Empowering Trial by Multi-source Multi-site Real-world Data: A Negative Control-Calibrated Digital Twin Approach

Dazheng Zhang[1,2], PhD, Huiyuan Wang[1,2], PhD, Yiwen Lu[1,2], BS, Yong Chen[1,2], PhD
[1] The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA,
[2] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA.

Background

Randomized controlled trials (RCTs) ensure robust causal inference through randomization but often lack generalizability due to strict inclusion criteria, high costs, and limited scalability. Real-world data (RWD)—including electronic health records (EHRs), insurance claims, and registries—offer broader population coverage and longitudinal follow-up at lower cost, but are prone to bias due to non-randomization and incomplete data access[1–3]. Integrating external controls from RWD can enhance the statistical efficiency of RCTs by reducing required sample sizes and costs[4,5].

Digital twin technology, originating from engineering, uses machine learning on high-dimensional RWD to simulate patient-specific counterfactuals and treatment responses[6]. The PROCOVA method applies this by creating virtual control groups to improve trial efficiency[7]. However, discrepancies between trial and real-world populations can introduce bias, necessitating careful calibration.

Despite the promise of digital twins in strengthening RCTs, the validity of digital twin-based analyses depends on precise calibration, as discrepancies between trial and real-world populations can introduce systematic biases (**Figure 1a**). Existing methods from OHDSI community, such as empirical calibration from negative control outcomes[8–10], offer potential strategies for mitigating systematic errors, yet challenges remain in calibrating the distribution shift between the RCT and RWD populations.

To address this, we propose a novel framework that corrects model shift bias in RCT-RWD integration using digital twins and negative control outcomes. By leveraging negative control outcomes in a data-driven calibration, our approach ensures robust, generalizable treatment effect estimates across data sources.

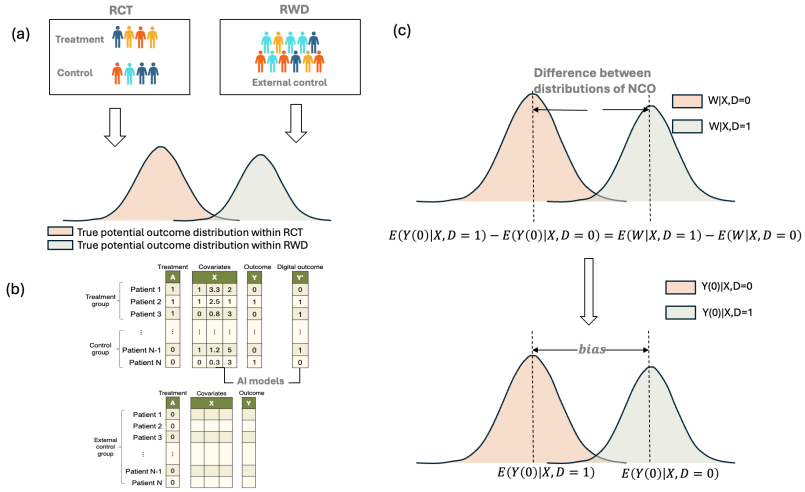

Figure 1. The working flow of the proposed method. a) the indication of the scenario, that is

counterfactually, the outcome distribution given control for the RCT and RWD are different. The difference between the distribution of the outcome for the RCT and the RWD are known as the model-shift bias. b) negative control outcomes (NCOs) are those outcomes whose treatment response are hypothetically zero. The difference between the RCT and RWD for the NCOs helps to identify the model-shift for the outcomes of interest. c) The overview of the proposed method. The model used to predict the digital twin (shown in the last column for the RCT data) can be trained within the RWD data.

Methods

Fig 1 (b) and **(c)** showed the overall working proposed method. Fig 1 (b) introduces the concept of NCOs, outcomes that are theoretically unaffected by treatment assignment, allowing for an empirical estimate of the model shift bias. The observed discrepancy in NCO distributions between RCT and RWD serves as a diagnostic tool for identifying systematic shifts in outcome distributions. Finally, **Fig 1 (c)** outlines our methodological framework: we utilize predictive modeling within the RWD to construct a *digital twin*—an individualized outcome model for RCT participants. By systematically calibrating for model shift, our method facilitates more robust and generalizable treatment effect estimation in hybrid trial designs.

To evaluate the proposed method in real-world data, we applied it to a combined dataset from the SPRINT-MIND trial and the iSTAGING Consortium[11,12]. The SPRINT-MIND trial[13] investigated the effects of intensive blood pressure control on cerebral white matter lesions. Participants underwent magnetic resonance imaging (MRI) to assess changes in brain structure. The study found that intensive systolic blood pressure control was associated with a smaller increase in white matter lesion volume compared to standard treatment. The treatment group was defined as those receiving intensive antihypertensive therapy targeting a systolic blood pressure (SBP) of 120 mm Hg, while the control group received standard antihypertensive therapy targeting an SBP of 140 mm Hg.

Results

The iSTAGING Consortium is an international collaboration that conducted studies across diverse geographic locations between 1995 and 2020, utilizing various scanners and acquisition parameters. For our analysis, we focused on cross-sectional scans from a subgroup of participants (n= 43,202) within the iSTAGING dataset who had complete MRI measures available. We integrate the RWD from four data sources: ADNI (n = 1,764), OASIS (n = 617), and Penn (n = 913), UK Biobank (n=39, 908). Key covariates for training the digital twins included age, sex, race, and education and other medication history. The machine learning model used for digital twin construction was based on random forests. We selected the deep white matter lesion volume as the main outcome of interest and the gray matter volume as the negative control outcomes.

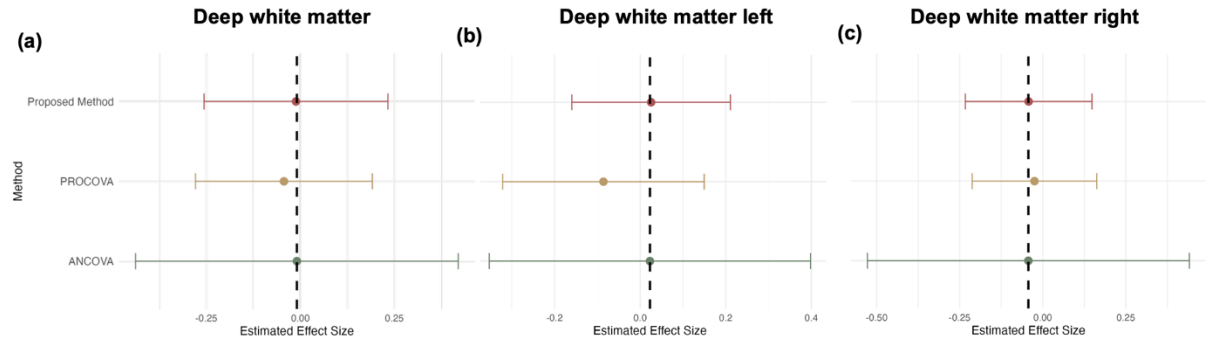

Figure 2. Estimated effect sizes and confidence intervals for three brain regions—(a) deep cerebral white matter, (b) deep cerebral white matter left, and (c) deep cerebral white matter right—using the proposed method (**dark red**), PROCOVA (**gold**), and ANCOVA (**gray**).

Figure 2 illustrates the estimated effect sizes across three brain regions—(a) deep cerebral white matter, (b) deep cerebral white matter left, and (c) deep cerebral white matter right—comparing the proposed method with two baseline approaches, PROCOVA and ANCOVA. As ANCOVA relies solely on RCT data, its estimates serve as the gold standard due to the benefits of randomization. The PROCOVA method incorporates external real-world data (RWD) but does not account for model-shift bias, leading to potential inaccuracies. The proposed method, by addressing model-shift bias, produces estimates that align closely with ANCOVA while achieving notable efficiency gains. Across all panels, the proposed method consistently yields more precise estimates, reflected in its narrower confidence intervals compared to ANCOVA, and avoids the potential distortions observed in PROCOVA.

Conclusion

In summary, our proposed digital twin framework with negative control outcome calibration effectively addresses model shift bias in RCT-RWD integration. Applied to brain imaging data from SPRINT-MIND and iSTAGING, the method achieved treatment effect estimates that closely aligned with RCT-based ANCOVA while offering improved statistical precision. This approach enhances the validity and efficiency of hybrid trial designs, paving the way for more reliable real-world evidence for OHDSI community.

References

1 Rothwell PM. External validity of randomised controlled trials:"to whom do the results of this trial apply?" *The Lancet* 2005; **365**: 82–93.
2 Miksad RA, Abernethy AP. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin Pharmacol Ther* 2018; **103**: 202–5.
3 Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* 2018; **320**: 867–8.
4 Viele K, Berry S, Neuenschwander B, *et al.* Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat* 2014; **13**: 41–54.
5 Schmidli H, Häring DA, Thomas M, Cassidy A, Weber S, Bretz F. Beyond randomized clinical trials: use of external controls. *Clin Pharmacol Ther* 2020; **107**: 806–16.
6 Katsoulakis E, Wang Q, Wu H, *et al.* Digital twins for health: a scoping review. *NPJ Digit Med* 2024; **7**: 77.
7 Schuler A, Walsh D, Hall D, Walsh J, Fisher C, Initiative ADN. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *Int J Biostat* 2022; **18**:

329–56.
8 Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med* 2014; **33**: 209–18.
9 OHDSI/EmpiricalCalibration: An R package for performing empirical calibration of observational study estimates. https://github.com/OHDSI/EmpiricalCalibration (accessed Oct 3, 2024).
10 Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A* 2018; **115**: 2571–7.
11 Habes M, Erus G, Toledo JB, *et al.* White matter hyperintensities and imaging patterns of brain ageing in the general population. *Brain* 2016; **139**: 1164–79.
12 Govindarajan ST, Mamourian E, Erus G, *et al.* Machine learning reveals distinct neuroanatomical signatures of cardiovascular and metabolic diseases in cognitively unimpaired individuals. *Nat Commun* 2025; **16**: 2724.
13 Nasrallah IM, Gaussoin SA, Pomponio R, *et al.* Association of intensive vs standard blood pressure control with magnetic resonance imaging biomarkers of alzheimer disease: secondary analysis of the SPRINT MIND randomized trial. *JAMA Neurol* 2021; **78**: 568–77.