AI-generated synthetic cohorts for accelerated clinical trial design and collaboration: Data from 19,000 patients (Pts) with metastatic breast cancer (MBC)

Data Mining

Breast

Eddy Saad1, Chris Labaki1, Ziad Bakouny2, Thibaut Sanglier3, Marta Batlle3, Miguel Valero Martin3, Paolo Tarantino1, David A. Braun1, Fanny Bouquet3, Maureen Joyce3, Eliezer M. Van Allen1, Wanling Xie1, Sara M. Tolaney1, Toni K. Choueiri1

1Dana-Farber Cancer Institute, USA 2Memorial Sloan Kettering Cancer Center, USA 3 Roche

ABSTRACT

Background

Rapid advances in MBC research necessitate enhanced collaboration to expedite real-world treatment evaluation and uncover personalized approaches. AI-generated synthetic real-world data (sRWD) can mimic real-world cohorts while addressing privacy and legal concerns. Herein, we assessed the utility of sRWD in a large cohort of MBC pts, focusing on survival outcome fidelity and patient privacy.

Methods

sRWD were generated for 19,164 pts with MBC from the Flatiron database, using conditional generative adversarial networks (CTGANs) with different privacy controls, or classification and regression trees (CART). We assessed univariate faithfulness via absolute standardized mean differences (aSMD). Cox models for real-world progression-free survival (rwPFS) evaluated hazard ratio agreement using estimate agreement (EA), 95%CI overlap (POCI), and ratio of HRs (rHR). Sample-to-population re-identification risk was quantified.

Results

CART most closely replicated the original cohort's baseline characteristics and survival outcomes, while CTGAN models progressively deviated as privacy constraints increased. For rwPFS associations, CART achieved the highest HR agreement of all models. Despite CART's slightly higher re-identification risk, all datasets remained within established regulatory thresholds (<0.09) (Table). Table: 3136ODistributions of agreement measures for hazard ratios (HR) estimates of real-world progression-free survival (rwPFS) between each synthetic real-world data (sRWD) dataset and the source cohort (SC), and re-identification risk for each sRWDsRWD model

Conclusions

CART-based sRWD offers an optimal balance between data utility and privacy, faithfully reproducing variables and association patterns with acceptable re-identification risk. Thus, sRWD has the potential to accelerate collaborative real-world analyses and serve as external comparators for clinical trials.

SCIENTIFIC IMPACT

This study proves that AI-generated synthetic real-world data (sRWD) can accurately replicate large oncology cohorts while ensuring patient privacy. By analyzing 19,000 metastatic breast cancer patients, researchers identified CART-based models as the superior method for maintaining survival outcome fidelity.