N° MSR138

ISPOR

Evaluating The Potential Of Synthetic Patient Data Generation To Accelerate Real-World Evidence (RWE) Generation

Margaux Törnqvist[1], Louise Dry[1], Gaëtan Pinon[1], Antoine Movschin[1]
[1]Quinten Health – Quinten – France

INTRODUCTION

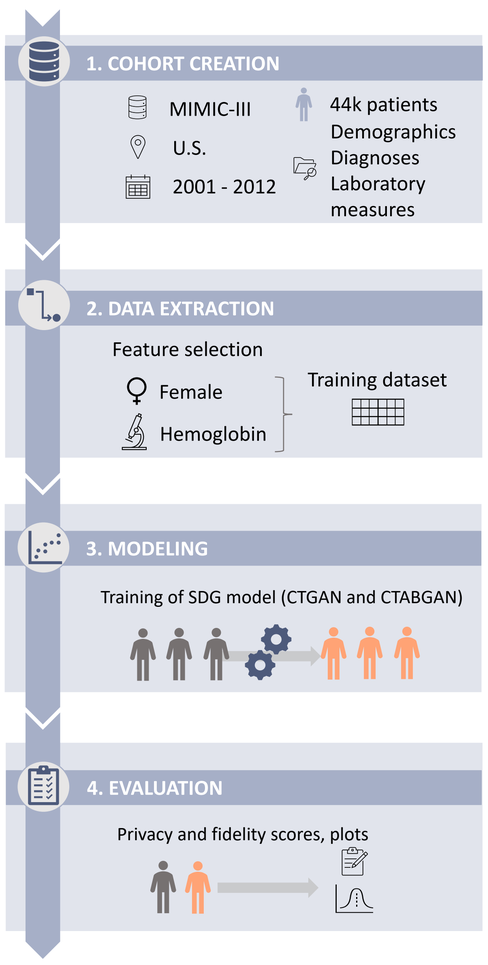
- The use of real-world data (RWD) has become pivotal in the development and evaluation of pharmaceuticals.
- Machine learning (ML) has shown potential in generating new evidence in medical research but requires access to high-quality, large-scale patient data.
- However, the use of such data is constrained by privacy concerns, high costs, limited availability, and accessibility barriers.
- To overcome these limitations, synthetic data generation (SDG) has emerged as an alternative.

OBJECTIVE

This work aims to evaluate the potential of SDG models, particularly a deep learning approach known as generative adversarial networks (GANs), to mimic tabular patient data by capturing and replicating its underlying statistical properties and patterns.

Method

Figure 1: Framework to evaluate SDG models



1. Cohort creation

The MIMIC-III [1] database, a publicly accessible collection of EHRs from intensive care, was selected for SDG. This dataset contains approximately 44k patients with demographic variables, diagnoses, and laboratory measures, recorded between 2001 and 2012 in the U.S.

2. Variables

To evaluate the models on various data types and patient profiles, both continuous features (e.g. hemoglobin, age at last admission) and discrete features (e.g. gender) were extracted, ensuring a minimum prevalence of over 10% within the database.

3. Modeling

This study explores the SDG of tabular data using two GAN models, CTGAN [2] and CTABGAN [3], both of which have shown efficiency in generating synthetic data across various non-healthcare domains [2] [3].

CTGAN handles well imbalanced data, while CTABGAN allows efficient modelling of both imbalanced or skewed data and enhanced management of mixed-type variables (e.g., both discrete and continuous variables) [2] [3].

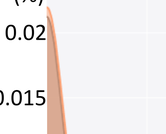
4. Evaluation

The quality of the generated synthetic data was assessed *via* fidelity metrics (similarity between datasets in terms of statistical distributions and correlation), privacy metrics (ability of the model to generate genuinely new data and to prevent data leaking) and visual comparisons (see table 1 and figures 2 and 3).

Results

- Both GAN models exhibited strong capabilities in generating realistic synthetic health data (see Table 1).
- Each model showed distinct strengths: CTABGAN excelled in generating data with high similarity in terms of distribution shape to the original data (see Figure 2) and preserving correlations (see Figure 3), while CTGAN outperformed in preserving data privacy.

Table 1: Evaluation metrics of SDG models

	Fidelity		Privacy	
Model	Distribution[1] 	Correlation[2]	New patients (%)	Distance score[3]
CTGAN	0.90	0.93	**100%**	**0.92**
CTABGAN	**0.94**	**0.96**	96%	0.85

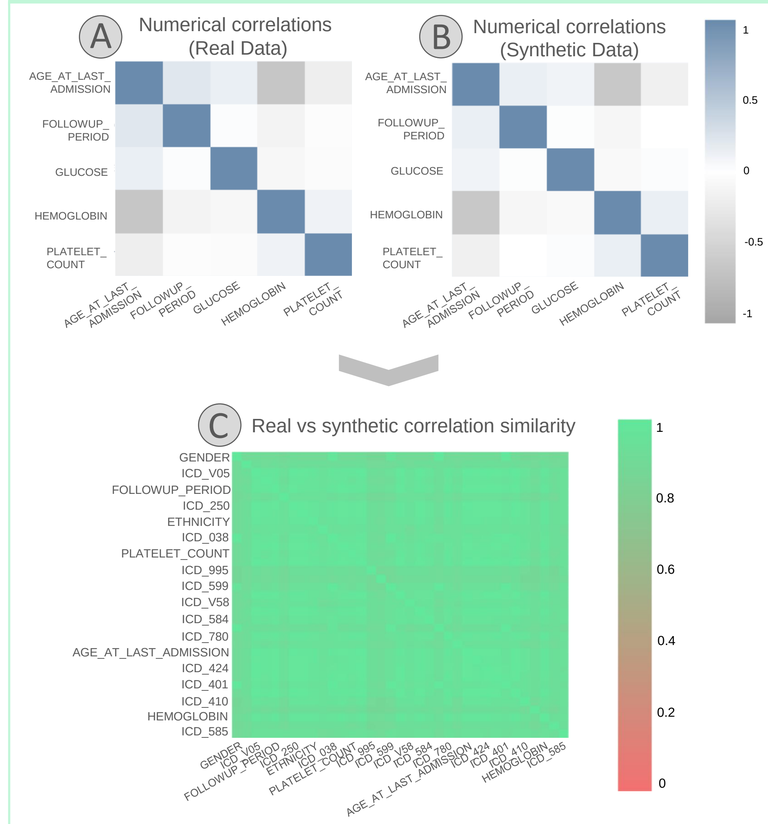
1. Distance shape score: average between KSComplement (1 minus the Kolmogorov-Smirnov statistic) and TVComplement (1 minus the Total Variation Distance), which measure the similarity between real and synthetic columns, for continuous and discrete features respectively.
2. Distribution correlation score: average between correlation and contingency similarities, which measure the similarity between real and synthetic data in terms of correlation and contingency tables, for numerical and categorial features respectively.
3. DCR: Distance to Closest Record (computed on 10% of the synthetic data for computational reasons): the minimal Euclidian distance between a synthetic and a real patient. Larger values indicate that synthetic data are not mere copies of the real data.

- As CTABGAN outperformed CTGAN in terms of fidelity, plots and scores are presented on the data generated by CTABGAN (see Figures 2 and 3).

Figure 2: Comparison of distribution shape plots of variables from real data and synthetic data

Distributions of both continuous and discrete features are very similar between the real and synthetic data.

Figure 3: Comparison of correlation plots between real and synthetic data from CTABGAN



Heatmaps A and B show the correlations between variables in the real and synthetic data. The heatmap C is built as 1 minus the absolute difference between heatmaps A and B. A value of 1 (perfect green) in heatmap C indicates perfect preservation of correlation during SDG.
Note: the heatmaps A and B presented here are extractions of the whole correlation matrices, for visualization purposes.

DISCUSSION

- GANs are complex models that require large datasets, significant computational resources, and can be challenging to train [4].
- The absence of standardized evaluation metrics for synthetic data poses a significant challenge to establish a robust assessment framework. This lack of universal criteria hinders the ability to reliably validate the quality and trustworthiness of the generated data, ultimately limiting their usability in critical applications [4].
- Moreover, the various available models possess unique properties that may be advantageous for different tasks [4], underscoring the importance of selecting the appropriate model depending on the available data and on the specific intended use of synthetic data.

CONCLUSION

Deep learning generative models offer a promising solution for synthesizing tabular health data. They address ~~the~~ challenges of accessing RWD while preserving their key characteristics. By preserving data privacy, balancing datasets, and providing more diverse training data for ML models, they may ~~can~~ accelerate real-world studies without compromising patient information.

REFERENCES

[1] Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. Sci. Data 3, 160035 (2016)

[2] Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. Modeling Tabular data using Conditional GAN. (2019) doi:10.48550/ARXIV.1907.00503

[3] Zhao, Z., Kunar, A., Van der Scheer, H., Birke, R. & Chen, L. Y. CTAB-GAN: Effective Table Data Synthesizing. (2021) doi:10.48550/ARXIV.2102.08369

[4] Chakraborty, T., Reddy K S, U., Naik, S. M., Panja, M. & Manvitha, B. Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art. Mach. Learn. Sci. Technol. 5, 011001 (2024)

The authors declare no conflict of interest.

quinten health