

Early Prediction of Clinical Trial Outcomes Using Generative Data Augmentation: A Retrospective Simulation Study

(The NAEC – Nantes Anticipation of Clinical Trials – Project)



BACKGROUND

Patient recruitment is the most time-consuming and costly phase of clinical trials, often delaying access to effective therapies. Reliable methods to anticipate trial outcomes before full enrollment could substantially improve decision-making, trial efficiency, and resource allocation in clinical research. Synthetic data generation for data augmentation may offer a promising approach to predict treatment effects at earlier stages of a trial, support interim analyses, and maximize the value of studies that do not achieve their planned recruitment targets.

OBJECTIVES

- To evaluate whether Avatar-based generative data augmentation method can reliably reproduce final treatment effect estimates in a survival clinical trial using partially recruited cohorts.
- To identify the minimum recruitment threshold associated with performance comparable to that observed at 90% recruitment.

CASE STUDY

- ACTG 175 clinical trial in HIV infection, Open-access dataset from the AIDS Clinical Trials Group (ACTG)
- Design: randomized, controlled, double-blind, parallel-group, clinical trial, comparing zidovudine monotherapy (n = 619) versus zidovudine + didanosine combination therapy (n = 513)
- Outcome: time to $\geq 50\%$ decline in CD4 cell count
- Primary analysis: Cox proportional hazards model
- Results: hazard ratio (HR) = 0.50 (95% CI: 0.39-0.63), β coefficient (log HR) = -0.7037

Références

Hammer SM, Katzenstein DA, Hughes MD, Gundacker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, Hirsch MS, Merigan TC. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. AIDS Clinical Trials Group Study 175 Study Team. N Engl J Med. 1996 Oct 10;335(15):1081-90. doi: 10.1056/NEJM199610103351501. PMID: 8813038.

<https://search.r-project.org/CRAN/refmans/speff2trial/html/ACTG175.html>

<https://kmcconeghy.github.io/jumble/reference/ACTG175.html>

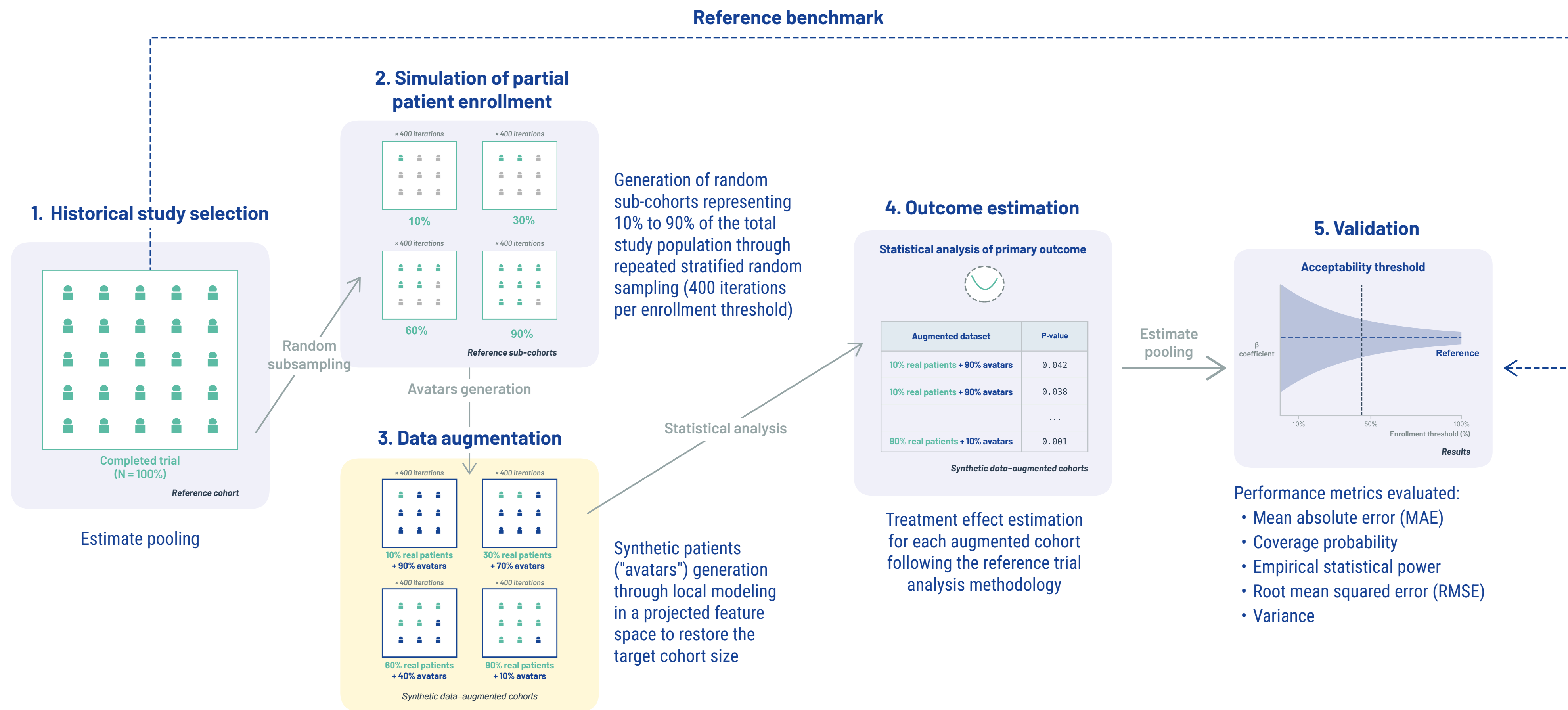
<https://r-packages.io/datasets/ACTG175>

CONCLUSION

In this retrospective simulation study, Avatar-based generative data augmentation progressively improved the agreement between interim and final treatment effect estimates as the proportion of observed data increased. Performance appeared to stabilize from approximately 50% patient enrollment onward, suggesting that reliable outcome estimation may be achievable well before full recruitment.

These preliminary results suggest that synthetic data generation may help strengthen interim analyses, support earlier go/no-go decisions and trial decision-making before full enrollment is achieved. Although validation across additional datasets and therapeutic areas is ongoing, this approach offers promising new opportunities to accelerate clinical research while maximizing the value of partially recruited studies.

METHODS



Avatar-based generative data augmentation method

Avatar is a privacy-preserving, individual-centered, synthetic data generation method based on three steps:

- Projection of the original data into a standardized continuous feature space
- Identification of the k nearest neighbors for each individual
- Generation of a synthetic "avatar" for each individual through local perturbation technique within the neighborhood structure

References

<https://docs.octopize.io/docs/principles/method/tabular/>

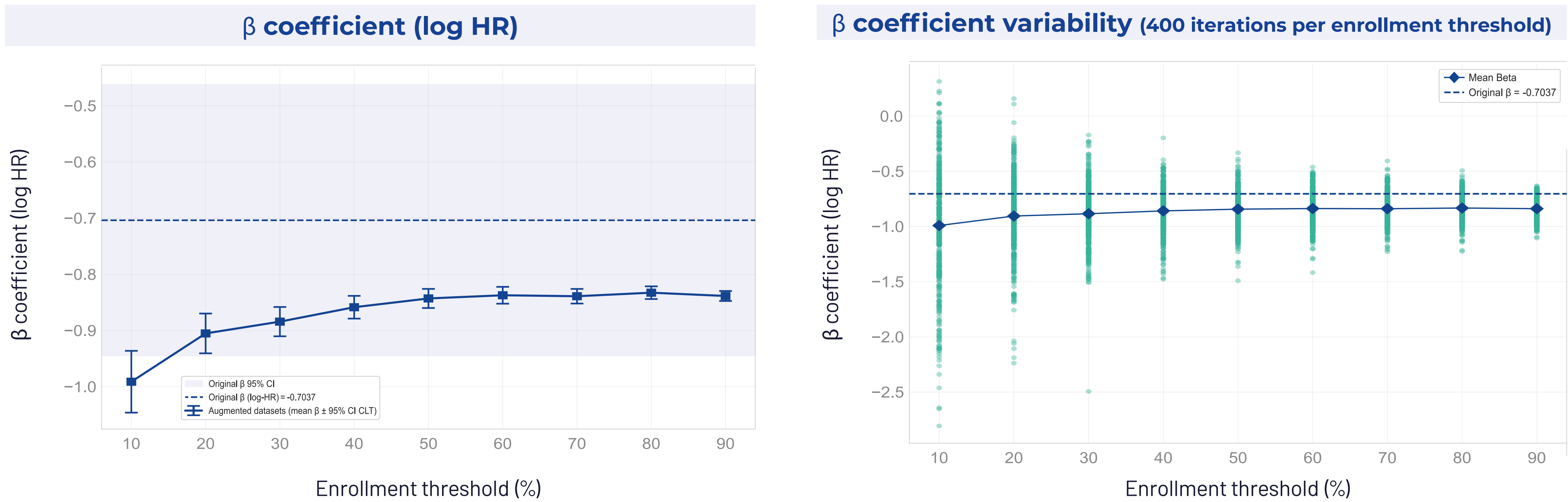
Guillaudeux M, Rousseau O, Petot J, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. npj Digit. Med. 6, 37 (2023). <https://doi.org/10.1038/s41746-023-00771-5>

Key advantages:

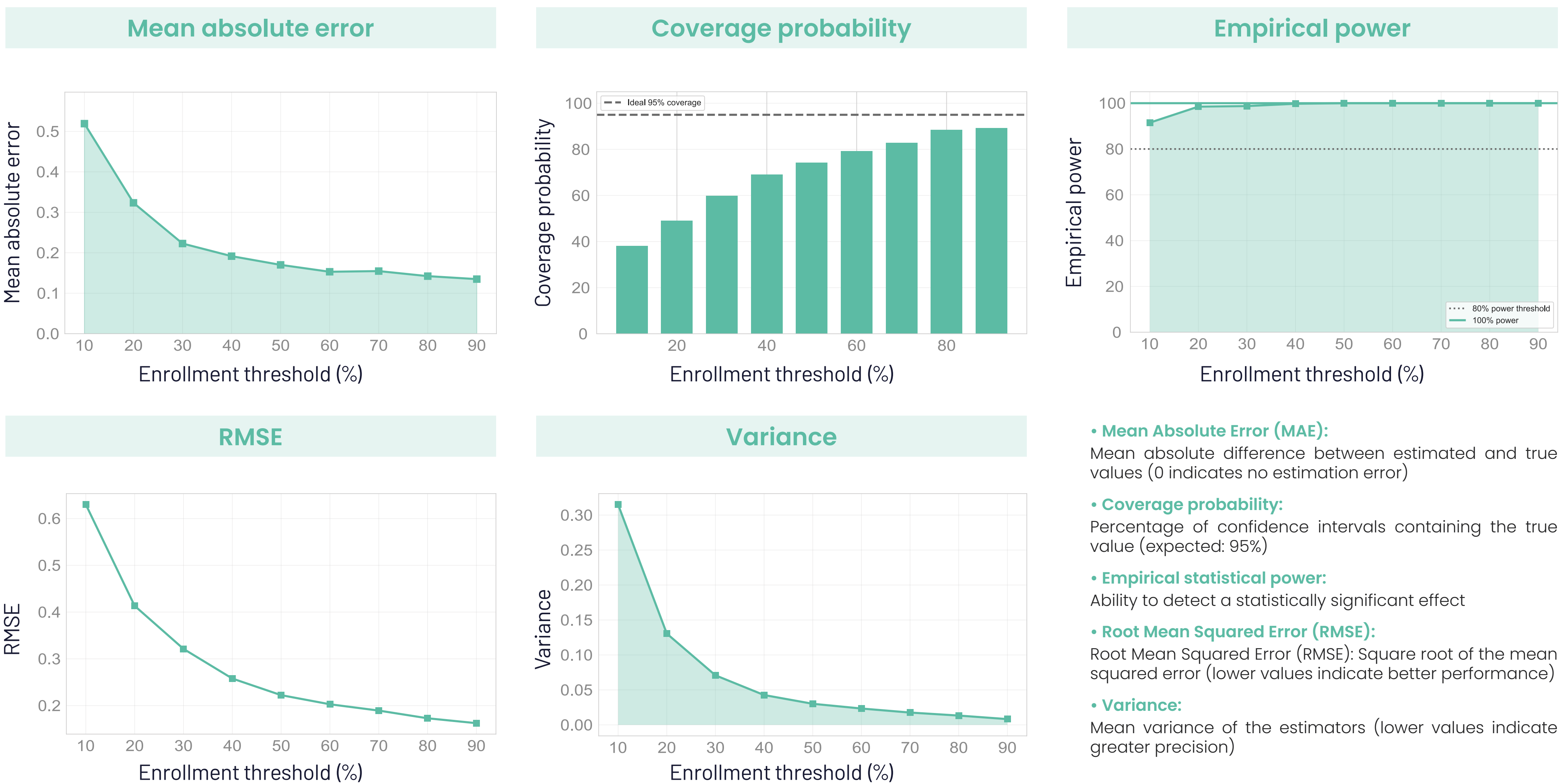
- Preservation of the joint statistical structure within local neighborhoods, including marginal distributions, correlations, and higher-order dependencies
- Formal protection against re-identification, as each synthetic avatar is concealed within a group of k original individuals
- Iterative data augmentation, using a predefined multiplication factor, to achieve the target sample size while preserving the original multivariate data structure

RESULTS

Outcome estimation



Validation: Performance metrics evaluation



- Mean Absolute Error (MAE):** Mean absolute difference between estimated and true values (0 indicates no estimation error)
- Coverage probability:** Percentage of confidence intervals containing the true value (expected: 95%)
- Empirical statistical power:** Ability to detect a statistically significant effect
- Root Mean Squared Error (RMSE):** Root Mean Squared Error (RMSE): Square root of the mean squared error (lower values indicate better performance)
- Variance:** Mean variance of the estimators (lower values indicate greater precision)

Overall, performance metrics stabilized from approximately 50% enrollment onward.