Synthetic Data Generation Methods using AI: A New AI tool that generates synthetic data using Generative Adversarial Network (CTGAN), Copula GANs and Sequential Decision Tree Methods.

Developed by Regulatory Scientific and Health Solutions Ltd (R-S-S)

R-S-S Regulatory Scientific and Health Solutions

Background & Introduction

There is an increasing interest in synthetically generated data (SGD). Reasons for this include earlier decision making, oversight of analyses undertaken, predicting future trends, augmenting small sample sizes in rare diseases and avoiding sharing of company data assets while complying with data privacy regulations. Moreover, guidelines from the joint clinical assessment (JCA) may require use of complete patient level data availability for indirect treatment comparisons (ITCs) for the Health Technology Assessment (HTA) process. ECLIPTICA® is a powerful Artificial Intelligence (AI) agent that generates SGD simply, using state of the art AI techniques.

Methods

ECLIPTICA® employs multiple data synthesis models: Conditional Tabular Generative Adversarial Network (CTGAN), a Gaussian Copula (GCP), a Copula GAN synthesizer (CGS) and Tabular Variational Autoencoder (TVAE) and Sequential Decision Trees (SDT)[1-8]. For the GANs, a dual-network adversarial architecture comprising a generator and a discriminator was used.

The performance of these methods were evaluated using three case studies (datasets): (i) Vital Signs data at a single timepoint (N=500); (ii) Data from Non-Small Cell Lung Cancer (NSCLC) patients in a randomized phase III trial (N=670)[9]; (iii) longitudinal EQ-5D-5L utility data from a randomized controlled trial (RCT) in ophthalmology (N=125).

ECLIPTICA®

SYNTHETIC DATA GENERATION

UNLOCKING THE POWER OF SYNTHETIC DATA USING AI

- Data Privacy & Compliance
- Data Accessibility
- Cost Efficiency & Speed
- Bias Reduction
- Rare Disease Modelling & Trial Simulation
- Data Augmentation
- Regulatory Testing & Validation
- Full Data Security & Risk Mitigation

Results

Case Study 1 : Vital Signs Data

Table 1: Results from Synthetically Generated Vital Signs dataset using CTGAN model

Variable	Original Data (N=500)	Synthetic Data (N=500)
Age (Years)		
n	500	500
Mean (SD)	44.20 (8.22)	45.71 (7.86)
Median	44	45
Min, Max	29.00, 60.00	29.00, 60.00
Sex, n(%)		
Female	301 (60.2%)	271 (54.2%)
Male	199 (39.8%)	229 (45.8%)
Weight (lb)		
n	499	489
Mean (SD)	152.08 (27.80)	150.98 (24.77)
Median	148	150
Min, Max	91.00, 247.00	98.00, 247.00
Diastolic BP (mmHG)		
n	500	500
Mean (SD)	84.39 (12.93)	87.05 (13.0)
Median	82	86
Min, Max	58.00, 150.00	58.00, 150.00
Systolic BP (mmHG)		
n	500	500
Mean (SD)	138.72 (23.45)	138.68 (28.13)
Median	134	131.5
Min, Max	98.00, 272.00	98.00, 272.00

Abbreviations: BP: Blood Pressure; mmHG: millimeters of mercury; Min: Minimum; Max: Maximum; SD: Standard Deviation.

When tested on the vital signs dataset, the CTGAN model was able to generate synthetic data closely aligned with the original data, as demonstrated by the summary statistics shown in Table 1.

Case Study 2: NSCLC RCT Data

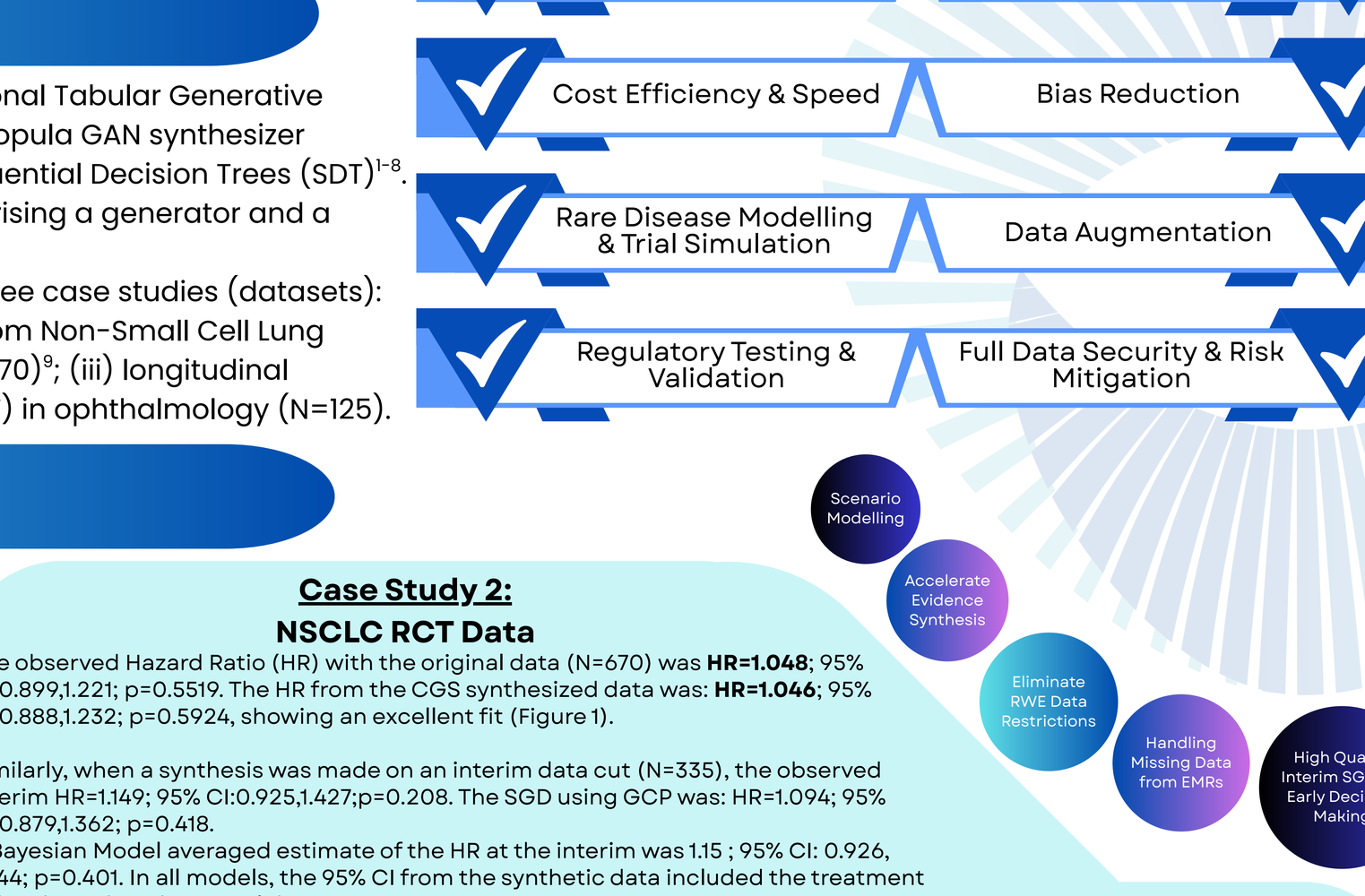
The observed Hazard Ratio (HR) with the original data (N=670) was **HR=1.048**; 95% CI:0.899,1.221; p=0.5519. The HR from the CGS synthesized data was: **HR=1.046**; 95% CI:0.888,1.232; p=0.5924, showing an excellent fit (Figure 1).

Similarly, when a synthesis was made on an interim data cut (N=335), the observed interim HR=1.149; 95% CI:0.925,1.427;p=0.208. The SGD using GCP was: HR=1.094; 95% CI:0.879,1.362; p=0.418.

A Bayesian Model averaged estimate of the HR at the interim was 1.15 ; 95% CI: 0.926, 1.444; p=0.401. In all models, the 95% CI from the synthetic data included the treatment effect based on the actual data.

- Scenario Modelling
- Accelerate Evidence Synthesis
- Eliminate RWE Data Restrictions
- Handling Missing Data from EMRs
- High Quality Interim SGD for Early Decision Making

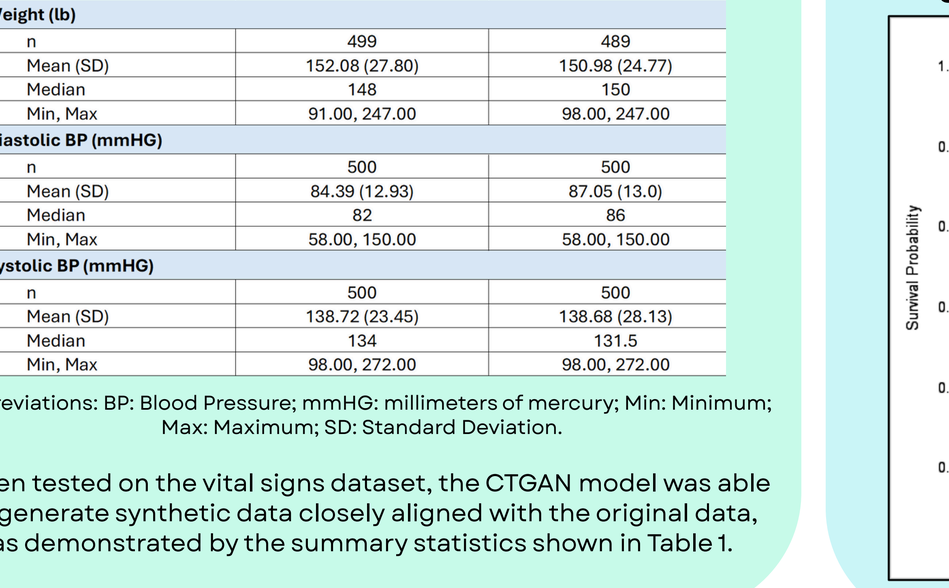
Figure 1: Synthetically generated Hazard Ratio (HR) from the TOPICAL RCT using CGS Model



*Model Average HR across all SDG Models: 1.013 ; 95% CI: 0.871, 1.181; p=0.710

Case Study 3: Opthalmology RCT Data

Figure 2: Synthetically generated Longitudinal Data using CGS model (EQ-5D-5L Utilities)



Note: EQ-5D-5L utilities estimated using the UK Alava Value set

When tested on longitudinal data (Figure 2), the CGS model produced EQ-5D-5L utility values aligned with the original data at baseline and months 12 and 24 for both the treatment and control groups.

Conclusions

- ECLIPTICA® generated high quality synthetic data from three different datasets in different disease settings. This was true for the full data set as well as for interim data cuts.
- ECLIPTICA® can be used to generate synthetic data from real world databases allowing consumers a mechanism for quality control and checks of vendor analyses.
- ECLIPTICA® can be used to predict future trends from the interim data allowing for earlier decision making and minimizing the risks of investment decisions for clinical trial, real world data and health technology (HTA) decision making.
- With ECLIPTICA®, sponsors need not give out their data assets to 3rd parties unless there is value in doing so – synthetic data can be shared instead.
- The new JCA requirement from EMA mandates the use of RWD for indirect treatment comparisons (ITCs) where possible. This may require sponsors to share their data assets to 3rd parties. Sponsors may not always wish to share highly sensitive data with 3rd parties. ECLIPTICA® offers a solution.

SCAN ME

References:

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS).
2. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular Data using Conditional GAN (CTGAN). arXiv:1907.00503.
3. Nelsen, R. B. (2006). An Introduction to Copulas (2nd ed.). Springer.
4. Torfi, A., Iranmanesh, S. M., Keneshloo, Y., Reddy, C. K., & Liu, N. (2022). Generating Synthetic Tabular Data via Diffusion Models. arXiv:2207.00068.
5. Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating Synthetic Data with Differential Privacy. ICLR 2019 Workshop.
6. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data Synthesis Based on Generative Adversarial Networks. arXiv:1806.03384.
7. Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2022). CTAB-GAN+: Enhancing Tabular Data Synthesis. arXiv:2204.00401.
8. Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-Series Generative Adversarial Networks (TimeGAN). NeurIPS 2019.
9. Lee SM, Khan I, Upadhyay S, Lewanski C, Falk S, Skailes G, Marshall E, Woll PJ, Hatton M, Lal R, Jones R, Toy E, Chao D, Middleton G, Bulley S, Ngai Y, Rudd R, Hackshaw A, Boshoff C. First-line erlotinib in patients with advanced non-small-cell lung cancer unsuitable for chemotherapy (TOPICAL): a double-blind, placebo-controlled, phase 3 trial. Lancet Oncol. 2012 Nov;13(11):1161-70. doi: 10.1016/S1470-2045(12)70412-6. Epub 2012 Oct 16. PMID: 23078958; PMCID: PMC3488187.

ecliptica@r-s-s.com

www.r-s-s.com/ecliptica