Integrating digital twins with randomized experiments

Yanping Li
Division of Biostatistics
University of California, Berkeley
yanpingli@berkeley.edu

Xinwei Ma
Department of Economics
University of California, San Diego
x1ma@ucsd.edu

Jingshen Wang *
Division of Biostatistics
University of California, Berkeley
jingshenwang@berkeley.edu

Abstract

Randomized controlled trials (RCTs) identify treatment effects for an enrolled trial population by assigning treatment independently of potential outcomes, conditional on baseline covariates. These trial-specific effects may not generalize to a broader target population when the target and trial covariate distributions differ. This problem becomes harder when individual-level target population data are unavailable. Pre-trained large language models (LLMs) offer one way to address this data limitation by generating digital twins (DTs) with synthetic covariates intended to resemble the target population. Synthetic covariates alone, however, are insufficient, because neither the RCT covariate distribution nor the DT covariate distribution necessarily matches the target covariate distribution. We therefore propose a statistical inference procedure that integrates RCTs with calibrated DTs using external target population summaries. The procedure represents the target covariate distribution as a calibrated mixture of the RCT and DT covariate distributions, allowing DTs to contribute covariate information while limiting the influence of poorly calibrated synthetic data. The procedure also incorporates LLM-generated auxiliary outcome predictions through calibrated outcome regressions, improving precision without changing the estimand. Theoretical results and simulation studies show that the proposed estimators reduce bias relative to RCT-only estimation and improve efficiency for overall and subgroup target treatment effects.

1 Introduction

Randomized controlled trials (RCTs) provide a strong basis for causal inference in the population that enters the study. Randomization makes treatment assignment independent of potential outcomes under the trial design, so an RCT can identify the average treatment effect (ATE) for enrolled participants. Many scientific and policy questions, however, concern a broader target population. Trial participants may differ from that population because eligibility criteria, recruitment, consent, site selection, and study logistics shape who appears in the randomized data [Rothwell, 2005, Buchanan et al., 2018]. When treatment effects vary with baseline covariates, this difference can make the trial ATE differ from the target population ATE, creating an external validity problem.

Target population inference addresses this problem by combining randomized evidence with information about the population of interest. Under a transportability assumption, the trial and target population share the same conditional outcome distributions given baseline covariates, although their

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

covariate distributions may differ [Pearl and Bareinboim, 2014, Dahabreh et al., 2020]. The RCT identifies how treatment response varies with covariates, and the remaining task averages this conditional response over the target covariate distribution. Existing approaches often use individual-level target covariates from registry samples, surveys, or observational data to reweight or standardize trial evidence [Cole and Stuart, 2010, Stuart et al., 2011, Buchanan et al., 2018, Dahabreh et al., 2020]. Recent reviews place these methods within the broader literature on generalizability and transportability [Degtiar and Rose, 2023]. In many applications, however, researchers observe only partial target population information, such as marginal means, subgroup proportions, or registry aggregates.

Digital twins (DTs) offer a candidate source of covariate information when individual-level target data are unavailable. In this paper, DTs are synthetic units generated to resemble the target population through baseline covariates [Katsoulakis et al., 2024, Laubenbacher et al., 2024]. They describe who is in the target population, not how those individuals respond to treatment. The RCT remains the source of causal identification, while the DT sample helps represent the covariate distribution over which the RCT-identified conditional treatment effect is averaged. Recent work on clinical DTs and LLM-based virtual patients suggests that synthetic patient representations can support clinical modeling and trial simulation [Das et al., 2023, Wang et al., 2024].

DTs alone, however, do not solve the target representation problem. A generated covariate distribution may miss important target population features, and the RCT covariate distribution may also differ from the target covariate distribution because trial enrollment is selective. Therefore, neither the RCT covariate distribution nor the DT covariate distribution necessarily equals the target covariate distribution. A useful estimator should allow DTs to contribute covariate information while limiting their influence when they conflict with external target summaries.

We propose a statistical inference procedure that integrates RCTs with calibrated DTs to estimate target population ATEs. The method represents the target covariate distribution as a calibrated mixture of the RCT and DT covariate distributions. It uses external summaries to constrain the implied target distribution, keeps the calibrated target close to the RCT covariate distribution, and penalizes weight concentration. Randomized outcomes identify conditional treatment response, and calibrated DT covariates help determine the population over which that response is averaged. The procedure also uses LLM-generated auxiliary outcome predictions through calibrated outcome regressions, improving precision while preserving the target population ATE as the estimand.

We make three contributions.

1. We introduce a calibrated mixture representation for the target covariate distribution that combines the RCT covariate distribution with a reweighted DT covariate distribution anchored by external target summaries.
2. We formulate target population treatment effect estimation under limited target data access by separating randomized outcomes used for causal identification from DT covariates used for target population representation.
3. We develop efficient-influence-function-based estimators that combine randomized outcomes, calibrated DTs, and LLM-generated auxiliary outcome predictions without changing the target estimand. We prove calibration-weight consistency, derive an efficient influence function representation, establish consistency, asymptotic normality, and variance consistency, and evaluate the method in a BRIGHTEN-based simulation.

The rest of the paper proceeds as follows. Section 2 introduces the data structure, target estimand, calibrated mixture representation, estimators, calibration weight estimation, and subgroup extension. Section 3 gives the assumptions and theoretical results. Section 4 reports the BRIGHTEN-based simulation study. Section 5 discusses the limitations and future work.

2 Method

In this section, we define the target population ATE and the calibrated mixture representation used to estimate it. The RCT provides randomized outcomes for causal identification. The DT sample provides synthetic baseline covariates that may help represent the target population after calibration to external summaries. LLM-generated auxiliary predictions enter only through outcome adjustment

and do not define the causal estimand. We first define the data structure, target estimand, and mixture representation. We then present the estimators, describe calibration-weight estimation, and extend the framework to subgroup analysis.

2.1 Improving External Validity of RCTs with Generative Agents

We consider an RCT with binary treatment. Let $\{(X_i, T_i, Y_i)\}_{i=1}^n$ denote an independent and identically distributed sample drawn from the distribution P_1 induced by the randomized experiment. Here $X_i \in \mathcal{X} \subset \mathbb{R}^p$ represents baseline covariates, $T_i \in \{0, 1\}$ the treatment indicator, and Y_i the observed outcome. In the RCT, the treatment is randomly assigned conditional on the baseline covariates X_i. Let $Y_i(0)$ and $Y_i(1)$ denote the corresponding potential outcomes. For each unit i and treatment level $t \in \{0, 1\}$, let $\widehat{Y}_i(t)$ denote the LLM-generated auxiliary outcome prediction under treatment t.

Randomization in an RCT secures internal validity, but it does not ensure external validity. In this setting, external validity refers to whether the treatment effect in the trial also applies to the target population. The two ATEs may differ when the trial sample and the target population have different baseline covariate distributions, especially if treatment effects vary across covariate-defined subgroups. This can arise when eligibility criteria define a narrow study population or when enrolled participants differ systematically from the target population. The concern is especially acute in small trials and when clinically important subgroups are underrepresented.

To address this external validity problem, we use a generative model to construct DTs, which are synthetic individuals intended to resemble the target population in baseline covariates. The DT sample contributes covariate information only. It does not provide observed outcomes or identify treatment response. We also generate LLM-based auxiliary outcome predictions for the DT sample using the same procedure as in the RCT sample. The synthetic sample is $\{(\widetilde{X}_j, \widetilde{Y}_j(1), \widetilde{Y}_j(0))\}_{j=1}^N$, where $\widetilde{X}_j$ denotes generated covariates and $\widetilde{Y}_j(t)$ denotes auxiliary outcome predictions obtained using the same LLM-based procedure. These synthetic observations follow a distribution P_0, whose covariate distribution is allowed to approximate the target covariate distribution.

We denote the target population distribution by $P = P_{\mathtt{X}} \cdot P_{\mathtt{Y}|\mathtt{X}}$, where $P_{\mathtt{X}}$ is the target covariate distribution and $P_{\mathtt{Y}|\mathtt{X}}$ is the conditional outcome model. Assuming transportability, the RCT and the target population share the same conditional outcome model given baseline covariates, although their covariate distributions may differ. To combine the RCT and DT samples, we introduce a source indicator $S \in \{0, 1\}$, where $S = 1$ corresponds to RCT observations and $S = 0$ to DT observations. For $s \in \{0, 1\}$, let $P_s = P_{\mathtt{X},s} \cdot P_{\mathtt{Y}|\mathtt{X}}$, where $P_{\mathtt{X},s}$ denotes the covariate distribution for source s. Because neither the RCT covariate distribution $P_{\mathtt{X},1}$ nor the DT covariate distribution $P_{\mathtt{X},0}$ necessarily equals the target covariate distribution $P_{\mathtt{X}}$, we introduce a reweighting function $\rho(x)$ for the DT sample. The target covariate distribution is represented as

$$\mathrm{d}P_{\mathtt{X}}(x) = \eta \, \mathrm{d}P_{\mathtt{X},1}(x) + (1 - \eta)\rho(x)\mathrm{d}P_{\mathtt{X},0}(x), \tag{1}$$

where $\eta \in (0, 1)$ is a fixed design proportion with $n/(n + N) \to \eta$, so $\mathbb{P}(S = 1) = \eta$ in the pooled population. In implementation, we set η by the relative sample sizes or vary it as a sensitivity parameter. The weight ρ is normalized so that $E_{P_{\mathtt{X},0}}[\rho(X)] = 1$, which makes the DT component a probability measure. The unified observed data are $O = \big(S, X, S\,T, S\,Y, Y^\dagger(0), Y^\dagger(1)\big)$, where X is observed in both sources, T and Y are observed only for RCT units, and $Y^\dagger(t) := S\,\widehat{Y}(t) + (1 - S)\widetilde{Y}(t)$ for $t \in \{0, 1\}$.

In practice, the target covariate distribution $P_{\mathtt{X}}$ is rarely fully observed. External data sources often report only partial summaries, such as marginal means and subgroup proportions. To incorporate this information, we construct $\rho(x)$ so that the implied target distribution matches the available summaries while remaining close to the RCT covariate distribution. Specifically, we obtain $\rho(x)$ by solving

$$\begin{aligned}
\min_{\rho} \quad & \mathtt{Dist}(P_{\mathtt{X},1}, P_{\mathtt{X}}(\rho)) && \leftarrow \text{Encourage similarity to RCT} \\
\text{s.t.} \quad & \rho(x) \in (0, C],\ \forall x \in \mathcal{X}, && \leftarrow \text{Boundedness} \\
& \mathbb{E}_{P_{\mathtt{X},0}}[\rho(X)] = 1, && \leftarrow \text{Valid density} \\
& \max_{a \in \mathcal{A}} \|\mathbb{E}_{P_{\mathtt{X}}}[g_a(X)] - m_a\|_2 \le \delta. && \leftarrow \text{Summary statistics alignment}
\end{aligned} \tag{2}$$

Here $P_{\mathtt{X}}(\rho)$ denotes the implied target covariate distribution induced by the reweighting function ρ. The objective $\mathtt{Dist}(\cdot,\cdot)$ is the Wasserstein-1 distance between the RCT covariate distribution and the implied target covariate distribution. The constant $C > 0$ bounds the weights, and the normalization constraint makes the reweighted DT component a valid probability measure. For covariates indexed by $\mathcal{A}$, m_a denotes externally reported summary statistics and $g_a : \mathcal{X} \to \mathbb{R}$ denotes the corresponding transformation. The tolerance parameter $\delta \geq 0$ allows each external summary constraint indexed by $a \in \mathcal{A}$ to be matched within a controlled deviation.

Our parameter of interest is the ATE in the target population,

$$\tau = \mathbb{E}_P[Y(1) - Y(0)] = \mathbb{E}_{P_{\mathtt{X}}}\left[\mathbb{E}_{P_{\mathtt{Y}|\mathtt{X}}}[Y(1) - Y(0) \mid X]\right].$$

For $t \in \{0, 1\}$, define the conditional mean of the RCT potential outcomes as

$$\mu_t(x) = \mathbb{E}[Y(t) \mid X = x, S = 1]. \tag{3}$$

Under the mixture representation of $P_{\mathtt{X}}$, the target estimand can be written as

$$\tau = \eta\, \mathbb{E}_{P_{\mathtt{X},1}}\left[\mu_1(X) - \mu_0(X)\right] + (1-\eta)\mathbb{E}_{P_{\mathtt{X},0}}\left[\rho(X)(\mu_1(X) - \mu_0(X))\right].$$

2.2 Efficient inference with calibrated DTs

We estimate the target ATE τ using the combined RCT and DT sample together with several nuisance functions. Let $e(x) = \mathbb{P}(T = 1 \mid X = x, S = 1)$ and $\pi(x) = \mathbb{P}(S = 1 \mid X = x)$ denote the treatment propensity score and the source propensity score, respectively. Recall that $\mu_t(x)$ is defined in Eq. (3), and that the calibration weight $\rho(x)$ is obtained by solving Problem (2). To incorporate the auxiliary outcome predictions, let $\mu^\dagger_{t,s}(x) = \mathbb{E}\big[Y^\dagger(t) \mid X = x, S = s\big]$ denote the source-specific conditional mean of the auxiliary prediction under treatment level t. We also define

$$\omega_t(x) = \frac{\mathbb{C}\mathrm{ov}\big(Y, Y^\dagger(t) \mid T = t, X = x, S = 1\big)}{\mathbb{V}\mathrm{ar}\big(Y^\dagger(t) \mid T = t, X = x, S = 1\big)},$$

which is the conditional projection coefficient of the outcome on the auxiliary prediction in the RCT under treatment level t. We define $\omega_t(x)$ with respect to the RCT because only the RCT identifies the relationship between the observed outcome and its auxiliary prediction.

We estimate the nuisance functions with flexible machine learning methods and combine them through influence-function-based estimators. Let $\widehat{e}(\cdot)$, $\widehat{\pi}(\cdot)$, $\widehat{\mu}_t(\cdot)$, $\widehat{\mu}^\dagger_{t,1}(\cdot)$, $\widehat{\mu}^\dagger_{t,0}(\cdot)$, $\widehat{\omega}_t(\cdot)$, and $\widehat{\rho}(\cdot)$ denote the corresponding estimators. We estimate e, π, μ_t, $\mu^\dagger_{t,1}$, $\mu^\dagger_{t,0}$, and ω_t using flexible machine learning methods within a double machine learning framework. We estimate ρ by solving the empirical calibration problem described in the next subsection.

The efficient influence function (EIF) for the target-population ATE is

$$\begin{aligned}\phi(O) =& S\Big(\frac{T\big(Y - \mu_1(X)\big)}{e(X)} - \frac{(1-T)\big(Y - \mu_0(X)\big)}{1 - e(X)}\Big)\Big(1 + \rho(X)\frac{1-\pi(X)}{\pi(X)}\Big) \\ &+ \Big(S + (1-S)\rho(X)\Big)\Big(\mu_1(X) - \mu_0(X)\Big) - \tau.\end{aligned} \tag{4}$$

This representation leads to the EIF-based estimator $\widehat{\tau} = (n+N)^{-1}\sum_{i=1}^{n+N}\widehat{\varphi}_i$, where $\widehat{\varphi}_i$ is the uncentered plug-in score contribution

$$\begin{aligned}\widehat{\varphi}_i = S_i\Big(\frac{T_i\big(Y_i - \widehat{\mu}_1(X_i)\big)}{\widehat{e}(X_i)} - \frac{(1-T_i)\big(Y_i - \widehat{\mu}_0(X_i)\big)}{1 - \widehat{e}(X_i)}\Big)\Big(1 + \widehat{\rho}(X_i)\frac{1-\widehat{\pi}(X_i)}{\widehat{\pi}(X_i)}\Big) \\ + \Big(S_i + (1-S_i)\widehat{\rho}(X_i)\Big)\Big(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big).\end{aligned}$$

The first term corrects for treatment assignment in the RCT and reweights the residual contribution toward the target population. The second term averages the estimated conditional treatment effect over the calibrated pooled sample.

Auxiliary predictions can be incorporated more directly through a calibrated outcome model. Define

$$m_{t,s}(x, y^\dagger) = \mu_t(x) + \omega_t(x)\{y^\dagger - \mu^\dagger_{t,s}(x)\}$$

for $t \in \{0,1\}$ and $s \in \{0,1\}$, where s indexes the data source rather than the treatment arm. The source-specific centering by $\mu^{\dagger}_{t,s}(x)$ allows the auxiliary prediction distribution to differ between RCT and DT units. The coefficient $\omega_t(x)$ is learned from the RCT, where outcomes are observed. Replacing $\mu_t(x)$ with this calibrated outcome model gives $\widetilde{\tau} = (n+N)^{-1}\sum_{i=1}^{n+N}\widetilde{\varphi}_i$, where

$$\widetilde{\varphi}_i = S_i\Big(\frac{T_i\big(Y_i - \widehat{m}_{1,S_i}(X_i, Y_i^{\dagger}(1))\big)}{\widehat{e}(X_i)} - \frac{(1-T_i)\big(Y_i - \widehat{m}_{0,S_i}(X_i, Y_i^{\dagger}(0))\big)}{1-\widehat{e}(X_i)}\Big)\Big(1 + \widehat{\rho}(X_i)\frac{1-\widehat{\pi}(X_i)}{\widehat{\pi}(X_i)}\Big)$$
$$+ \Big(S_i + (1-S_i)\widehat{\rho}(X_i)\Big)\Big(\widehat{m}_{1,S_i}(X_i, Y_i^{\dagger}(1)) - \widehat{m}_{0,S_i}(X_i, Y_i^{\dagger}(0))\Big).$$

Relative to $\widehat{\tau}$, this estimator uses auxiliary predictions through source-specific calibrated regressions.

Both estimators admit standard variance estimation through their estimated influence-score contributions. Let $\widehat{\psi}_i^*$ denote either $\widehat{\varphi}_i$ or $\widetilde{\varphi}_i$, and let $\widehat{\tau}^* = (n+N)^{-1}\sum_{i=1}^{n+N}\widehat{\psi}_i^*$ denote the corresponding estimator. We estimate its asymptotic variance by

$$\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\tau}^*) = \frac{1}{(n+N)(n+N-1)}\sum_{i=1}^{n+N}\big(\widehat{\psi}_i^* - \widehat{\tau}^*\big)^2. \tag{5}$$

We construct a two-sided level-α confidence interval as $\Big[\widehat{\tau}^* \pm \Phi^{-1}(1-\alpha/2)\cdot\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\tau}^*)}\Big]$, where Φ denotes the cumulative distribution function of the standard normal distribution, and $\alpha \in (0,1)$ is the nominal significance level.

2.3 Estimation of the calibration weight function

The EIF-based estimators use the calibration weight $\widehat{\rho}$ as an input. We estimate this weight from the RCT and DT covariates. Because Problem (2) may have multiple population solutions, we select the solution that stays closest to uniform weighting of the DT sample. Let $\mathcal{C}^*$ be the set of population calibration weights that solve Problem (2). We define the selected oracle calibration weight as

$$\rho^* \in \arg\min_{\rho\in\mathcal{C}^*}\mathbb{E}_{P_{\mathrm{X},0}}\big[\{\rho(X)-1\}^2\big]. \tag{6}$$

In the theoretical analysis, the population calibration weight ρ appearing in the target distribution and EIF is understood as this selected oracle weight ρ^*. Since feasible weights satisfy $\mathbb{E}_{P_{\mathrm{X},0}}[\rho(X)] = 1$, this rule is equivalent to minimizing $\mathbb{E}_{P_{\mathrm{X},0}}[\rho(X)^2]$ over $\mathcal{C}^*$. Thus ρ^* satisfies the calibration constraints while avoiding unnecessary concentration of the DT weights.

We estimate ρ^* by solving a penalized empirical version of Problem (2). To ensure positivity and normalization, we use the normalized exponential tilt $\rho_\theta(x) = \exp\{h_\theta(x)\}/\mathbb{E}_{P_{\mathrm{X},0}}[\exp\{h_\theta(X)\}]$, where $h_\theta : \mathcal{X} \to \mathbb{R}$ is bounded. In the sample, the normalizing constant is replaced by its empirical counterpart, $\widehat{\rho}_\theta(x) = \exp\{h_\theta(x)\}/\{N^{-1}\sum_{j=1}^{N}\exp\{h_\theta(\widetilde{X}_j)\}\}$.

We estimate θ by solving

$$\widehat{\theta} \in \arg\min_{\theta}\Big\{\widehat{\Phi}(\widehat{\rho}_\theta) + \beta\widehat{\mathcal{S}}(\widehat{\rho}_\theta) + \gamma\widehat{\mathcal{R}}(\widehat{\rho}_\theta)\Big\}, \tag{7}$$

and set $\widehat{\rho}(x) = \widehat{\rho}_{\widehat{\theta}}(x)$. The discrepancy term is the empirical Wasserstein-1 distance,

$$\widehat{\Phi}(\widehat{\rho}_\theta) = \sup_{\|f\|_L\le 1}\Big\{\frac{1}{n}\sum_{i=1}^{n}f(X_i) - \frac{1}{N}\sum_{j=1}^{N}\widehat{\rho}_\theta(\widetilde{X}_j)f(\widetilde{X}_j)\Big\}.$$

The summary-balance term matches external summaries after accounting for the RCT component of the mixture. For each available summary m_a of $g_a(X)$, define $m_{a,\eta} = \{m_a - \eta\,\mathbb{E}_{P_{\mathrm{X},1}}[g_a(X)]\}/(1-\eta)$ and set $\widehat{\mathcal{S}}(\widehat{\rho}_\theta) = \sum_{a\in\mathcal{A}}\|N^{-1}\sum_{j=1}^{N}\widehat{\rho}_\theta(\widetilde{X}_j)g_a(\widetilde{X}_j) - m_{a,\eta}\|_2^2$. The dispersion term controls concentration of the DT weights, $\widehat{\mathcal{R}}(\widehat{\rho}_\theta) = N^{-1}\sum_{j=1}^{N}\{\widehat{\rho}_\theta(\widetilde{X}_j)-1\}^2$. Because the sample-normalized exponential tilt imposes empirical normalization, this penalty also controls the empirical second moment of the weights. It provides a finite-sample analogue of the minimum-dispersion selection rule in Eq. (6). The tuning parameters β and γ may depend on (n,N), but we suppress this dependence in the notation. Appendix D gives computational details for approximating $\widehat{\Phi}$ and optimizing Eq. (7).

2.4 Broad use of DTs for other causal estimands

The same framework can be adapted to other causal estimands, and we illustrate this extension with subgroup average treatment effects. Let $\{\mathcal{X}_j\}_{j=1}^J$ be a pre-specified partition of the covariate space $\mathcal{X}$. For each subgroup $j \in \{1, \ldots, J\}$ with $P_{\mathtt{X}}(X \in \mathcal{X}_j) > 0$, define

$$\tau_j = \mathbb{E}_P\big[Y(1) - Y(0) \mid X \in \mathcal{X}_j\big].$$

This estimand is the treatment effect within subgroup j in the target population.

Under the mixture representation of $P_{\mathtt{X}}$, the subgroup estimand can be written as

$$\tau_j = \frac{\eta\, \mathbb{E}_{P_{\mathtt{X},1}}\left[(\mu_1(X) - \mu_0(X))\mathbb{1}\{X \in \mathcal{X}_j\}\right] + (1-\eta)\, \mathbb{E}_{P_{\mathtt{X},0}}\left[\rho(X)(\mu_1(X) - \mu_0(X))\mathbb{1}\{X \in \mathcal{X}_j\}\right]}{\eta\, \mathbb{E}_{P_{\mathtt{X},1}}\left[\mathbb{1}\{X \in \mathcal{X}_j\}\right] + (1-\eta)\, \mathbb{E}_{P_{\mathtt{X},0}}\left[\rho(X)\mathbb{1}\{X \in \mathcal{X}_j\}\right]}.$$

Thus, the subgroup average treatment effect restricts the target population ATE to units in $\mathcal{X}_j$ and normalizes by the target subgroup probability. Subgroup estimation requires enough RCT outcome information and enough calibrated target mass within each subgroup.

3 Theoretical investigations

In this section, we study the large-sample behavior of the proposed procedure. The results show that the empirical calibration rule recovers the selected oracle calibration weight. They also characterize the influence-function representation for the target-population ATE and establish consistency, asymptotic normality, and variance consistency for the EIF-based and auxiliary-enhanced estimators.

Assumption 1 (Sampling, transportability, identification, and positivity)**.** *The RCT and DT samples are independent, with sample sizes n and N satisfying $n/(n+N) \to \eta \in (0,1)$. For first-order analysis, the pooled observation is represented by $O = (S, X, ST, SY, Y^\dagger(0), Y^\dagger(1))$ under the limiting mixture law with $\mathbb{P}(S=1) = \eta$, where the conditional law given $S = 1$ is the RCT population P_1 and the conditional law given $S = 0$ is the DT population P_0. The conditional potential outcome distribution is transportable from the RCT to the target population. For $t \in \{0,1\}$, the conditional distribution of $Y(t)$ given X in the target population equals that in the RCT population. In the RCT sample, $T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X, S = 1$. There exists a constant $c > 0$ such that $c \le e(x) \le 1 - c$ and $c \le \pi(x) \le 1 - c$ for all x in the support of the pooled covariate distribution.*

Assumption 1 gives the sampling and causal identification conditions under which the target ATE is identified from randomized outcomes and the calibrated target covariate distribution.

Assumption 2 (Calibration regularity)**.** *The population solution set $\mathcal{C}^*$ of Problem* (2) *is nonempty. The minimum-dispersion oracle calibration weight ρ^* defined in Eq.* (6) *is unique. The target covariate distribution admits the mixture representation in Eq.* (1) *under this calibration weight. The normalized exponential-tilt class is compact and uniformly bounded. The empirical calibration criterion uniformly approximates its population counterpart. The tuning parameters β and γ, which may depend on (n, N), are chosen so that the summary-balance penalty enforces the calibration constraints asymptotically and the weight-dispersion penalty implements the minimum-dispersion selection rule asymptotically.*

Assumption 2 connects the penalized empirical calibration criterion to the selected population weight ρ^*, allowing the original calibration problem to have multiple solutions while requiring a unique minimum-dispersion selection.

Assumption 3 (Nuisance estimation and cross-fitting)**.** *For $t \in \{0,1\}$, the nuisance estimators $\widehat{e}(\cdot)$, $\widehat{\pi}(\cdot)$, $\widehat{\mu}_t(\cdot)$, $\widehat{\mu}^\dagger_{t,1}(\cdot)$, $\widehat{\mu}^\dagger_{t,0}(\cdot)$, and $\widehat{\omega}_t(\cdot)$ are obtained by cross-fitting. The population nuisance functions are uniformly bounded. The estimated nuisance functions satisfy the same boundedness and positivity conditions, when those conditions apply, with probability tending to one. The observed outcome Y and the auxiliary predictions $Y^\dagger(t)$ have finite second moments. Each nuisance estimator is consistent in the L^2 norm corresponding to its source distribution. The nuisance estimation errors also satisfy the product-rate conditions required for root-n inference.*

Assumption 3 states high-level conditions for flexible nuisance estimation. Cross-fitting controls empirical-process effects, and the product-rate conditions make the nuisance-induced remainders second order.

Assumption 4 (First-order calibration remainder)**.** *The first-order effect of estimating the calibration weight is negligible for the target ATE. In particular,*

$$\sqrt{n}\left|\mathbb{E}_{P_{\mathcal{I},o}}\left[\{\widehat{\rho}(X)-\rho^*(X)\}\{\mu_1(X)-\mu_0(X)\}\right]\right| = o_p(1).$$

For the auxiliary-enhanced estimator, the same condition holds after replacing $\mu_1(X)-\mu_0(X)$ with the corresponding auxiliary-enhanced contrast.

Assumption 4 imposes a projected first-order condition on the calibration-weight error. It requires the calibration error to be negligible only along the treatment-effect contrast, not root-n consistent in full L_2 norm. This condition is plausible when calibrated summaries capture covariates that drive treatment-effect heterogeneity, and it excludes errors aligned with that contrast.

Theorem 1 (Consistency of the calibration-weight estimator)**.** *Under Assumption 2,*

$$\|\widehat{\rho}-\rho^*\|_{L^2(P_{\mathcal{I},o})} \xrightarrow{p} 0.$$

Theorem 1 states that the empirical calibration procedure consistently recovers the selected oracle calibration weight.

Lemma 1 (Influence-function representation)**.** *Under Assumptions 1 and 2, and treating ρ^* as fixed, the function $\phi(O)$ defined in Eq.* (4) *is the efficient influence function for the target-population ATE τ and satisfies $\mathbb{E}[\phi(O)]=0$. Replacing μ_t with the calibrated outcome model $m_{t,s}$ gives a mean-zero auxiliary-enhanced influence score $\widetilde{\phi}(O)$ for the same estimand τ.*

Lemma 1 identifies the first-order influence-function representation for the target ATE and verifies that the auxiliary-enhanced score remains centered at the same estimand.

Theorem 2 (Consistency and asymptotic normality)**.** *Suppose Assumptions 1–4 hold. Then the EIF-based estimator satisfies $\widehat{\tau} \xrightarrow{p} \tau$ and*

$$\sqrt{n}(\widehat{\tau}-\tau) \xrightarrow{d} N\big(0, \eta\,\mathbb{V}\mathrm{ar}\{\phi(O)\}\big).$$

The auxiliary-enhanced estimator satisfies $\widetilde{\tau} \xrightarrow{p} \tau$ and

$$\sqrt{n}(\widetilde{\tau}-\tau) \xrightarrow{d} N\big(0, \eta\,\mathbb{V}\mathrm{ar}\{\widetilde{\phi}(O)\}\big).$$

The plug-in variance estimators in Eq. (5) *are consistent, with*

$$n\,\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\tau}) \xrightarrow{p} \eta\,\mathbb{V}\mathrm{ar}\{\phi(O)\}, \quad n\,\widehat{\mathbb{V}\mathrm{ar}}(\widetilde{\tau}) \xrightarrow{p} \eta\,\mathbb{V}\mathrm{ar}\{\widetilde{\phi}(O)\}.$$

Theorem 2 uses the RCT sample size n as the scaling parameter. Since $n/(n+N) \to \eta$, pooled-sample scaling gives the same limit after the corresponding variance rescaling. The two estimators target the same ATE, with asymptotic variances determined by their influence scores.

4 Simulation

In this section, we use simulation to evaluate the proposed estimators in a BRIGHTEN-based setting. The simulation constructs RCT and DT superpopulations, defines an oracle target distribution through calibration, and compares RCT-only, naive target-population, and EIF-based estimators. We focus on bias, standard deviation, and efficiency gains from auxiliary outcome predictions for overall and subgroup target ATEs.

4.1 BRIGHTEN-based simulation design

We base the simulation on the BRIGHTEN study, a fully remote, smartphone-based randomized trial program for depression conducted in the United States between 2016 and 2018 [Pratap et al., 2022]. The outcome is the PHQ-9 score, with higher values indicating greater depressive symptom severity. The baseline variables include gender, education, employment status, marital status, race, age, baseline PHQ-9 total score, and a screening-based psychiatric risk summary.

We construct two superpopulations of size $20{,}000$. For the RCT distribution P_1, we fit a CTGAN to the structured BRIGHTEN covariates and sample synthetic baseline covariates from the fitted generator.

We generate potential outcomes $Y(0)$ and $Y(1)$ using arm-specific random forest models fitted on the BRIGHTEN data. For the DT distribution P_0, we generate synthetic baseline covariates using a local Qwen3-8B digital-twin generator with an output schema aligned to the RCT covariates. The DT sample contains covariates only. It does not contain potential outcomes or observed randomized outcomes. For both RCT and DT units, we use Qwen3-8B to generate treatment-specific auxiliary outcome predictions $Y^\dagger(0)$ and $Y^\dagger(1)$ from baseline covariates and treatment-specific prompts. These auxiliary predictions are used only as auxiliary measurements for efficiency improvement, not as causal outcomes.

We define the target covariate distribution through the mixture representation in Eq. (1). The oracle calibration weight $\rho(x)$ applies a pre-specified tilt to selected baseline covariates, so the target population differs from both the RCT and DT populations. We draw a target sample of size $50{,}000$ from this oracle target distribution and use only its empirical summaries as external calibration constraints. The estimator does not observe individual-level target covariates. The benchmark estimand is the target ATE τ. We estimate its reference value by fitting separate random forest models for $\mu_1(x)$ and $\mu_0(x)$ using the RCT superpopulation, where potential outcomes are available by construction, and then averaging $\mu_1(x) - \mu_0(x)$ over the oracle target covariate distribution. This reference value is a simulation benchmark induced by the superpopulation outcome model.

In each Monte Carlo replication, we independently draw an RCT sample of size n and a DT sample of size N. Treatment is randomized with probability $1/2$ in the RCT sample, and the observed outcome equals the realized potential outcome. We compute all estimators with two-fold cross-fitting and estimate $\rho(\cdot)$ from the target summary constraints. We report empirical bias and Monte Carlo standard error over $B = 500$ replications.

4.2 Estimators considered for comparison

We compare four estimators. The first is an RCT-only benchmark. The second is a naive plug-in target-population estimator that estimates the conditional treatment effect from the RCT and averages it over the calibrated target mixture. The last two are the proposed EIF-based estimators, $\widehat{\tau}$ and $\widetilde{\tau}$, where $\widetilde{\tau}$ further uses the calibrated outcome model.

The RCT-only benchmark is

$$\widehat{\tau}_{\texttt{RCT}} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i\big(Y_i - \widehat{\mu}_1(X_i)\big)}{\widehat{e}(X_i)} - \frac{(1-T_i)\big(Y_i - \widehat{\mu}_0(X_i)\big)}{1-\widehat{e}(X_i)} + \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big).$$

This estimator is the standard augmented inverse-probability-weighted (AIPW) estimator computed only in the trial sample. It targets the trial-population ATE and provides a benchmark for the cost of ignoring covariate shift.

The naive plug-in target-population estimator is

$$\begin{aligned}\widehat{\tau}_{\texttt{naive}} =& \eta\,\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i\big(Y_i - \widehat{\mu}_1(X_i)\big)}{\widehat{e}(X_i)} - \frac{(1-T_i)\big(Y_i - \widehat{\mu}_0(X_i)\big)}{1-\widehat{e}(X_i)} + \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big) \\ &+ (1-\eta)\,\frac{1}{N}\sum_{j=1}^{N}\widehat{\rho}(\widetilde{X}_j)\Big(\widehat{\mu}_1(\widetilde{X}_j) - \widehat{\mu}_0(\widetilde{X}_j)\Big).\end{aligned}$$

This plug-in estimator substitutes the estimated conditional mean functions into the target estimand without adding the full EIF correction for the RCT score contribution.

The proposed estimators are $\widehat{\tau}$ and $\widetilde{\tau}$. The estimator $\widehat{\tau}$ applies the EIF-based score in Eq. (4) and combines calibrated DT weights with the estimated source propensity score. The estimator $\widetilde{\tau}$ adds the calibrated outcome model $m_{t,s}(x, y^\dagger)$. Comparing $\widehat{\tau}_{\texttt{naive}}$ with $\widehat{\tau}$ measures the gain from EIF-based correction, while comparing $\widehat{\tau}$ with $\widetilde{\tau}$ measures the additional gain from auxiliary predictions.

4.3 Simulation results

We evaluate the estimators by varying the common sample size $n = N$ and by considering age-defined target subgroups. We summarize the contribution of auxiliary predictions through

$$\mathrm{EffGain}(\widetilde{\tau}, \widehat{\tau}) = \frac{\mathbb{V}\mathrm{ar}(\widehat{\tau}) - \mathbb{V}\mathrm{ar}(\widetilde{\tau})}{\mathbb{V}\mathrm{ar}(\widehat{\tau})} \times 100\%.$$

Positive values indicate empirical variance reduction from auxiliary outcome prediction. For the naive baseline, we compute the same measure by comparing $\widetilde{\tau}_{\texttt{naive}}$ with $\widehat{\tau}_{\texttt{naive}}$, where $\widetilde{\tau}_{\texttt{naive}}$ replaces $\widehat{\mu}_t(X)$ with $\widehat{m}_{t,s}(X, Y^{\dagger}(t))$ in the naive target-population estimator. Appendix E reports an additional analysis that fixes the RCT sample size and varies the DT sample size.

Figure 1 reports overall target ATE performance when $n = N$ ranges from 2,000 to 5,000. The proposed target-population estimators have small scaled bias across sample sizes, whereas the RCT-only benchmark shows increasing scaled bias due to covariate mismatch. The auxiliary-enhanced estimator $\widetilde{\tau}$ has smaller scaled Monte Carlo standard error than the non-auxiliary EIF estimator $\widehat{\tau}$. Figure 2 reports subgroup efficiency gains for age-defined target ATEs. The proposed auxiliary adjustment gives positive gains across all age groups and sample sizes, and generally outperforms the auxiliary-enhanced naive benchmark.

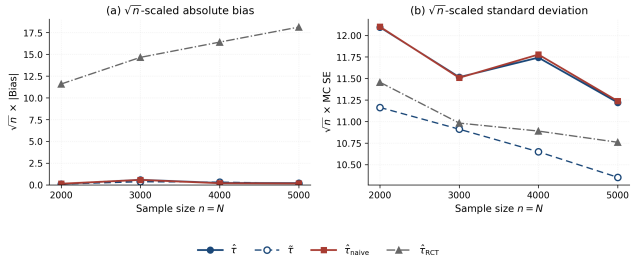

Figure 1: Overall target ATE performance across 500 Monte Carlo replications. Panel (a) reports $\sqrt{n}$-scaled absolute bias, and panel (b) reports $\sqrt{n}$-scaled standard deviation across sample sizes.

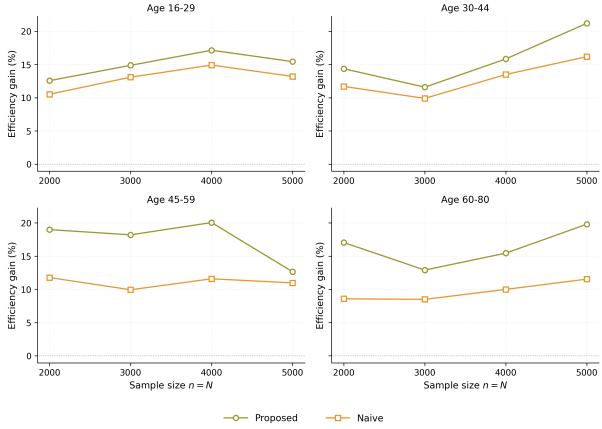

Figure 2: Efficiency gains for age-defined subgroup target ATEs across 500 Monte Carlo replications. Each panel corresponds to one age subgroup. The proposed curve compares $\widetilde{\tau}$ with $\widehat{\tau}$, and the naive curve compares $\widetilde{\tau}_{\texttt{naive}}$ with $\widehat{\tau}_{\texttt{naive}}$.

5 Discussion

We proposed a framework for target-population inference from randomized experiments when individual-level target covariates are unavailable. The framework uses randomized outcomes to identify conditional treatment effects and calibrated digital twins, with external summaries, to

represent the target covariate distribution. DTs therefore define the population over which the RCT-identified effect is averaged, rather than generate causal outcomes. LLM-generated auxiliary predictions enter only through outcome adjustment, so they may improve precision without changing the estimand or identification source.

The main limitation is transportability. The trial and target population must share the same conditional outcome model given observed baseline covariates, and the RCT alone cannot verify this condition. More realistic DT covariates do not guarantee transportability, especially when important effect modifiers are unobserved, omitted from summaries, or poorly represented by DT covariates. Calibrated DTs help only when their support matches the external summaries. Concentrated weights can signal weak overlap and increased variance, so effective sample size is a useful diagnostic. Future work can study weaker conditions using limited target outcomes, multiple trials, or sensitivity models, and can develop diagnostics for calibration quality and weak overlap.

References

Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Katie R Mollan, Paul E Sax, Eric S Daar, Adaora A Adimora, Joseph J Eron, and Michael J Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(4):1193–1209, 2018.

Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1):107–115, 2010.

Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39 (14):1999–2014, 2020.

Trisha Das, Zifeng Wang, and Jimeng Sun. Twin: Personalized clinical trial digital twin generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 402–413, 2023.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, 2023.

Evangelia Katsoulakis, Qi Wang, Huanmei Wu, Leili Shahriyari, Richard Fletcher, Jinwei Liu, Luke Achenie, Hongfang Liu, Pamela Jackson, Ying Xiao, et al. Digital twins for health: a scoping review. *NPJ digital medicine*, 7(1):77, 2024.

Reinhard Laubenbacher, Borna Mehrad, Ilya Shmulevich, and Natalia Trayanova. Digital twins in medicine. *Nature Computational Science*, 4(3):184–191, 2024.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486.

Abhishek Pratap, Ava Homiar, Luke Waninger, Calvin Herd, Christine Suver, Joshua Volponi, Joaquin A Anguera, and Pat Areán. Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. *Scientific Data*, 9(1):522, 2022.

Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386, 2011.

Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. Twin-gpt: digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.