VIRTUAL PATIENT DATA GENERATOR: SYNTHETIC ECMO DATA FOR ENHANCED MEDICAL DEVICE DEVELOPMENT

Micha Landoll (1,2,3), Yifei Huang (1), Ulrich Steinseifer (1), Stephan Strassmann (3), Christian Karagiannidis (3), Michael Neidlin (1)

1. Department of Cardiovascular Engineering, Institute of Applied Medical Engineering, Helmholtz Institute, RWTH Aachen University, Germany;
2. ARDS and ECMO Centre Cologne-Merheim, Dept. of Pneumology and Critical Care Medicine, Kliniken der Stadt Köln gGmbH, Witten/Herdecke University, Germany;
3. Germany Institute for Computational Biomedicine II, University Hospital Aachen, RWTH Aachen University, Germany

Introduction

In silico clinical trials for Extracorporeal Membrane Oxygenation (ECMO) devices, by offering risk-free testing environments, could significantly enhance medical device development. A reliable generation of virtual patient cohorts from clinical data is the first step in such in silico clinical trials [1,2]. The challenge of integrating complex patient data, however, limits these models' predictive power [3]. This study aims to overcome current limitations such as patient privacy concerns and inadequate dataset sizes by generating high quality synthetic patient data from Electronic Health Records (EHR) of ECMO therapy.

Methods

This study employs a Conditional Tabular Generative Adversarial Network (CTGAN) to generate synthetic data from the Electronic Health Records (EHR) of 767 patients during ECMO. It includes time-series data across 59 selected therapy parameters, such as vital signs, blood gases, and organ function indicators. A Conditional Tabular Generative Adversarial Network (CTGAN) was employed to synthesize data (Figure 1), incorporating steps such as data preprocessing, imputation, and hyperparameter tuning. Through its generative adversarial process, this model refines synthetic data to closely mimic real patient records.

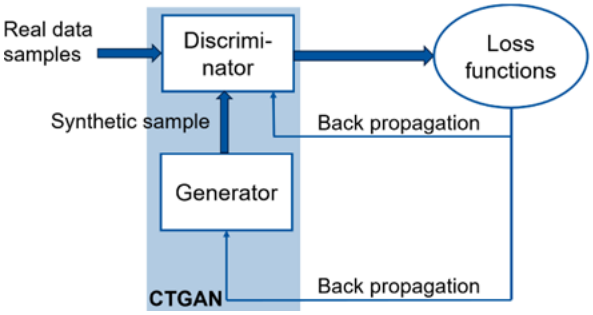

Figure 1: CTGAN Model Architecture with two components: a generator creating synthetic data and a discriminator evaluating its authenticity, working in tandem to produce high-fidelity synthetic data.

Results

The synthetic data showed strong alignment with the original dataset, evidenced by coverage scores averaging 95% (min. 58%) and boundary and synthesis scores of 100%. The mean absolute differences in correlations coefficients between the synthetic and original data were minimal, averaging 5.9% with a maximum deviation of 44.7% (exemplary data subset in figure 2).

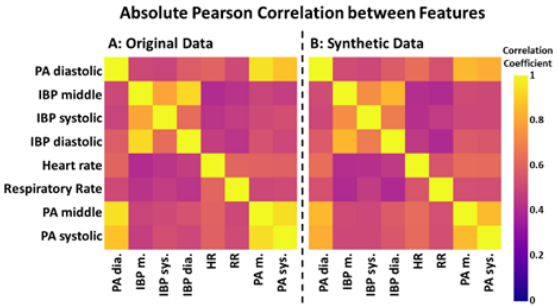

Figure 2: Comparison on exemplary data subset of absolute Pearson correlation coefficients between original (A) and synthetic (B) datasets to evaluate synthetic data generation precision and mirror inter-variable relationships.

Discussion

This synthetic data accurately reflects the original dataset, demonstrating the method's potential to enhance in silico ECMO trials, crucial for patient safety. While scaling to higher dimensional datasets poses challenges, these findings lay a solid foundation for future development. Addressing scalability will further enhance this research's impact, potentially shortening ECMO device development cycles, increasing patient safety and advancing clinical practice. Alongside future explorations in synthetic data benchmarking and patient trajectory modelling, the development of an online platform for virtual patient data generation based on EHR could enhance accessibility to patient data and drive innovation towards in silico clinical trials in ECMO research.

References

1. Simalatsar A, Front. Big Data, 6:1085571, 2023.
2. Van Breugel B et al., arXiv:2304.03722, 2023.
3. Chen RJ et al., Nat Biomed Eng, 5:493–497, 2021.