

Synthetic Alzheimer's Disease Patient Generation to Improve Clinical Trial Representativeness in The All-Embraced Framework

Alberto Boccali¹, Claudio Crema¹, Claudio Demaria¹, Matteo Lupini², Ilaria Passeggia², Monica Almici², Cesare Michele Baronio¹, Silvia De Francesco¹, Damiano Archetti¹, Alberto Redolfi¹

1. Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy

2. Laboratory Alzheimer's Neuroimaging & Epidemiology, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy

INTRODUCTION

Clinical trials (CTs) results are frequently affected by suboptimal study designs or ineffective recruitment strategies, leading to discrepancies between enrolled participants and the intended target populations. The ALL-EMBRACED project (www.allembraced.eu) seeks to enhance inclusiveness and transparency in CTs through data-driven and AI-based methodologies. In this context, we investigated the application of synthetic data generation for Alzheimer's disease (AD) using the open-source **Synthea** platform (Walonoski et al., 2018). The objective was to generate realistic synthetic cohorts with pathological markers across different disease stages, suitable for assessing how eligibility criteria relaxation impacts CTs outcomes and population representativeness.

REFERENCES

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3), 230–238. <https://doi.org/10.1093/jamia/ocx079>

METHODS

The Synthea AD module state machine was adapted to represent disease progression according to evidence from the literature. We extended the AD module by integrating 15 biomarkers to synthetically reproduce microdata populations from 203 CTs retrieved from **ClinicalTrials.gov**, modeling cognitive decline trajectories (including MMSE, NPI, CDR-SB, among others), ethnicity (White, Black, Asian), pharmacological interventions (cholinesterase inhibitors, anti-amyloid monoclonal antibodies, and other treatments), and mortality patterns. Reproducibility and accuracy of the synthetic data were evaluated against original datasets using Welch's t-test. These data were then integrated into the ALL-EMBRACED platform to support in silico simulations of eligibility criteria relaxation guided by SHAP-based analyses.

CONTACTS

E-mail:
darchetti@fatebenefratelli.eu,
aboccali@fatebenefratelli.eu

Tel: +39.030.350.1553

Postal address: via Pilastroni 4, Brescia (BS),
25125, Italy

RESULTS

Initial simulations demonstrated that the customized Synthea state machine successfully generated AD populations suitable for testing the relaxation of CT eligibility criteria. To date, about **500,000** synthetic AD patients have been generated. Preliminary results tested on 29 CTs showed that the resulting cohorts reproduced biomarker distributions observed in the selected CTs subgroup, with mean values and standard deviations statistically comparable ($p > 0.05$) in approximately 25% of the 25 ClinicalTrials.gov studies analyzed.

CONCLUSIONS

Synthetic patient generation represents a valuable strategy to improve CTs design within the **ALL-EMBRACED** framework, enabling realistic simulations and optimization of eligibility criteria, and ultimately fostering more inclusive, transparent, and representative clinical research in Alzheimer's disease.

