HIGH-FIDELITY SYNTHETIC PATIENT GENERATION AND VALIDATION FOR NSCLC USING TVAE AND OCTOPUS CLINICAL DATA

8471

Abir Tadmouri, MPH, PhD[1], Salma Barkaoui, PhD[2], Mohammed BENNANI[3], Jerome Vetillard, PhD, MD[2], Hadhami Mejbri, Master[2].
[1]Pierre Fabre, Boulogne Billancourt, France, [2]Qualees, Paris, France, [3]PhD, QUALEES, PARIS, France.

INTRODUCTION

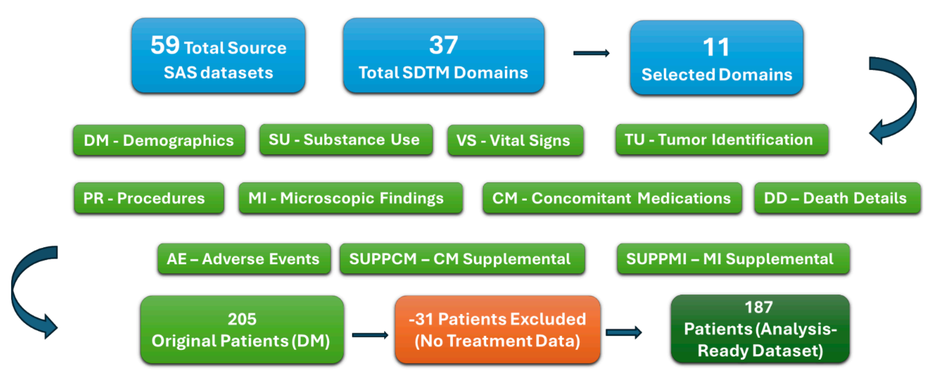
The generation of high-fidelity synthetic health data plays a crucial role in advancing disease prediction models and enabling data sharing while protecting patient privacy. However, traditional generative models often face limitations in capturing complex correlations and maintaining realistic data distributions within heterogeneous medical datasets. In oncology research, where data scarcity and patient heterogeneity pose major challenges—particularly for rare disease subtypes such as non-small cell lung cancer (NSCLC)—synthetic data generation offers a promising strategy to enhance statistical power and enable more comprehensive analyses. Despite numerous NSCLC studies, limited sample sizes and incomplete data often hinder reproducibility and generalizability.

OBJECTIVES

In this study, we aimed to enhance statistical power in evaluating oncology treatment efficacy by augmenting a real-world NSCLC dataset with high-quality synthetic data. Using the OCTOPUS study cohort (~200 patients), we applied advanced generative AI methods to expand the dataset while preserving its clinical and statistical integrity. We first describe the NSCLC cohort and the preprocessing steps used to improve data quality, followed by the TVAE method and a validation pipeline designed to assess synthetic data quality. Evaluation metrics included distributional similarity, correlation preservation, and machine learning utility. Finally, we present key results demonstrating the effectiveness of synthetic data augmentation for NSCLC research.

METHODS

OCTOPUS Data Description and Preprocessing



Data Integration Strategy

1. Central Identifier and Reference Table
- Identified USUBJID as the unique patient ID.
- Used the Demographics (DM) table as the central hub for merging.

2. Table Selection and Exploration
Selected key SDTM domains relevant for survival modeling:
- **DM –** Demographics
- **CM –** Concomitant Medications
- **MI –** Microscopic Findings
- **DD –** Death Details
- **SU –** The use of Substances
- **PR –** Procedures
- **TU –** Tumor Identification
- **VS –** Vital Signs
- **+ SUPP –** Tables (e.g., SUPPMI, SUPPCM)
- **AE –** Adverse Effects

3. Supplementary Table Integration
- Joined SUPP tables using USUBJID, IDVAR, IDVAL.
- Pivoted and merged supplementary variables into the corresponding main domains.

Data Processing Flow

1. Treatment Flow Mapping
Per patient, we mapped the treatment trajectory in the following order:
- Systemic treatment
- First-line Systemic treatment
- Second-line Systemic treatment
- Subsequent Systemic treatments

2. Treatment Outcome Categorization
Patients were classified into four treatment outcome groups:
- Completed treatment (positive reason)
- Discontinued due to death
- Discontinued by physician (negative reason)
- Other treatment interruptions

3. Missing Data Strategy
- Identified meaningful vs. structured missing data.
- Used rule-based + MICE hybrid imputation.
- MICE showed the best longitudinal coherence compared to KNN and SoftImpute.
- Verified statistical coherence of imputed values.

4. Dataset Final Composition:
- Baseline demographics and clinical features
- Treatment timelines and status
- Survival and progression indicators
- Harmonized variables across domains
- Cleaned and imputed values for robust downstream modeling

Tabular Variational AutoEncoder (TVAE)

In this study, we enhanced TVAE by optimizing its loss function to better preserve feature relationships. TVAE is a generative model tailored for tabular data.

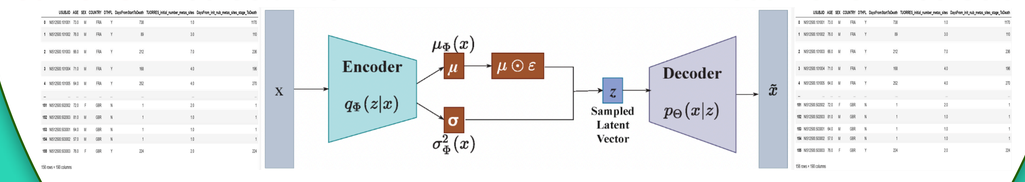
TVAE uses a variational autoencoder architecture to learn the underlying distribution of input features and generate new synthetic samples.

Preprocessed Real Data

156 Rows X 190 Columns

Synthetic Data Generation

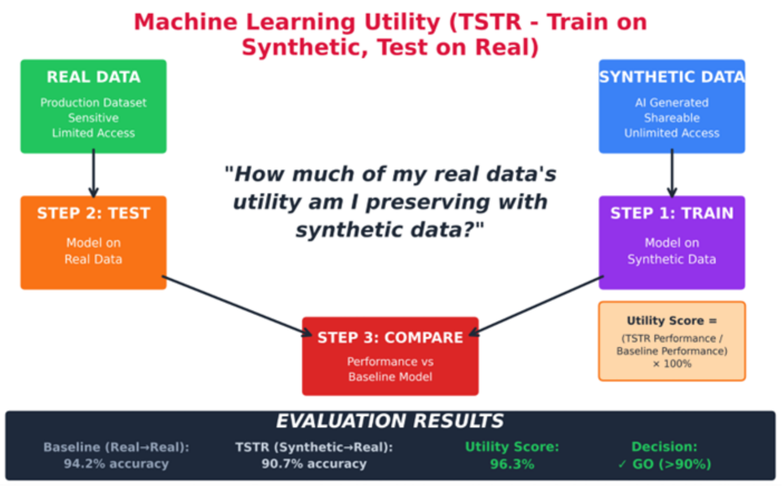
500 Rows X 190 Columns



TVAE

RESULTS

Synthetic Data Validation Metrics



Machine learning (ML) utility refers to a predictive model comparison metric that evaluates whether models trained on synthetic data achieve performance comparable to those trained on real data. This metric addresses the practical question:

Can synthetic data be used to train effective machine learning models?

Statistical Tests

Kolmogorov-Smirnov (KS) Test: Tests whether two samples come from the same continuous distribution by comparing their cumulative distribution functions (CDFs) between real and synthetic data. *Healthcare: Critical for ensuring synthetic patient measurements (e.g., blood pressure, lab values, treatment durations) follow realistic distributions.*

Chi-Square (χ^2) Test: Tests independence and goodness of fit for categorical variables. *Healthcare: Essential for categorical variables such as diagnosis codes, treatment types, patient demographics, and medication classes.*

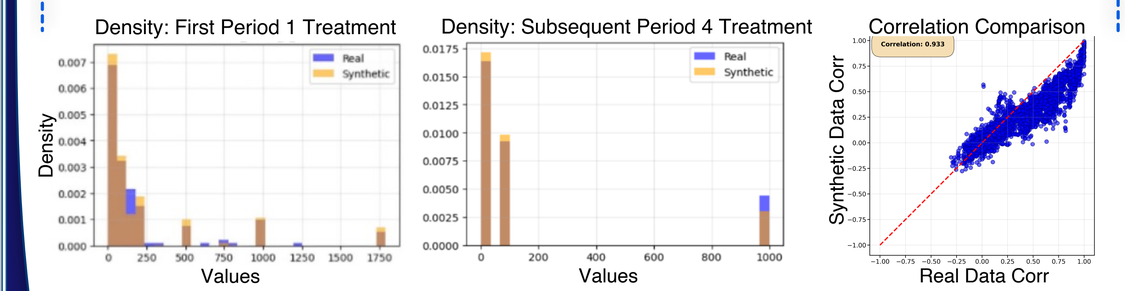
Entropy Comparison: Measures the information content and uncertainty in data distributions. *Healthcare:* Ensures synthetic data maintains the complexity and variability of real patient populations.

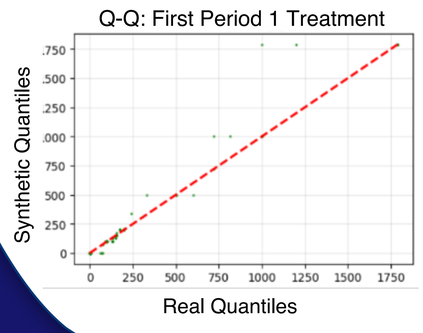
ANOVA (Analysis of Variance) Test: Tests whether the means of multiple groups are statistically different.

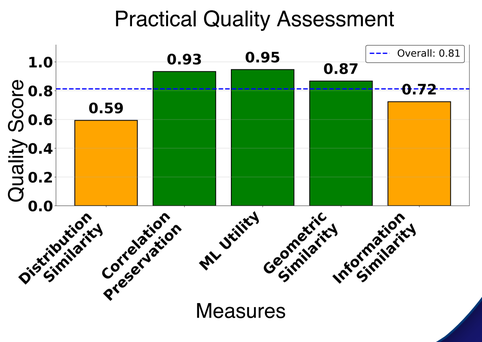
Mutual Information: Measures statistical dependence between variables, capturing both linear and non-linear relationships. *Healthcare: Critical for preserving complex medical relationships, such as comorbidity patterns, drug interactions, and disease progression dependencies.*

Correlation Preservation: Measures how well the relationships between features in the synthetic data match those in the real data.

Validation Metrics Results







CONCLUSION

This study demonstrates a robust, validated pipeline for generating high-fidelity synthetic NSCLC patient data from structured clinical datasets. The synthetic OCTOPUS cohort of 500 patients achieved excellent quality scores across all validation tiers, with an overall practical score of 0.802. Notably, the synthetic data demonstrated outstanding performance in ML utility (0.907, marked as the most important metric), correlation preservation (0.935), and geometric similarity (0.844), while maintaining good distribution similarity (KS: 0.604) and information similarity (KL: 0.721).

These results confirm that the synthetic dataset successfully preserves complex clinical relationships and predictive signals essential for downstream research applications. The validated synthetic data enables advanced clinical research, including predictive modeling, treatment optimization, and hypothesis generation with statistically reliable outcomes. This scalable framework opens promising avenues for patient-specific survival risk prediction and personalized treatment selection based on individual patient trajectories, advancing precision oncology research with a high-fidelity dataset ready for production use.

REFERENCES

D'Amico, S., et al. (2023). Synthetic data generation by artificial intelligence to support clinical trials in oncology. JCO Clinical Cancer Informatics. https://doi.org/10.1200/CCI.23.00021

Gonzalez-Abril, L., Angulo, C., Ortega, J. A., & Lopez-Guerra, J.-L. (2022). Statistical validation of synthetic data for lung cancer patients generated by using generative adversarial networks. Electronics, 11(20), 3277. https://doi.org/10.3390/electronics11203277

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114. https://arxiv.org/abs/1312.6114

Krishnaswamy, D., Bontempi, D., Thiriveedhi, V., Punzo, D., Clunie, D., Bridge, C. P., Aerts, H. J. W. L., Kikinis, R., & Fedorov, A. (2023). Enrichment of the NLST and NSCLC-Radiomics computed tomography collections with AI-derived annotations. arXiv. https://arxiv.org/abs/2306.00150

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using a conditional GAN. Advances in Neural Information Processing Systems, 32, 11990–12000. https://proceedings.neurips.cc/paper/2019/hash/5e1f2d734e6aebd2d0f90d33f6c3d2d2-Abstract.html

Yoon, J., Jordon, J., & van der Schaar, M. (2019). TVAE: Tabular variational autoencoder for tabular data. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence. https://ojs.aaai.org/index.php/AAAI/article/view/4404