MSR15

Advancing Target Trial Emulation with Synthetic Data: The Target Trial Optimization Framework

Paolo Messina[1], Pauline Bambury[1], Mario Torchia[1], Luca Emili[1], Daniel Roeshammar[1]

Affiliations:
1. InSilicoTrials Technologies S.p.A., Trieste, Italy

InSilicoTrials

Introduction

Target trial emulation (TTE)[1] provides a structured framework for causal inference from real-world data by specifying the hypothetical randomized trial that observed data aim to replicate. While TTE has advanced retrospective analysis rigor, clinical development lacks prospective tools to systematically optimize trial designs before implementation.

Objective

Develop and demonstrate the Target Trial Optimization (TTO) framework, integrating TTE concept with synthetic data generation, and modeling & simulation (M&S) to enable systematic exploration and optimization of trial specifications across all TTE domains prior to study initiation for quantitative decision making.

Table 1. Target Trial Space Exploration

TTE Domain	TTE Description	TTO Exploration
Eligibility Criteria	Define inclusion/exclusion characteristics	Evaluate alternative population definitions and boundary conditions
Treatment Strategies	Specify interventions under comparison	Explore hypothetical treatment regimens and intervention sequences
Assignment Procedures	Assume randomization conditional on covariates	Simulate allocation mechanisms, design features, and confounding scenarios
Outcome(s)	Specify primary and secondary endpoints	Assess alternative endpoint(s) (definitions) and measurement approaches
Follow-up Period	Define observation start, duration, and end	Test varying observation windows and dropout mechanisms
Causal Contrast(s)	State estimand(s) of interest: e.g., ITT, PP.	Test alternative estimands and intercurrent event handling strategies
Analysis Plan	Describe statistical method	Explore analytical methods

Method

Stage 1: Target Trial Specification

- Apply TTE principles to define trial protocol aligned with the clinical question
- Identify data/knowledge sources: real-world data (RWD) for data-driven approaches, literature, mechanistic models or expert knowledge for knowledge-based approaches

Stage 2: Synthetic Data Generation

- **Data-driven approach:** When sufficient RWD, employ generative models
- **Knowledge-based approach:** When clinical data are sparse, mechanistic disease progression models, PK/PD models

Stage 3: Design Space Simulation

Systematic evaluation of protocol variants through what-if scenario (Table 1)

Stage 4: Monte Carlo Simulation Architecture

Each scenario is evaluated through independent Monte Carlo simulations

Stage 5: Operating Characteristics Quantification

E.g. Measure the effect, type I and type II errors and probability of success

Stage 6: Multi-Criteria Optimization & Selection

Compare scenarios using operating characteristics paired with feasibility constraints to identify optimal trial specifications

TTO answers:
What trial design maximizes statistical power and probability of success by optimally targeting populations with the greatest treatment effect?

Results

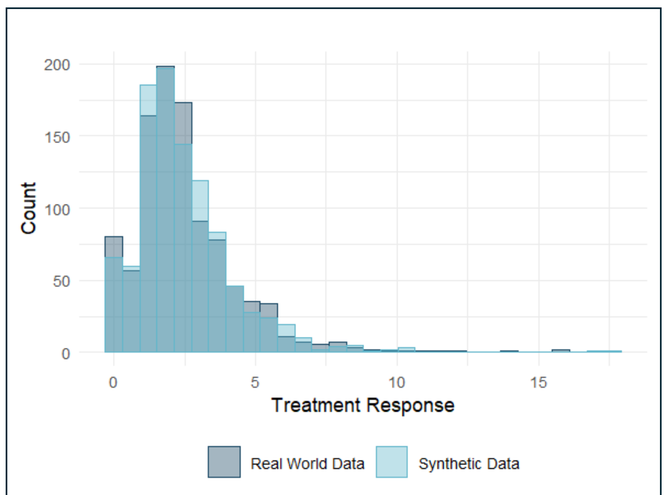
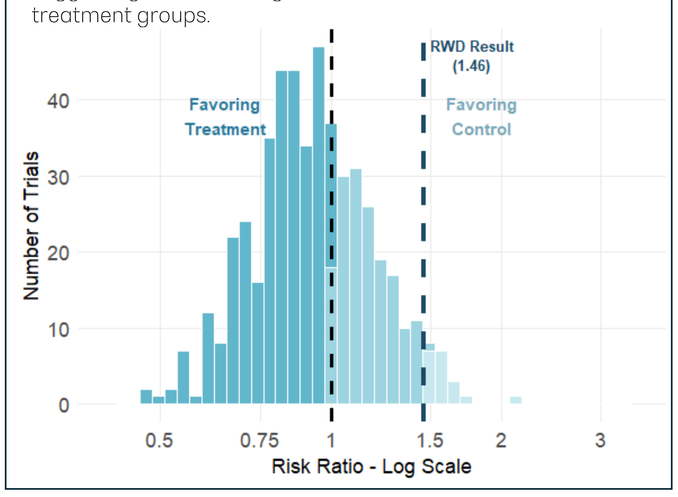


Figure 1. Synthetic Data Fidelity on Treatment Response. Distribution of treatment response showing high overlap between real-world data (darker bars) and synthetic data (lighter bars), demonstrating fidelity of the synthetic cohort.

Figure 2. Robustness assessment for a low-prevalence endpoint. Distribution of risk ratios from Monte Carlo simulations (n=500 trials) showing treatment effect with 95% confidence intervals. RWD point estimate (1.46, dashed line) falls in the right tail of the synthetic data-derived distribution, suggesting rebalancing toward similar risk between treatment groups.

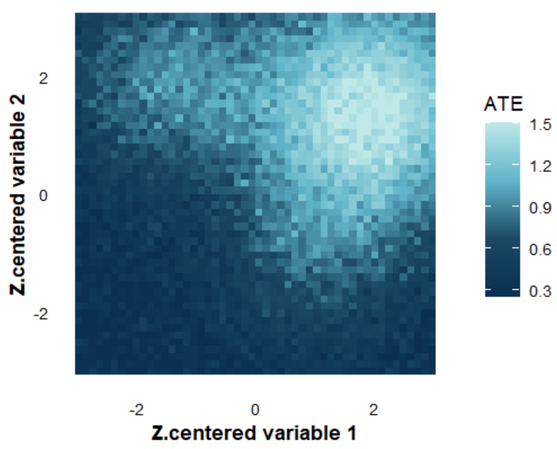


- A purely data-driven synthetic data generation method was implemented using TabPFN, a recently introduced tabular foundation model[2]
- The optimized configuration demonstrated high fidelity (Figure 1) and utility with marginal privacy risk (the validation metrics protocol will be part of a separate publication)
- Simulated results aligned with real-world data (RWD) findings while providing enhanced precision for subgroups and outcomes where RWD precision was constrained by limited observed events (Figure 2)
- What-if scenarios identified combinations of patient characteristics that maximized average treatment effect (Figure 3), enabling exploration of optimal treatment strategies beyond the constraints of the observed dataset

Conclusion

The Target Trial Optimization Framework transforms target trial emulation from a retrospective analytical tool into a prospective design optimization engine, enabling evidence-based trial planning through synthetic data generation and systematic scenario evaluation. By maximizing probability of success through optimal population selection and design parameters, this framework enables precision trial design for precision medicine.

Figure 3. What-if Average Treatment Effect.
Lighter regions indicate higher treatment benefit. Analysis conducted using Monte Carlo simulations with synthetic cohorts to enable comprehensive scenario exploration.



References:
1. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758-64. doi: 10.1093/aje/kwv254.
2. Hollmann, N., Müller, S., Purucker, L. et al. Accurate predictions on small data with a tabular foundation model. Nature 637, 319–326 (2025). https://doi.org/10.1038/s41586-024-08328-6.