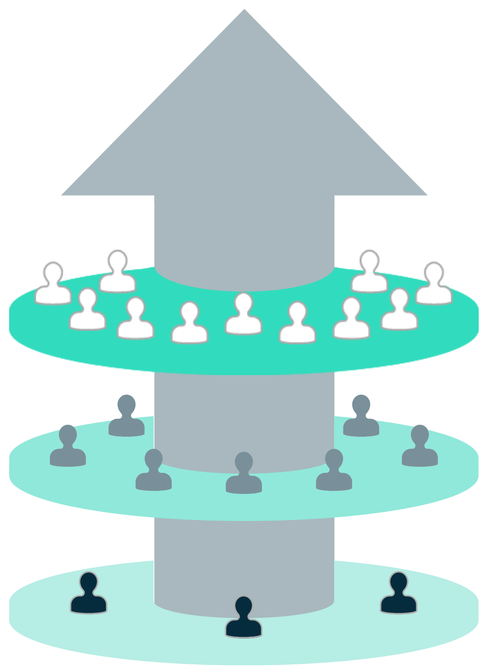
GANs (n=3)
Generative Adversarial Networks: neural network generating realistic data via adversarial training.

SMOTE (n=1)
Synthetic Minority Over-sampling Technique: synthetic sample created by interpolating minority class points.

CTGAN (n=1)
Conditional Tabular GAN: GAN model for tabular data with mixed types and conditional sampling.

Bayesian simulation (n=1)
Bayesian simulation: data generated by sampling from probabilistic models.

How synthetic data improved model performance?

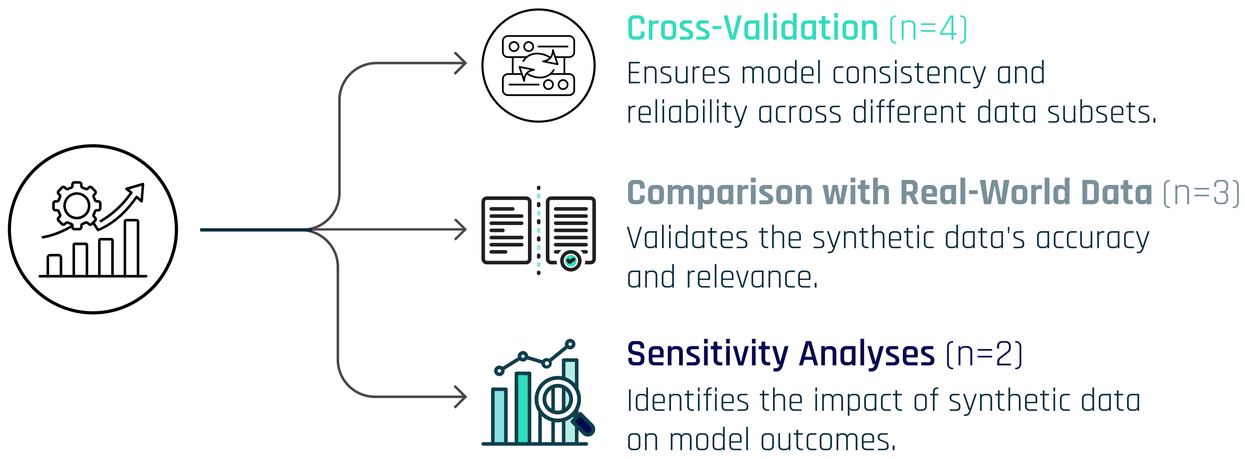


Synthetic Data Enrichment
Adding synthetic data to improve dataset quality

Improved Model Training
Training models with enriched datasets

Enhanced Model Evaluation
Evaluating models with improved data

How to assess the robustness of synthetic data ?



Cross-Validation (n=4)
Ensures model consistency and reliability across different data subsets.

Comparison with Real-World Data (n=3)
Validates the synthetic data's accuracy and relevance.

Sensitivity Analyses (n=2)
Identifies the impact of synthetic data on model outcomes.

CONCLUSION

Synthetic patient data generation is a promising strategy to improve the performance and robustness of machine learning models in healthcare. The reviewed studies show that synthetic data can effectively address data limitations while supporting privacy-preserving model development. Standardized evaluation frameworks and real-world implementation are needed to fully unlock its potential in clinical decision-making and health technology assessment.

REFERENCES

DuMont Schütte et al., 2021; García-Domínguez et al., 2023; Nelson et al., 2023; Rodriguez-Almeida et al., 2023; Scroggins et al., 2025; Son et al., 2023

CONTACT: Morgane Switat, PhD, msw@rcts.fr