ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

Evaluation of Artificial Intelligence-Generated Synthetic Data for Clinical Research in Secondary Cardiovascular Prevention of Patients with Dyslipidemia

Bonomi Alice[1], Saccani Sebastiano[2], Coser Andrea[2], Valsecchi Carola[3], Galotta Arianna[1], Scatigna Marco[1], Werba Pablo[1]

(1) Centro Cardiologico Monzino IRCCS, Milan
(2) Aindo Spa, Trieste
(3) Università degli Studi di Milano-Bicocca, Milan

CORRESPONDING AUTHOR: *Bonomi Alice, abonomi@cardiologicomonzino.it*

INTRODUCTION

Data is crucial in modern healthcare, bearing the potential to improve patient care by powering clinical research and enhancing public health. However, the promise of real-world data to enable personalized medicine, guide policymaking, and respond to rapidly evolving healthcare needs is often constrained by challenges in accessing high-quality datasets. Synthetic data offers a compelling alternative, addressing privacy concerns, simplifying ethics reviews, reducing costs, and ensuring access to sufficiently large and reliable patient cohorts [1]. While synthetic data has already been successfully validated in domains outside healthcare, its application in the medical field remains limited.

OBJECTIVES

This study aims to compare synthetic and real-world data in healthcare by applying various statistical methodologies to both types of datasets.

METHODS

Data synthesis techniques will be used to create cohorts of patients with specific attributes that are statistically similar to real patients by using AI tools released by by Aindo SpA. The real datasets originate from the clinical platform of Centro Cardiologico Monzino and include structured data systematically collected by the Atherosclerosis Prevention Unit from 2002 to 2024 during outpatient visits of 1000 patients in secondary cardiovascular prevention and with a personal history of dyslipidemia. A comparison was performed between real and synthetic datasets. For categorical variables (n = 49), Jensen–Shannon Divergence (JSD) was used. For continuous variables (n = 25), differences in means were evaluated using 95% confidence intervals (CIs). A logistic regression model with stepwise selection and cross-validation was developed using both real and synthetic data.

RESULTS

Only 1 out of 49 categorical variables showed a statistically significant difference (~2%), which is below the expected 5% due to type I error. For continuous variables, 4 out of 25 (16%) showed CIs that did not include zero. Logistic regression with cross-validation identified the same predictors in both datasets: CPK (UI/l) and Therapy modification. Odds Ratio (OR) for CPK was 1.01 (real) and 1.01 (synthetic), with a 70% overlap of CIs. OR for Therapy modification was 0.10 (real) and 0.28 (synthetic). The model developed on the synthetic database was used for training and then validated on the real database. All variables selected by the stepwise model on synthetic data were validated on real data, confirming the model's transferability.

DOI: 10.54103/2282-0930/29350

CONCLUSIONS

The synthetic dataset demonstrated reliability and comparability with real-world data. While having a slightly higher margin of error, cross-validation of SAMS (statin-associated muscle symptoms) predictors indicates that models trained on synthetic data can be transferred effectively to real-world data applications, supporting wider use of synthetic datasets in clinical and epidemiological studies.

This study has been realized thanks to the support of NOVARTIS.

REFERENCES

1. Echo Wang, Katrina Mott, Hongtao Zhang et al. Validation Assessment of Privacy-Preserving Synthetic Electronic Health Record Data: Comparison of Original Versus Synthetic Data on Real-World COVID-19 Vaccine Effectiveness. Pharmacoepidemiology and Drug Safety, 2024; 33:e70019.

DOI: 10.54103/2282-0930/29350