Cleveland Clinic Innovations

PRODUCT
Fully automated platform for generating and validating synthetic tabular data.

INDICATION
Reduces barriers to research using healthcare data.

VALUE PROPOSITION
- User-friendly interface streamlines synthetic data production and compliance.
- The model can increase the data generated between 2-5x.
- Scalability reduces time, cost and risk.

DEVELOPMENT STAGE
STNG is currently being used in four (4) active IRB studies at Cleveland Clinic.

INTELLECTUAL PROPERTY
Patent application pending.

RELATED PUBLICATIONS
Rashidi, H. H. et al. Prediction of Tuberculosis Using an Automated Machine Learning Platform for Models Trained on Synthetic Data. J Pathol Inform 13, 10 (2022).

CONTACT INFORMATION
Jerry Wilmink, PhD
Director of Business Development & Licensing
wilminj@ccf.org
216.314.6397
CCF ref: IDF 2022-185

Synthetic Tabular Neural Generator (STNG)

Inventors: Hooman Rashidi, MD, MS, Samer Albahra, MD, Bo Hu, PhD, Brian Rubin, MD, PhD - *Cleveland Clinic Pathology and Laboratory Medicine (PLMI), PLMI's Center for Artificial Intelligence & Data Science*

UNMET NEED
Synthetic data generation is a technique that involves creating artificial data that resembles real data but doesn't compromise privacy. It has emerged as a solution to address the challenges of traversing internal policies and procedures governing the use of protected health information datasets. Synthetic data can maintain the collective relationships of their real data counterparts. They can be made immediately available in a complete, non-biased, representative dataset for discovery and innovation. Existing platforms for generating synthetic data have limited utility because they require extensive technical knowledge, machine learning (ML) and statistical expertise. They do not always represent real data performance when validated.

SOLUTION
STNG is a fully automated enabling platform that allows users to quickly generate and evaluate tabular synthetic data from original reference datasets. The synthetically generated data maintains the collective relationships of real data counterparts in complete, non-biased, representative datasets for discovery and innovation. STNG's ability to rapidly optimize and maximize synthetic data production is invaluable in grant funding, speed-to-discovery, intellectual property, licensing, collaborative research, and publications.

- STNG's multi-function generator creates eight (8) viable synthetic datasets, each validated within a separate Auto-ML validation process incorporating scaling, feature selection, and hyper-parameter tuning.
- STNG outperformed generic synthetic data generators for eight (8) of the nine (9) binary dataset studies.
- Twelve (12) datasets were used to validate STNG's performance. The number of features (independent variables) varied from 6 to 98, and the sample sizes ranged from 280 to 13,611.