Integrating Digital Twins with Randomized Experiments

Yanping Li
UC Berkeley
yanpingli@berkeley.edu

Xinwei Ma
UC San Diego
x1ma@ucsd.edu

Jingshen Wang*
UC 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 their influence when they conflict with the summaries. The procedure also incorporates LLM-generated auxiliary outcome predictions through calibrated outcome regressions, which can improve precision without changing the estimand. We establish consistency and asymptotic normality of the proposed estimators. In a semisynthetic simulation, they reduce bias relative to RCT-only estimation and attain near-nominal coverage, and the auxiliary predictions 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 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

*Corresponding author

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

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], and Degtiar and Rose [2023] review these methods within the literature on generalizability and transportability. In many applications, however, researchers observe only partial target 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]. Because DTs carry no outcomes, they can help represent target profiles that the trial population also contains, but they cannot identify treatment effects for profiles absent from it.

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. A useful estimator should allow DTs to contribute covariate information while limiting their influence when they conflict with external target summaries. A natural alternative reweights the RCT alone to the external summaries, as in calibration weighting and entropy balancing [Deville and Särndal, 1992, Hainmueller, 2012]. Such reweighting can only redistribute mass over trial covariate profiles. When the summaries leave outcome-relevant target structure unspecified, DT profiles that resemble the target can supply that structure.

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. The procedure also uses LLM-generated auxiliary outcome predictions through calibrated outcome regressions, which can improve 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 derive the efficient influence function of the calibrated-mixture estimand, rather than inserting calibration weights into an existing augmented inverse-probability-weighted (AIPW) formula, and develop estimators that combine randomized outcomes, calibrated DTs, and LLM-generated auxiliary outcome predictions without changing the target estimand. We prove calibration-weight consistency, establish consistency, asymptotic normality, and variance consistency, and evaluate the method in a BRIGHTEN-based simulation that includes a comparison with direct calibration of the trial and sensitivity analyses for the auxiliary predictions.

Section 2 introduces the framework and estimators, Section 3 gives the theory, Section 4 reports the simulation, and Section 5 discusses limitations and future work.

2 Method

In this section, we define the target population ATE and the calibrated mixture representation used to estimate it, present the estimators and the estimation of the calibration weight, and extend the framework to subgroup effects.

2.1 Improving external validity of RCTs with digital twins

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 not external validity, which concerns whether the treatment effect in the trial also applies to the target population. The two ATEs may differ when the trial and target covariate distributions differ and treatment effects vary across covariate-defined subgroups, as when eligibility criteria define a narrow study population or 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 provides no observed outcomes. Besides covariates, it carries auxiliary outcome predictions generated with the same LLM-based 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 the auxiliary outcome predictions. These synthetic observations follow a distribution P_0, whose covariate distribution is intended to approximate the target covariate distribution.

We denote the target population distribution by $P = P_{\mathtt{X}} \cdot P_{\mathtt{Y|X}}$, where $P_{\mathtt{X}}$ is the target covariate distribution and $P_{\mathtt{Y|X}}$ 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 $P_{\mathtt{X},s}$ denote 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, so η records the source composition rather than DT quality. Eq. (1) is a representation of the target distribution. Whenever it holds, different DTs can change $P_{\mathtt{X},0}$, η, and ρ, but not $P_{\mathtt{X}}$ or the target ATE. The weight ρ is normalized so that $\mathbb{E}_{P_{\mathtt{X},0}}[\rho(X)] = 1$, 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 $t \in \{0, 1\}$.

External data sources often report only partial summaries of $P_{\mathtt{X}}$, 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

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

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. For summaries indexed by $\mathcal{A}$, m_a denotes externally reported summary statistics and $g_a : \mathcal{X} \to \mathbb{R}^{d_a}$ denotes the corresponding transformation. The tolerance parameter $\delta \ge 0$ 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

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

This representation separates the roles of the two samples. Randomized RCT outcomes identify μ_t, the calibrated mixture determines the population over which the conditional effect is averaged, and the DT sample supplies no observed outcomes.

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. 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. The predictions never enter the identification of τ.

We estimate the nuisance functions 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 [Chernozhukov et al., 2018], and estimate ρ by the empirical calibration of Section 2.3.

With the calibration weight held fixed, the efficient influence function (EIF) for the target-population ATE based on (S, X, ST, SY) 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 enter 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. Replacing $\mu_t(x)$ with this calibrated outcome model gives $\widetilde{\tau} = (n+N)^{-1}\sum_{i=1}^{n+N}\widetilde{\varphi}_i$, where

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

Because ω_t is estimated from RCT outcomes, the data determine how much the predictions are used. A prediction that is uncorrelated with the outcome given the covariates gives $\omega_t(x) = 0$, and $\widetilde{\tau} = \widehat{\tau}$ when $\widehat{\omega}_0$ and $\widehat{\omega}_1$ are identically zero.

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 $(1-\alpha)$-level confidence interval as $\big[\widehat{\tau}^* \pm z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\tau}^*)}\big]$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile 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. The summaries carry no information about the reliability of individual DT records, so ρ^* reweights the DT sample to match the summaries rather than scoring each record. Low-quality DTs can instead appear as insufficient support, infeasible constraints, or weights that must concentrate on a few records to match the summaries.

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 $\widehat{m}_{a,\eta} = \{m_a - \eta\, n^{-1}\sum_{i=1}^{n}g_a(X_i)\}/(1-\eta)$ and set $\widehat{\mathcal{S}}(\widehat{\rho}_\theta) = \sum_{a\in\mathcal{A}}\|N^{-1}\sum_{j=1}^{N}\widehat{\rho}_\theta(\widetilde{X}_j)g_a(\widetilde{X}_j) - \widehat{m}_{a,\eta}\|_2^2$. 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

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

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

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

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 show that the empirical calibration rule recovers the selected oracle calibration weight, characterize the influence-function representation for the target-population ATE, and establish consistency, asymptotic normality, and variance consistency for both proposed estimators.

Assumption 1 (Sampling, transportability, identification, and positivity)**.** *The RCT and DT samples are independent, with sample sizes n and N satisfying $n/(n+N) \to \eta \in (0,1)$. The pooled observations $O = (S, X, ST, SY, Y^\dagger(0), Y^\dagger(1))$ are independent draws from the 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. For $t \in \{0,1\}$, the conditional distribution of $Y(t)$ given X is the same in the target and RCT populations, and the observed RCT outcome is $Y = TY(1) + (1-T)Y(0)$. In the RCT sample, $T \perp\!\!\!\perp \{Y(0), Y(1), Y^\dagger(0), Y^\dagger(1)\} \mid X, S = 1$. There exists a constant $c > 0$ such that $c \le e(x) \le 1 - c$ and $\pi(x) \ge c$ for all x in the support of the pooled covariate distribution.*

Assumption 1 gives the sampling and causal identification conditions under which randomized outcomes and the calibrated target covariate distribution identify the target ATE. The positivity condition on π requires every DT profile to also occur in the trial population. For a profile absent from the trial, calibration could still match the summaries, but the RCT has no outcomes to identify μ_t there without extrapolation.

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, uniformly bounded, satisfies the uniform entropy condition, and contains ρ^*. 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^*. Problem (2) may have multiple solutions, but the minimum-dispersion selection among them must be unique. These calibration conditions hold when the mixture can match the summaries with well-spread weights, which small calibration residuals and a high effective sample size indicate in practice, as Appendix H shows for the simulation. The requirement that ρ^* represent $P_{\mathtt{X}}$ in Eq. (1)

holds when $P_{\mathtt{X}}$ is itself a feasible mixture that the summaries single out, and τ is then identified. Without this requirement, the results below hold with τ replaced by the summary-constrained target ATE, the ATE under the mixture defined by ρ^*.

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 stated in Appendix C. For $\widetilde{\tau}$, the auxiliary-mean estimators satisfy $\sqrt{n}\,|\mathbb{E}_{P_{\mathtt{I},o}}[\rho^*\omega_t\{(\widehat{\mu}^{\dagger}_{t,1}-\mu^{\dagger}_{t,1})-(\widehat{\mu}^{\dagger}_{t,0}-\mu^{\dagger}_{t,0})\}]| = o_p(1)$ for each t.*

Assumption 3 states high-level conditions for flexible nuisance estimation. Cross-fitting controls empirical-process effects, the product-rate conditions make the nuisance-induced remainders second order, and the last condition makes source-specific centering negligible to first order.

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

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

Assumption 4 lets Theorem 2 apply the influence function of Lemma 1, which treats the calibration weight as known, to the estimators that use $\widehat{\rho}$. It restricts the calibration error only along the treatment effect contrast. To see what it requires, write $\mu_1(x)-\mu_0(x) = \sum_{a\in\mathcal{A}} c_a^{\top} g_a(x) + r(x)$, where r is the part of the contrast that the summary functions do not capture, so that the quantity in Assumption 4 splits into a term along the summary functions and a term along r. The term along r vanishes when $r = 0$. The term along the summary functions need not be negligible, but it only cancels chance fluctuations of the sample averages of these functions, because calibration matches the summaries in every sample. The variance estimator in Eq. (5) still counts these fluctuations, so when $r = 0$ the intervals remain valid and are conservative, as Appendix C shows.

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

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

Theorem 1, proved in Appendix A, shows 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 based on (S, X, ST, SY) 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, derived in Appendix B, treats ρ^* as known. The estimators use $\widehat{\rho}$ instead, and Theorem 1 and Assumption 4 make this replacement negligible to first order, so Theorem 2 applies to the estimators actually computed.

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, proved in Appendix C, uses the RCT sample size n as the scaling parameter. When n and N are set by design rather than drawn from the mixture, as in the simulation, the intervals are asymptotically conservative.

4 Simulation

In this section, we evaluate the proposed estimators in a BRIGHTEN-based simulation through bias, standard deviation, coverage, and efficiency gains for overall and subgroup target ATEs, and we compare the mixture with direct calibration of the trial.

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 [Pratap et al., 2022]. The treatment $T = 1$ is iPST, an app that delivers problem-solving therapy. The control $T = 0$ pools the other two study apps, a cognitive training game and an app with general health tips. The outcome is the score on the Patient Health Questionnaire-9 (PHQ-9), 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 conditional tabular generative adversarial network (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 no 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 all baseline covariates with the prompt in Appendix J, which requests both predictions. These auxiliary predictions are used only as auxiliary measurements for efficiency improvement, not as causal outcomes.

We construct the target covariate distribution as the mixture in Eq. (1) with $\eta = 1/2$ and a data-generating weight that tilts employment status, marital status, and race, so the target population differs from both the RCT and DT populations. We draw a target sample of size $50{,}000$ from this distribution. The external summaries are its proportions in four age bands, 18 to 29, 30 to 44, 45 to 59, and 60 to 80 years, and the estimators observe no individual-level target covariates. The estimators adjust for age band, so the reference value of the target ATE is the RCT age-band effect averaged over the target age-band distribution, which these proportions determine. We compute it from random forest fits of the potential outcomes on age band in the RCT superpopulation, where both potential outcomes are available by construction.

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 cross-fit the outcome models over two folds, use the known treatment probability, and estimate $\rho(\cdot)$ by the penalized calibration in Section 2.3 with $\beta = 500$. Because exact matching of the four age-band proportions determines the band-constant ρ, the dispersion penalty has no competing solutions to select among, and we set $\gamma = 0$, with the remaining settings in Appendix D. We report empirical bias, standard deviation, and coverage of 95% intervals across $B = 500$ 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 $\widetilde{\tau}$.

The RCT-only benchmark $\widehat{\tau}_{\texttt{RCT}} = n^{-1}\sum_{i=1}^{n}\widehat{\psi}_i$ is the standard AIPW estimator computed only in the trial sample, with score

$$\widehat{\psi}_i = \frac{T_i\big(Y_i - \widehat{\mu}_1(X_i)\big)}{\widehat{e}(X_i)} - \frac{(1-T_i)\big(Y_i - \widehat{\mu}_0(X_i)\big)}{1-\widehat{e}(X_i)} + \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)$$

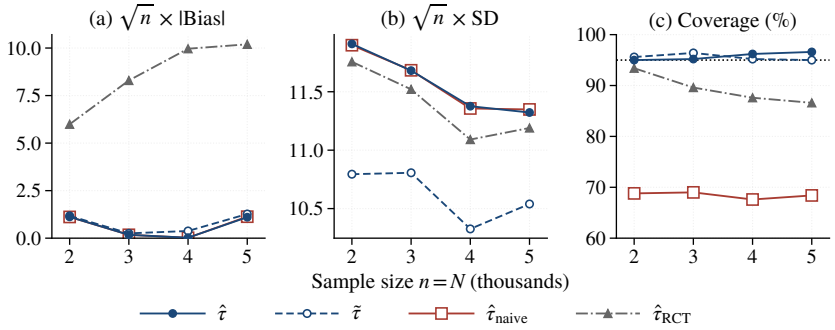


Figure 1: Overall target ATE over 500 replications, with (a) $\sqrt{n}$-scaled absolute bias, (b) $\sqrt{n}$-scaled SD, and (c) coverage of 95% intervals. The curves of $\widehat{\tau}$ and $\widehat{\tau}_{\texttt{naive}}$ nearly coincide in (a) and (b).

for RCT unit i. 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

$$\widehat{\tau}_{\texttt{naive}} = \eta \frac{1}{n} \sum_{i=1}^{n} \widehat{\psi}_i + (1-\eta) \frac{1}{N} \sum_{j=1}^{N} \widehat{\rho}(\widetilde{X}_j)\big(\widehat{\mu}_1(\widetilde{X}_j) - \widehat{\mu}_0(\widetilde{X}_j)\big).$$

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. Its nominal 95% interval uses the sample variances of the two averages, treating $\widehat{\mu}_t$ as fixed.

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. To isolate what the DT profiles add, we also compare $\widehat{\tau}$ with direct calibration of the trial, which averages the scores $\widehat{\psi}_i$ with entropy-balancing weights that reweight the RCT alone to the same summaries [Hainmueller, 2012].

4.3 Simulation results

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

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

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 fixes $n = 2{,}000$ and varies N from 1,000 to 4,000, so η ranges from $2/3$ to $1/3$, and finds stable estimators and efficiency gains of about 15% to 23%.

Figure 1 reports overall target ATE performance. The proposed target-population estimators have small scaled bias across sample sizes, whereas the RCT-only benchmark shows much larger scaled bias, which generally increases with sample size due to covariate mismatch, and its interval coverage falls from 93% to 87%. The auxiliary-enhanced estimator $\widetilde{\tau}$ has smaller scaled standard deviation than $\widehat{\tau}$. The naive estimator has nearly the same bias and standard deviation as $\widehat{\tau}$, but its interval covers the target ATE in only 68% to 69% of replications, whereas the intervals of $\widehat{\tau}$ and $\widetilde{\tau}$ cover it in 95% to 97%, as Appendix F details. Figure 2 reports subgroup efficiency gains for age-defined target ATEs. The proposed auxiliary adjustment gives positive gains, from 8.0% to 25.8%, across all age groups and sample sizes, and generally outperforms the naive benchmark.

The DT profiles reduce bias relative to direct calibration of the trial when the summaries do not determine the target distribution of the covariates in the outcome adjustment. This comparison keeps the DTs within the support of the trial and uses the ATE over the full target covariate distribution.

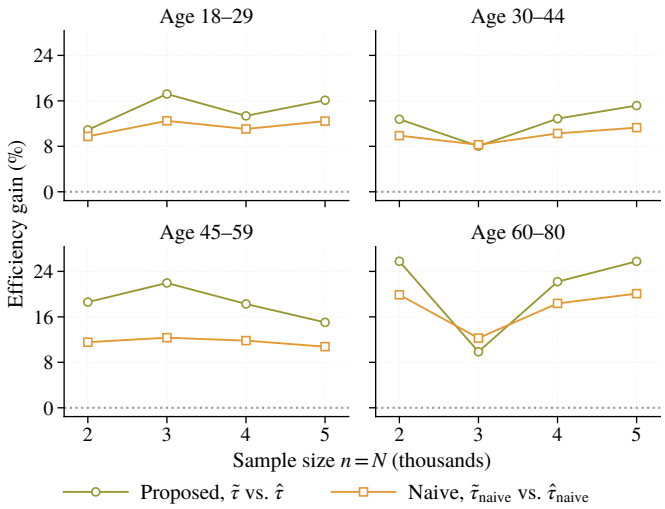


Figure 2: Efficiency gains for the age-band subgroup target ATEs over 500 replications. Positive values indicate variance reduction from the auxiliary predictions.

Direct calibration matches $\widehat{\tau}$ when the outcome adjustment and the summaries use only age band, because both average the same age-band effects. When the adjustment also uses demographics, the DT profiles supply the demographic composition of the target within age bands, which reweighting the trial to the summaries cannot recover. With the age-band proportions as summaries, alone or with the race proportions or the employment, marital-status, and race proportions, the large-sample absolute bias of $\widehat{\tau}$ is then between 0.02 and 0.08, against 0.27 to 0.31 for direct calibration. At $n = N = 5{,}000$, $\widehat{\tau}$ has an absolute bias between 0.15 and 0.21, against 0.28 to 0.33, and its intervals cover this ATE in 88% to 89% of replications, against 62% to 66%, as Appendix G details.

Appendix H checks the support of the DT covariates, the calibration residuals, and the effective sample size of the weights, and finds nearly uniform weights. In Appendix I, zero-shot, few-shot, and GPT-5-mini predictions give efficiency gains of 19.5%, 19.4%, and 18.7%, whereas randomly permuted zero-shot predictions give no gain.

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 define the population over which these effects are averaged, so DTs contribute covariate profiles rather than causal outcomes. Within common support, DTs help when the summaries leave outcome-relevant target structure unspecified, the outcome model adjusts for it, and the DT sample represents it. Otherwise, direct calibration of the trial is the simpler choice. Direct calibration matches the mixture when the summaries give the joint target distribution of the adjustment covariates, which external sources rarely report. LLM-generated auxiliary predictions enter only through outcome adjustment, so they may improve precision without changing the estimand or identification source, in the spirit of prediction-powered inference [Angelopoulos et al., 2023, 2026].

The main limitation is transportability. The trial and target populations must share the same conditional outcome model given observed baseline covariates, which the RCT alone cannot verify, and DTs cannot identify treatment effects outside the support of the trial. The summaries also cannot verify whether the DT profiles resemble the target in the structure they leave unspecified. Our semisynthetic evaluation is built on BRIGHTEN covariates and outcome models and makes the target effects known, but it does not validate the method against an individual-level target cohort. Future work can study weaker conditions using limited target outcomes, multiple trials, or sensitivity models, can develop diagnostics for calibration quality and weak overlap, and can apply the framework to trials whose target populations are described by registry or census summaries.

Acknowledgments and Disclosure of Funding

This work was supported by NIH grants R01AG096488 and R01AG089512 and NSF grant DMS-2239047. The authors declare no competing interests.

References

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *The Annals of Applied Statistics*, 20(3):2235–2252, 2026.

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.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 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 Hernán. 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.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30, 2017.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

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

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of Wasserstein GANs. In *International Conference on Learning Representations*, 2018.

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.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Cédric Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

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.

Technical appendices

A Proof of Theorem 1

Proof. Let Θ index the normalized exponential-tilt class. For $\theta \in \Theta$, let ρ_θ be the population-normalized tilt. Its sample-normalized analogue in Section 2.3 is

$$\widehat{\rho}_\theta(x) = \frac{\rho_\theta(x)}{N^{-1}\sum_{j=1}^{N} \rho_\theta(\widetilde{X}_j)}.$$

Write the empirical criterion in Eq. (7) as

$$\widehat{M}_n(\theta) = \widehat{\Phi}(\widehat{\rho}_\theta) + \beta\widehat{\mathcal{S}}(\widehat{\rho}_\theta) + \gamma\widehat{\mathcal{R}}(\widehat{\rho}_\theta),$$

and let $M_n(\theta)$ denote its population counterpart, which replaces the sample averages by expectations and $\widehat{\rho}_\theta$ by ρ_θ and uses the same tuning parameters β and γ. Both criteria may depend on (n, N) through β and γ. By Assumption 2, there is $\theta^* \in \Theta$ with $\rho_{\theta^*} = \rho^*$. All norms in this proof are those of $L^2(P_{\mathtt{X},0})$.

We use the last two conditions of Assumption 2 in the following form. There are constants $s_n > 0$ such that

$$\sup_{\theta\in\Theta} \left|\widehat{M}_n(\theta) - M_n(\theta)\right| = o_p(s_n),$$

and, for every $\epsilon > 0$, there is $c_\epsilon > 0$ such that, for all large n,

$$M_n(\theta) - M_n(\theta^*) \geq c_\epsilon s_n \quad \text{whenever} \quad \|\rho_\theta - \rho^*\| \geq \epsilon.$$

The second condition states that the penalties enforce the calibration constraints and select the minimum-dispersion weight ρ^* by a margin that the approximation error in the first condition cannot offset. When γ decreases to zero to break ties among the solutions of Problem (2), s_n is of the order of γ, so the approximation error must vanish faster than γ.

Let $\widehat{\theta}$ satisfy $\widehat{M}_n(\widehat{\theta}) \leq \widehat{M}_n(\theta^*) + o_p(s_n)$. An exact minimizer in Eq. (7) satisfies this condition, and so does an approximate minimizer whose optimization error is $o_p(s_n)$. Fix $\epsilon > 0$. For large n, on the event $\|\rho_{\widehat{\theta}} - \rho^*\| \geq \epsilon$,

$$\begin{aligned} c_\epsilon s_n &\leq M_n(\widehat{\theta}) - M_n(\theta^*) \\ &= \{M_n(\widehat{\theta}) - \widehat{M}_n(\widehat{\theta})\} + \{\widehat{M}_n(\widehat{\theta}) - \widehat{M}_n(\theta^*)\} + \{\widehat{M}_n(\theta^*) - M_n(\theta^*)\} \\ &\leq 2\sup_{\theta\in\Theta} \left|\widehat{M}_n(\theta) - M_n(\theta)\right| + \{\widehat{M}_n(\widehat{\theta}) - \widehat{M}_n(\theta^*)\}. \end{aligned}$$

Both terms on the right are $o_p(s_n)$, by the first condition and by the choice of $\widehat{\theta}$. The probability of the event therefore tends to zero, so $\|\rho_{\widehat{\theta}} - \rho^*\| \to 0$ in probability.

It remains to pass from $\rho_{\widehat{\theta}}$ to $\widehat{\rho} = \widehat{\rho}_{\widehat{\theta}}$. The tilt class is bounded by a constant C and satisfies the uniform entropy condition, so it is a Glivenko–Cantelli class [van der Vaart, 1998, Chapter 19]. Since $\mathbb{E}_{P_{\mathtt{X},0}}[\rho_\theta(X)] = 1$ for every θ, it follows that $\sup_{\theta\in\Theta} |N^{-1}\sum_{j=1}^{N} \rho_\theta(\widetilde{X}_j) - 1| \to 0$ in probability, and hence $\sup_{\theta\in\Theta}\sup_x |\widehat{\rho}_\theta(x) - \rho_\theta(x)| \to 0$ in probability. Therefore

$$\|\widehat{\rho} - \rho^*\| \leq \|\widehat{\rho}_{\widehat{\theta}} - \rho_{\widehat{\theta}}\| + \|\rho_{\widehat{\theta}} - \rho^*\| \xrightarrow{p} 0.$$

□

B Proof of Lemma 1

Proof. Throughout this proof, we treat the selected oracle calibration weight ρ^* as fixed and write it as ρ for notational simplicity. We derive the efficient influence function for the pooled observation (S, X, ST, SY) of Assumption 1, without the auxiliary predictions, whose source indicator satisfies $\mathbb{P}(S = 1) = \eta$. Because ρ is fixed, the submodels leave $f_{\mathtt{DT}}$ unrestricted. The mixing proportion η is a parameter of this observed-data distribution, and the target distribution in Eq. (1) uses the same η.

We start with the data generating process. Let $(S, X, S \cdot T, S \cdot Y)$ be a representative observation from the mixture distribution. That is,

$$\mathbb{P}[S = 1] = \mathbb{P}[\text{data belongs to RCT}] = \eta.$$

Conditioning on $S = 1$ and $S = 0$, the distributions of the covariates, treatment, and outcome satisfy

$$X \sim f_{\texttt{DT}}(x) \mid S = 0,$$

and

$$X \sim f_{\texttt{RCT}}(x) \mid S = 1, \quad T \sim \texttt{Bernoulli}(e(X)) \mid X, S = 1,$$
$$Y \sim g_1(y|X)^T g_0(y|X)^{1-T} \mid T, X, S = 1.$$

In the above, $f_{\texttt{DT}}$ and $f_{\texttt{RCT}}$ denote the covariate densities in the two sources, g_1 and g_0 denote the conditional densities of the potential outcomes given the covariates, and e is the propensity score. The variables T and Y are only available in the RCT data. As in Section 2, $\mu_1(X) = \mathbb{E}[Y(1) \mid X, S = 1]$ and $\mu_0(X) = \mathbb{E}[Y(0) \mid X, S = 1]$. The parameter of interest is the average treatment effect for the target population,

$$\tau = \int \Big(\mu_1(x) - \mu_0(x)\Big)\Big(\eta f_{\texttt{RCT}}(x) + (1-\eta)\rho(x) f_{\texttt{DT}}(x)\Big)\mathrm{d}x.$$

The density $f_{\texttt{DT}}$, unlike ρ, needs to be estimated from the digital twin sample.

To compute the efficiency bound, we first write the joint distribution of the data, which is simply

$$p(s,x,t,y) = \Big\{(1-\eta) f_{\texttt{DT}}(x)\Big\}^{1-s}\Big\{\eta e(x) f_{\texttt{RCT}}(x) g_1(y|x)\Big\}^{st}\Big\{\eta(1-e(x)) f_{\texttt{RCT}}(x) g_0(y|x)\Big\}^{s(1-t)}.$$

Now consider a parametric submodel indexed by ϵ. It takes the form

$$\begin{aligned} p(s,x,t,y;\epsilon) = &\Big\{(1-\eta_\epsilon) f_{\texttt{DT}}(x;\epsilon)\Big\}^{1-s}\Big\{\eta_\epsilon e(x;\epsilon) f_{\texttt{RCT}}(x;\epsilon) g_1(y|x;\epsilon)\Big\}^{st} \\ &\times \Big\{\eta_\epsilon(1-e(x;\epsilon)) f_{\texttt{RCT}}(x;\epsilon) g_0(y|x;\epsilon)\Big\}^{s(1-t)}, \end{aligned}$$

or alternatively, we write the log-likelihood as

$$\begin{aligned} \log p(s,x,t,y;\epsilon) &= (1-s)\log\Big\{(1-\eta_\epsilon) f_{\texttt{DT}}(x;\epsilon)\Big\} + (st)\log\Big\{\eta_\epsilon e(x;\epsilon) f_{\texttt{RCT}}(x;\epsilon) g_1(y|x;\epsilon)\Big\} \\ &\quad + s(1-t)\log\Big\{\eta_\epsilon(1-e(x;\epsilon)) f_{\texttt{RCT}}(x;\epsilon) g_0(y|x;\epsilon)\Big\} \\ &= (1-s)\log(1-\eta_\epsilon) + (1-s)\log f_{\texttt{DT}}(x;\epsilon) + s\log\eta_\epsilon + s\log f_{\texttt{RCT}}(x;\epsilon) \\ &\quad + (st)\log e(x;\epsilon) + s(1-t)\log(1-e(x;\epsilon)) \\ &\quad + (st)\log g_1(y|x;\epsilon) + s(1-t)\log g_0(y|x;\epsilon). \end{aligned}$$

Step 1. We first describe the tangent space. It consists of six subspaces, corresponding to the source proportion, the covariate distributions in the two sources, the propensity score, and the conditional distributions of the potential outcomes. Specifically, we have

$$\begin{aligned} \mathcal{T}_S &= \Big\{(s-\eta)c : \ c \in \mathbb{R}\Big\} \\ \mathcal{T}_{X,\texttt{DT}} &= \Big\{(1-s)\mathfrak{f}_{\texttt{DT}}(x) : \ \int \mathfrak{f}_{\texttt{DT}}(x) f_{\texttt{DT}}(x)\mathrm{d}x = 0\Big\} \\ \mathcal{T}_{X,\texttt{RCT}} &= \Big\{s\mathfrak{f}_{\texttt{RCT}}(x) : \ \int \mathfrak{f}_{\texttt{RCT}}(x) f_{\texttt{RCT}}(x)\mathrm{d}x = 0\Big\} \\ \mathcal{T}_T &= \Big\{s\Big(\frac{t}{e(x)} - \frac{1-t}{1-e(x)}\Big)\mathfrak{e}(x) : \ \mathfrak{e}(x) \text{ general function}\Big\} \\ \mathcal{T}_{Y(0)} &= \Big\{s(1-t)\mathfrak{g}_0(y|x) : \ \int \mathfrak{g}_0(y|x) g_0(y|x)\mathrm{d}y = 0\Big\} \\ \mathcal{T}_{Y(1)} &= \Big\{st\mathfrak{g}_1(y|x) : \ \int \mathfrak{g}_1(y|x) g_1(y|x)\mathrm{d}y = 0\Big\}. \end{aligned}$$

We remark that the above tangent subspaces are mutually orthogonal.

Step 2. We next compute the pathwise derivative. We write the parameter of interest in terms of the observed variables as

$$\tau = \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X) + (1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big].$$

The expectation is taken with respect to the distribution in the combined data, that is, the mixture of $f_{\texttt{RCT}}$ and $f_{\texttt{DT}}$ with the mixing probability η and $1-\eta$. Again, with ϵ denoting a parametric submodel, the above can be written as

$$\tau_\epsilon = \mathbb{E}_\epsilon\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X;\epsilon)} - \frac{(1-T)Y}{1-e(X;\epsilon)}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X;\epsilon) + (1-\eta)\rho(X)f_{\texttt{DT}}(X;\epsilon)}{f_{\texttt{RCT}}(X;\epsilon)}\Big].$$

To compute the pathwise derivative, that is, $\mathrm{d}\tau_\epsilon/\mathrm{d}\epsilon$, it is helpful to be precise about the parametric submodels. We first consider submodels with $\eta_\epsilon = \eta$ and treat the direction that varies η after the summary below. Such submodels are constructed by considering variations of the form

$$\begin{aligned}
f_{\texttt{DT}}(x;\epsilon) &= f_{\texttt{DT}}(x)\Big(1+\epsilon\mathfrak{f}_{\texttt{DT}}(x)\Big)\\
f_{\texttt{RCT}}(x;\epsilon) &= f_{\texttt{RCT}}(x)\Big(1+\epsilon\mathfrak{f}_{\texttt{RCT}}(x)\Big)\\
e(x;\epsilon) &= e(x) + \epsilon\mathfrak{e}(x)\\
g_0(y|x;\epsilon) &= g_0(y|x)\Big(1+\epsilon\mathfrak{g}_0(y|x)\Big)\\
g_1(y|x;\epsilon) &= g_1(y|x)\Big(1+\epsilon\mathfrak{g}_1(y|x)\Big).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\mathrm{d}\tau_\epsilon}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} &= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X) + (1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{s}(S,X,T,Y)\Big]\\
&\quad+\mathbb{E}\Big[\frac{S}{\eta}\Big(-\frac{TY\mathfrak{e}(X)}{e(X)^2} - \frac{(1-T)Y\mathfrak{e}(X)}{(1-e(X))^2}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X) + (1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big]\\
&\quad+\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\Big(\frac{\eta f_{\texttt{RCT}}(X)\mathfrak{f}_{\texttt{RCT}}(X)}{f_{\texttt{RCT}}(X)}\\
&\qquad\qquad-\frac{\Big(\eta f_{\texttt{RCT}}(X) + (1-\eta)\rho(X)f_{\texttt{DT}}(X)\Big)f_{\texttt{RCT}}(X)\mathfrak{f}_{\texttt{RCT}}(X)}{f_{\texttt{RCT}}(X)^2}\Big)\Big]\\
&\quad+\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)\mathfrak{f}_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big]\\
&= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X) + (1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{s}(S,X,T,Y)\Big] \quad \text{(I)}\\
&\quad+\mathbb{E}\Big[\frac{S}{\eta}\Big(-\frac{TY}{e(X)^2} - \frac{(1-T)Y}{(1-e(X))^2}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X) + (1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{e}(X)\Big] \quad \text{(II)}\\
&\quad+\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\Big(-\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big)\mathfrak{f}_{\texttt{RCT}}(X)\Big] \quad \text{(III)}\\
&\quad+\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{f}_{\texttt{DT}}(X)\Big]. \quad \text{(IV)}
\end{aligned}$$

To simplify, we start with (II), which by the law of iterated expectation, takes the form,

$$
\begin{aligned}
\text{(II)} &= -\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)^2}+\frac{(1-T)Y}{(1-e(X))^2}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{e}(X)\Big]\\
&= -\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{T\mu_1(X)}{e(X)^2}+\frac{(1-T)\mu_0(X)}{(1-e(X))^2}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{e}(X)\Big]\\
&= -\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{T\mu_1(X)}{e(X)}-\frac{(1-T)\mu_0(X)}{1-e(X)}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\\
&\qquad\cdot S\Big(\frac{T}{e(X)}-\frac{1-T}{1-e(X)}\Big)\mathfrak{e}(X)\Big]\\
&= -\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{(T-e(X))\mu_1(X)}{e(X)}-\frac{(1-T-(1-e(X)))\mu_0(X)}{1-e(X)}\Big)\\
&\qquad\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\cdot S\Big(\frac{T}{e(X)}-\frac{1-T}{1-e(X)}\Big)\mathfrak{e}(X)\Big]\\
&= -\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{(T-e(X))\mu_1(X)}{e(X)}-\frac{(1-T-(1-e(X)))\mu_0(X)}{1-e(X)}\Big)\\
&\qquad\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\cdot\mathfrak{s}(S,X,T,Y)\Big].
\end{aligned}
$$

Combining (I) and (II) gives

$$
\begin{aligned}
\text{(I) + (II)} &= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)}-\frac{(1-T)Y}{1-e(X)}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{s}(S,X,T,Y)\Big]\\
&-\mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{(T-e(X))\mu_1(X)}{e(X)}-\frac{(1-T-(1-e(X)))\mu_0(X)}{1-e(X)}\Big)\\
&\qquad\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\cdot\mathfrak{s}(S,X,T,Y)\Big]\\
&= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{T\Big(Y-\mu_1(X)\Big)}{e(X)}-\frac{(1-T)\Big(Y-\mu_0(X)\Big)}{1-e(X)}\Big)\\
&\qquad\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{s}(S,X,T,Y)\Big]\\
&+\mathbb{E}\Big[\frac{S}{\eta}\Big(\mu_1(X)-\mu_0(X)\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\cdot\mathfrak{s}(S,X,T,Y)\Big]\\
&= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{T\Big(Y-\mu_1(X)\Big)}{e(X)}-\frac{(1-T)\Big(Y-\mu_0(X)\Big)}{1-e(X)}\Big)\\
&\qquad\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{s}(S,X,T,Y)\Big]\\
&+\mathbb{E}\Big[\Big(\frac{S}{\eta}\Big(\mu_1(X)-\mu_0(X)\Big)\cdot\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}-\tau\Big)\cdot\mathfrak{s}(S,X,T,Y)\Big].
\end{aligned}
$$

We define

$$
\begin{aligned}
\phi_{(\mathrm{I})}(s,x,t,y) &= \frac{s}{\eta}\Big(\frac{t\big(y-\mu_1(x)\big)}{e(x)}-\frac{(1-t)\big(y-\mu_0(x)\big)}{1-e(x)}\Big)\cdot\frac{\eta f_{\texttt{RCT}}(x)+(1-\eta)\rho(x)f_{\texttt{DT}}(x)}{f_{\texttt{RCT}}(x)},\\
\phi_{(\mathrm{II})}(s,x,t,y) &= \frac{s}{\eta}\Big(\mu_1(x)-\mu_0(x)\Big)\cdot\frac{\eta f_{\texttt{RCT}}(x)+(1-\eta)\rho(x)f_{\texttt{DT}}(x)}{f_{\texttt{RCT}}(x)}-\tau.
\end{aligned}
$$

The function $\phi_{(\mathrm{I})}(s,x,t,y)$ belongs to $\mathcal{T}_{Y(0)}\oplus\mathcal{T}_{Y(1)}$, because its conditional mean given $(X,T,S=1)$ is zero.

Next, we analyze (III). To start, we write

$$\begin{aligned}
\text{(III)} &= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\Big(-\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big)\mathfrak{f}_{\texttt{RCT}}(X)\Big] \\
&= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\Big(-\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big)\cdot\Big(S\mathfrak{f}_{\texttt{RCT}}(X)\Big)\Big] \\
&= \mathbb{E}\Big[\frac{S}{\eta}\Big(\mu_1(X)-\mu_0(X)\Big)\Big(-\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big)\cdot\Big(S\mathfrak{f}_{\texttt{RCT}}(X)\Big)\Big] \\
&= \mathbb{E}\Big[\frac{S}{\eta}\Big(\mu_1(X)-\mu_0(X)\Big)\Big(-\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\Big)\cdot\mathfrak{s}(S,X,T,Y)\Big].
\end{aligned}$$

Term (IV) is slightly more complicated to analyze. Observe that

$$\begin{aligned}
\text{(IV)} &= \mathbb{E}\Big[\frac{S}{\eta}\Big(\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\Big)\cdot\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{f}_{\texttt{DT}}(X)\Big] \\
&= \mathbb{E}\Big[\frac{f_{\texttt{RCT}}(X)}{\eta f_{\texttt{RCT}}(X)+(1-\eta)f_{\texttt{DT}}(X)}\Big(\mu_1(X)-\mu_0(X)\Big)\cdot\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{f_{\texttt{RCT}}(X)}\mathfrak{f}_{\texttt{DT}}(X)\Big] \\
&= \mathbb{E}\Big[\Big(\mu_1(X)-\mu_0(X)\Big)\cdot\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{\eta f_{\texttt{RCT}}(X)+(1-\eta)f_{\texttt{DT}}(X)}\mathfrak{f}_{\texttt{DT}}(X)\Big] \\
&= \mathbb{E}\Big[(1-S)\frac{\eta f_{\texttt{RCT}}(X)+(1-\eta)f_{\texttt{DT}}(X)}{(1-\eta)f_{\texttt{DT}}(X)}\Big(\mu_1(X)-\mu_0(X)\Big) \\
&\qquad\cdot\frac{(1-\eta)\rho(X)f_{\texttt{DT}}(X)}{\eta f_{\texttt{RCT}}(X)+(1-\eta)f_{\texttt{DT}}(X)}\mathfrak{f}_{\texttt{DT}}(X)\Big] \\
&= \mathbb{E}\Big[\frac{1-S}{1-\eta}\Big(\mu_1(X)-\mu_0(X)\Big)\cdot(1-\eta)\rho(X)\mathfrak{f}_{\texttt{DT}}(X)\Big] \\
&= \mathbb{E}\Big[\frac{1-S}{1-\eta}\Big(\mu_1(X)-\mu_0(X)\Big)\cdot(1-\eta)\rho(X)\Big((1-S)\mathfrak{f}_{\texttt{DT}}(X)\Big)\Big] \\
&= \mathbb{E}\Big[\frac{1-S}{1-\eta}\Big(\mu_1(X)-\mu_0(X)\Big)\cdot(1-\eta)\rho(X)\mathfrak{s}(S,X,T,Y)\Big].
\end{aligned}$$

We therefore define

$$\phi_{(\text{III})}(s,x,t,y) = (1-\eta)\rho(x)\Big(\mu_1(x)-\mu_0(x)\Big)\Big(\frac{1-s}{1-\eta} - \frac{s}{\eta}\frac{f_{\texttt{DT}}(x)}{f_{\texttt{RCT}}(x)}\Big).$$

The function $\phi_{(\text{III})}(s,x,t,y)$ has mean zero.

To summarize, the efficient influence function is

$$\begin{aligned}
&\phi(s,x,t,y) = \phi_{(\text{I})}(s,x,t,y) + \phi_{(\text{II})}(s,x,t,y) + \phi_{(\text{III})}(s,x,t,y) \\
&= \frac{s}{\eta}\Big(\frac{t\big(y-\mu_1(x)\big)}{e(x)} - \frac{(1-t)\big(y-\mu_0(x)\big)}{1-e(x)}\Big)\cdot\Big(\eta+(1-\eta)\rho(x)\frac{f_{\texttt{DT}}(x)}{f_{\texttt{RCT}}(x)}\Big) \\
&\quad + \frac{s}{\eta}\Big(\mu_1(x)-\mu_0(x)\Big)\cdot\Big(\eta+(1-\eta)\rho(x)\frac{f_{\texttt{DT}}(x)}{f_{\texttt{RCT}}(x)}\Big) - \tau \\
&\quad + (1-\eta)\rho(x)\Big(\mu_1(x)-\mu_0(x)\Big)\Big(\frac{1-s}{1-\eta} - \frac{s}{\eta}\frac{f_{\texttt{DT}}(x)}{f_{\texttt{RCT}}(x)}\Big) \\
&= \frac{s}{\eta}\Big(\frac{t\big(y-\mu_1(x)\big)}{e(x)} - \frac{(1-t)\big(y-\mu_0(x)\big)}{1-e(x)}\Big)\cdot\Big(\eta+(1-\eta)\rho(x)\frac{f_{\texttt{DT}}(x)}{f_{\texttt{RCT}}(x)}\Big) \\
&\quad + \Big(s+(1-s)\rho(x)\Big)\Big(\mu_1(x)-\mu_0(x)\Big) - \tau \\
&= s\Big(\frac{t\big(y-\mu_1(x)\big)}{e(x)} - \frac{(1-t)\big(y-\mu_0(x)\big)}{1-e(x)}\Big)\cdot\Big(1+\rho(x)\frac{1-\pi(x)}{\pi(x)}\Big) \\
&\quad + \Big(s+(1-s)\rho(x)\Big)\Big(\mu_1(x)-\mu_0(x)\Big) - \tau,
\end{aligned}$$

where $\pi(x) = \mathbb{P}[S = 1|X = x]$, which is the conditional probability of an observation belonging to the RCT subpopulation.

The computation above represents the pathwise derivative along submodels with η fixed. Let $\kappa_1 = \mathbb{E}[\mu_1(X) - \mu_0(X) \mid S = 1]$ and $\kappa_0 = \mathbb{E}[\rho(X)\{\mu_1(X) - \mu_0(X)\} \mid S = 0]$, so that $\tau = \eta\kappa_1 + (1-\eta)\kappa_0$. Along a submodel $\eta_\epsilon = \eta + \epsilon\mathfrak{n}$ with the other components fixed, the score is $\mathfrak{n}(s-\eta)/\{\eta(1-\eta)\}$ and $\mathrm{d}\tau_\epsilon/\mathrm{d}\epsilon = \mathfrak{n}(\kappa_1 - \kappa_0)$. Since $\mathbb{E}[\phi(O) \mid S = 1] = \kappa_1 - \tau$ and $\mathbb{E}[\phi(O) \mid S = 0] = \kappa_0 - \tau$, we have $\mathbb{E}[\phi(O)(S-\eta)] = \eta(1-\eta)(\kappa_1 - \kappa_0)$, so ϕ also represents this derivative. Moreover, the last expression in the display decomposes ϕ as

$$\begin{aligned}\phi(s,x,t,y) = \phi_{(\mathrm{I})}(s,x,t,y) &+ s\big\{\mu_1(x) - \mu_0(x) - \kappa_1\big\} \\ &+ (1-s)\big\{\rho(x)\big(\mu_1(x) - \mu_0(x)\big) - \kappa_0\big\} + (\kappa_1 - \kappa_0)(s - \eta).\end{aligned}$$

The four terms lie in $\mathcal{T}_{Y(0)} \oplus \mathcal{T}_{Y(1)}$, $\mathcal{T}_{X,\mathtt{RCT}}$, $\mathcal{T}_{X,\mathtt{DT}}$, and $\mathcal{T}_S$, respectively, and each has mean zero. Hence ϕ lies in the tangent space, represents the pathwise derivative along every submodel, and is the efficient influence function, with $\mathbb{E}[\phi(O)] = 0$.

It is also useful to consider the first-order derivative of the efficient influence function with respect to the nuisance parameters. We have

$$\frac{\partial\phi(s,x,t,y)}{\partial e(x)} = \frac{s}{\eta}\Big(-\frac{t\big(y-\mu_1(x)\big)}{e(x)^2} - \frac{(1-t)\big(y-\mu_0(x)\big)}{(1-e(x))^2}\Big)\cdot\Big(\eta + (1-\eta)\rho(x)\frac{f_{\mathtt{DT}}(x)}{f_{\mathtt{RCT}}(x)}\Big),$$

which has mean zero given X, because $\mathbb{E}[T\{Y - \mu_1(X)\} \mid X, S = 1] = \mathbb{E}[(1-T)\{Y - \mu_0(X)\} \mid X, S = 1] = 0$. Similarly,

$$\begin{aligned}\frac{\partial\phi(s,x,t,y)}{\partial\mu_1(x)} &= \frac{s}{\eta}\,\frac{e(x)-t}{e(x)}\Big(\eta + (1-\eta)\rho(x)\frac{f_{\mathtt{DT}}(x)}{f_{\mathtt{RCT}}(x)}\Big) + (1-\eta)\rho(x)\Big(\frac{1-s}{1-\eta} - \frac{s}{\eta}\frac{f_{\mathtt{DT}}(x)}{f_{\mathtt{RCT}}(x)}\Big), \\ \frac{\partial\phi(s,x,t,y)}{\partial\mu_0(x)} &= \frac{s}{\eta}\,\frac{e(x)-t}{1-e(x)}\Big(\eta + (1-\eta)\rho(x)\frac{f_{\mathtt{DT}}(x)}{f_{\mathtt{RCT}}(x)}\Big) - (1-\eta)\rho(x)\Big(\frac{1-s}{1-\eta} - \frac{s}{\eta}\frac{f_{\mathtt{DT}}(x)}{f_{\mathtt{RCT}}(x)}\Big),\end{aligned}$$

both of which have mean zero given X, because $\mathbb{E}[T \mid X, S = 1] = e(X)$ and, since $\pi(X) = \eta f_{\mathtt{RCT}}(X)/\{\eta f_{\mathtt{RCT}}(X) + (1-\eta)f_{\mathtt{DT}}(X)\}$,

$$\mathbb{E}\Big[\frac{1-S}{1-\eta} - \frac{S}{\eta}\frac{f_{\mathtt{DT}}(X)}{f_{\mathtt{RCT}}(X)}\,\Big|\,X\Big] = \frac{1-\pi(X)}{1-\eta} - \frac{\pi(X)}{\eta}\frac{f_{\mathtt{DT}}(X)}{f_{\mathtt{RCT}}(X)} = 0.$$

Finally,

$$\begin{aligned}\frac{\partial\phi(s,x,t,y)}{\partial(f_{\mathtt{DT}}(x)/f_{\mathtt{RCT}}(x))} &= \frac{s}{\eta}\Big(\frac{t\big(y-\mu_1(x)\big)}{e(x)} - \frac{(1-t)\big(y-\mu_0(x)\big)}{1-e(x)}\Big)(1-\eta)\rho(x) \\ &\quad + \frac{s}{\eta}\Big(\mu_1(x) - \mu_0(x)\Big)\cdot(1-\eta)\rho(x) \\ &\quad + (1-\eta)\rho(x)\Big(\mu_1(x) - \mu_0(x)\Big)\Big(-\frac{s}{\eta}\Big) \\ &= \frac{s}{\eta}\Big(\frac{t\big(y-\mu_1(x)\big)}{e(x)} - \frac{(1-t)\big(y-\mu_0(x)\big)}{1-e(x)}\Big)(1-\eta)\rho(x),\end{aligned}$$

which again has mean zero given X. Hence ϕ is Neyman orthogonal with respect to e, μ_1, μ_0, and the density ratio.

For estimation, we employ the efficient influence function above, leading to

$$\begin{aligned}\widehat{\tau} = \frac{1}{n+N}\sum_{i=1}^{n+N}\Bigg\{ &S_i\Big(\frac{T_i\big(Y_i - \widehat{\mu}_1(X_i)\big)}{\widehat{e}(X_i)} - \frac{(1-T_i)\big(Y_i - \widehat{\mu}_0(X_i)\big)}{1-\widehat{e}(X_i)}\Big)\cdot\Big(1 + \widehat{\rho}(X_i)\frac{1-\widehat{\pi}(X_i)}{\widehat{\pi}(X_i)}\Big) \\ &+ \Big(S_i + (1-S_i)\widehat{\rho}(X_i)\Big)\Big(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big)\Bigg\}.\end{aligned}$$

It remains to verify the second claim of Lemma 1. Let $\widetilde{\phi}$ be ϕ with $\mu_t(X)$ replaced by the calibrated outcome model $m_{t,S}(X, Y^\dagger(t)) = \mu_t(X) + \omega_t(X)\{Y^\dagger(t) - \mu^\dagger_{t,S}(X)\}$. By the definition of $\mu^\dagger_{t,s}$,

$\mathbb{E}[m_{t,S}(X, Y^\dagger(t)) \mid X, S] = \mu_t(X)$, so the term $\{S + (1-S)\rho(X)\}\{m_{1,S} - m_{0,S}\}$ has mean τ. By Assumption 1, T is independent of $\{Y(1), Y^\dagger(1)\}$ given $(X, S = 1)$, so

$$\mathbb{E}\big[T\{Y - m_{1,1}(X, Y^\dagger(1))\} \mid X, S = 1\big] = e(X)\,\mathbb{E}\big[Y(1) - m_{1,1}(X, Y^\dagger(1)) \mid X, S = 1\big] = 0,$$

and the same argument applies to the control arm. Hence $\mathbb{E}[\widetilde{\phi}(O)] = 0$ for any bounded coefficient ω_t, and $\widetilde{\phi}$ is centered at the same estimand τ.

□

C Proof of Theorem 2

Proof. We first prove the result for the EIF-based estimator. Write the score in Eq. (4) as a function of the nuisance functions $\nu = (e, \pi, \mu_0, \mu_1)$ and the calibration weight,

$$\psi(O; \nu, \rho) = S\,R_\nu(O)\Big\{1 + \rho(X)\frac{1-\pi(X)}{\pi(X)}\Big\} + \{S + (1-S)\rho(X)\}\{\mu_1(X) - \mu_0(X)\},$$

where $R_\nu(O) = T\{Y - \mu_1(X)\}/e(X) - (1-T)\{Y - \mu_0(X)\}/\{1 - e(X)\}$. With ν^* the population nuisance functions and $\psi^* = \psi(\cdot\,; \nu^*, \rho^*)$, Lemma 1 gives $\phi = \psi^* - \tau$. The pooled sample is split into folds, and for an observation in fold k the estimator uses $\widehat{\psi}_k = \psi(\cdot\,; \widehat{\nu}_k, \widehat{\rho})$, where $\widehat{\nu}_k$ is fitted without fold k. Let P_n and $P_{n,k}$ denote the empirical distributions of the pooled sample and of fold k, which has n_k observations. For a function f that may depend on the estimates, let Pf denote its integral against the distribution of a new observation, with the estimated functions held fixed. Since $P\psi^* = \tau$,

$$\widehat{\tau} - \tau = (P_n - P)\psi^* + \sum_k \frac{n_k}{n+N}(P_{n,k} - P)(\widehat{\psi}_k - \psi^*) + \sum_k \frac{n_k}{n+N} P(\widehat{\psi}_k - \psi^*).$$

We show that the first term is asymptotically normal and that the other two terms are $o_p(n^{-1/2})$. Throughout, $\|\cdot\|_1$ and $\|\cdot\|_0$ denote the norms of $L^2(P_{\mathtt{X},1})$ and $L^2(P_{\mathtt{X},0})$, and C denotes a generic constant. Because $\pi \geq c$, the density ratio of $P_{\mathtt{X},0}$ to $P_{\mathtt{X},1}$ is bounded, so $\|g\|_0 \leq C\|g\|_1$ for every function g.

Leading term. Under Assumptions 1 and 3, $\phi(O)$ has finite second moment, so the central limit theorem gives $\sqrt{n+N}\,(P_n - P)\psi^* \xrightarrow{d} N(0, \mathbb{V}\mathrm{ar}\{\phi(O)\})$. Since $n/(n+N) \to \eta$, it follows that

$$\sqrt{n}\,(P_n - P)\psi^* \xrightarrow{d} N\big(0, \eta\,\mathbb{V}\mathrm{ar}\{\phi(O)\}\big).$$

Bias term. Fix the estimated functions, write $\delta_t = \widehat{\mu}_t - \mu_t$ for $t \in \{0, 1\}$, and let $\widehat{W} = 1 + \widehat{\rho}(1-\widehat{\pi})/\widehat{\pi}$ and $\widetilde{W} = 1 + \widehat{\rho}(1-\pi)/\pi$, with all functions evaluated at X. By Assumption 1,

$$\mathbb{E}\{R_{\widehat{\nu}}(O) \mid X, S = 1\} = -\frac{e}{\widehat{e}}\,\delta_1 + \frac{1-e}{1-\widehat{e}}\,\delta_0.$$

Since $\tau = \mathbb{E}[\{S + (1-S)\rho^*\}(\mu_1 - \mu_0)]$, this gives

$$\begin{aligned} P(\widehat{\psi}_k - \psi^*) = \mathbb{E}\Big[S\Big\{\widehat{W}\Big(-\frac{e}{\widehat{e}}\,\delta_1 + \frac{1-e}{1-\widehat{e}}\,\delta_0\Big) + \delta_1 - \delta_0\Big\}\Big] \\ + \mathbb{E}\{(1-S)\widehat{\rho}(\delta_1 - \delta_0)\} + \mathbb{E}\{(1-S)(\widehat{\rho} - \rho^*)(\mu_1 - \mu_0)\}. \end{aligned}$$

The identity $\mathbb{E}\{(1-S)f(X)\} = \mathbb{E}[Sf(X)\{1-\pi(X)\}/\pi(X)]$ turns the second expectation into $\mathbb{E}\{S(\widetilde{W} - 1)(\delta_1 - \delta_0)\}$, so

$$P(\widehat{\psi}_k - \psi^*) = \mathbb{E}\Big[S\Big\{\delta_1\Big(\widetilde{W} - \widehat{W}\frac{e}{\widehat{e}}\Big) - \delta_0\Big(\widetilde{W} - \widehat{W}\frac{1-e}{1-\widehat{e}}\Big)\Big\}\Big] + (1-\eta)\,\mathbb{E}_{P_{\mathtt{X},0}}\{(\widehat{\rho} - \rho^*)(\mu_1 - \mu_0)\}.$$

Since $\widetilde{W} - \widehat{W} = \widehat{\rho}(\widehat{\pi} - \pi)/(\pi\widehat{\pi})$, the two factors in parentheses equal $\widehat{\rho}(\widehat{\pi} - \pi)/(\pi\widehat{\pi}) + \widehat{W}(\widehat{e} - e)/\widehat{e}$ and $\widehat{\rho}(\widehat{\pi} - \pi)/(\pi\widehat{\pi}) - \widehat{W}(\widehat{e} - e)/(1 - \widehat{e})$. By positivity, boundedness, and the Cauchy–Schwarz inequality, the first expectation is therefore at most $C\sum_t \|\delta_t\|_1(\|\widehat{\pi} - \pi\|_1 + \|\widehat{e} - e\|_1)$. For $\widehat{\tau}$, the product-rate conditions of Assumption 3 are

$$\|\widehat{\mu}_t - \mu_t\|_1\big(\|\widehat{\pi} - \pi\|_1 + \|\widehat{e} - e\|_1\big) = o_p(n^{-1/2}), \qquad t \in \{0, 1\},$$

which make the first expectation $o_p(n^{-1/2})$. Assumption 4 makes the second term $o_p(n^{-1/2})$.

Empirical-process term. Because ψ is affine in ρ,

$$\widehat{\psi}_k - \psi^* = \{\psi(\cdot\,; \widehat{\nu}_k, \rho^*) - \psi^*\} + (\widehat{\rho} - \rho^*)\, a_{\widehat{\nu}_k},$$

where

$$a_\nu(O) = S\, R_\nu(O) \frac{1-\pi(X)}{\pi(X)} + (1-S)\{\mu_1(X) - \mu_0(X)\}.$$

Conditional on the observations outside fold k, the first term is a fixed function whose $L^2(P)$ norm tends to zero in probability, by the L^2 consistency, boundedness, and positivity in Assumption 3 and the finite second moment of Y. Chebyshev's inequality then gives $(P_{n,k} - P)\{\psi(\cdot\,; \widehat{\nu}_k, \rho^*) - \psi^*\} = o_p(n^{-1/2})$.

For the second term, write $\widehat{\rho} = c_N \rho_{\widehat{\theta}}$ with $c_N = \{N^{-1} \sum_{j=1}^N \rho_{\widehat{\theta}}(\widetilde{X}_j)\}^{-1}$, which tends to one in probability by the proof of Theorem 1, and $\|\rho_{\widehat{\theta}} - \rho^*\|_0 \to 0$ in probability. Then

$$(\widehat{\rho} - \rho^*)\, a_{\widehat{\nu}_k} = c_N(\rho_{\widehat{\theta}} - \rho^*)\, a_{\nu^*} + c_N(\rho_{\widehat{\theta}} - \rho^*)(a_{\widehat{\nu}_k} - a_{\nu^*}) + (c_N - 1)\rho^* a_{\widehat{\nu}_k}.$$

The last term contributes $(c_N - 1)\, O_p(n^{-1/2}) = o_p(n^{-1/2})$. For the first term, the class of functions $(\rho - \rho^*)a_{\nu^*}$, with ρ in the tilt class, has a square-integrable envelope and inherits the uniform entropy condition, so it is a Donsker class [van der Vaart, 1998, Chapter 19]. Moreover, $P\{(\rho_{\widehat{\theta}} - \rho^*)^2 a_{\nu^*}^2\} \to 0$ in probability, because the factor $(1-\pi)/\pi \le 1/c$ turns the RCT part of this integral into an integral against $P_{\mathtt{x},0}$, the weights are bounded, and $\|\rho_{\widehat{\theta}} - \rho^*\|_0 \to 0$. Lemma 19.24 of van der Vaart [1998] then gives $(P_n - P)\{(\rho_{\widehat{\theta}} - \rho^*)a_{\nu^*}\} = o_p(n^{-1/2})$. For the middle term, $g_k = a_{\widehat{\nu}_k} - a_{\nu^*}$ is fixed conditional on the observations outside fold k, and its $L^2(P)$ norm tends to zero in probability by the argument for the first term above. The functions $(\rho - \rho^*)g_k$, with ρ in the tilt class, have envelope $2C|g_k|$ and a uniform entropy integral bounded by that of the tilt class. The maximal inequality for such classes [van der Vaart, 1998, Section 19.6] bounds the conditional expectation of $\sup_\rho |(P_{n,k} - P)\{(\rho - \rho^*)g_k\}|$ by $C n_k^{-1/2} \|g_k\|_{L^2(P)}$, which is $o_p(n^{-1/2})$. Hence the empirical-process term is $o_p(n^{-1/2})$.

Conclusion for $\widehat{\tau}$. Combining the three terms,

$$\sqrt{n}(\widehat{\tau} - \tau) = \sqrt{n}\,(P_n - P)\psi^* + o_p(1) \xrightarrow{d} N\big(0, \eta\, \mathbb{V}\mathrm{ar}\{\phi(O)\}\big),$$

and in particular $\widehat{\tau} \xrightarrow{p} \tau$. For the variance, let $k(i)$ denote the fold of observation i. The same bounds show that $(n+N)^{-1} \sum_i \{\widehat{\psi}_{k(i)}(O_i) - \psi^*(O_i)\}^2 \to 0$ in probability. The cross-fitted part has conditional mean tending to zero, and the part involving $\widehat{\rho}$ is controlled by $c_N \to 1$, $\|\rho_{\widehat{\theta}} - \rho^*\|_0 \to 0$, and the Glivenko–Cantelli property of the functions $(\rho - \rho^*)^2 a_{\nu^*}^2$. Hence the empirical variance of the estimated score contributions converges to $\mathbb{V}\mathrm{ar}\{\phi(O)\}$. Since the variance estimator in Eq. (5) estimates the variance of a sample average over $n + N$ observations,

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

Auxiliary-enhanced estimator. The score $\widetilde{\psi}$ of $\widetilde{\tau}$ is ψ with μ_t replaced by $m_{t,S}$, and its nuisance functions add ω_t and $\mu_{t,s}^\dagger$ to ν. The score is again affine in ρ, and it has finite second moment because Y and $Y^\dagger(t)$ have finite second moments and ω_t is bounded. The arguments for the leading, empirical-process, and variance terms therefore apply unchanged, and only the bias term changes. Write $\zeta_{t,s} = \widehat{\mu}_{t,s}^\dagger - \mu_{t,s}^\dagger$ and $u_t = \delta_t - \widehat{\omega}_t \zeta_{t,1}$. Since T is independent of $\{Y(t), Y^\dagger(t)\}$ given $(X, S=1)$, the conditional mean of the estimated residual is the expression above with δ_t replaced by u_t. The conditional mean of $\widehat{m}_{1,s} - \widehat{m}_{0,s}$ given $(X, S=s)$ equals $\mu_1 - \mu_0 + u_1 - u_0$ for $s=1$, and for $s=0$ it contains the additional term $\sum_t (2t-1)\widehat{\omega}_t(\zeta_{t,1} - \zeta_{t,0})$. The same substitutions give

$$\begin{aligned} P(\widetilde{\psi}_k - \widetilde{\psi}^*) = \mathbb{E}\Big[S\Big\{u_1\Big(\widetilde{W} - \widehat{W}\frac{e}{\widehat{e}}\Big) - u_0\Big(\widetilde{W} - \widehat{W}\frac{1-e}{1-\widehat{e}}\Big)\Big\}\Big] \\ + (1-\eta)\, \mathbb{E}_{P_{\mathtt{x},0}}\Big\{\widehat{\rho} \sum_{t\in\{0,1\}} (2t-1)\, \widehat{\omega}_t\, (\zeta_{t,1} - \zeta_{t,0})\Big\} \\ + (1-\eta)\, \mathbb{E}_{P_{\mathtt{x},0}}\{(\widehat{\rho} - \rho^*)(\mu_1 - \mu_0)\}. \end{aligned}$$

The first term is bounded as before, with $\|u_t\|_1 \le \|\delta_t\|_1 + C\|\zeta_{t,1}\|_1$. In the second term, replacing $\widehat{\rho}\widehat{\omega}_t$ by $\rho^*\omega_t$ changes it by at most $C\sum_{t,s}\|\zeta_{t,s}\|_s(\|\widehat{\rho}-\rho^*\|_0 + \|\widehat{\omega}_t - \omega_t\|_1)$, and the term with $\rho^*\omega_t$ is $o_p(n^{-1/2})$ by the auxiliary-mean condition in Assumption 3. For $\widetilde{\tau}$, the product-rate conditions therefore also require, for $t \in \{0,1\}$,

$$\|\zeta_{t,1}\|_1\big(\|\widehat{\pi}-\pi\|_1 + \|\widehat{e}-e\|_1\big) + \sum_{s\in\{0,1\}}\|\zeta_{t,s}\|_s\big(\|\widehat{\rho}-\rho^*\|_0 + \|\widehat{\omega}_t-\omega_t\|_1\big) = o_p(n^{-1/2}).$$

With these conditions, the argument for $\widehat{\tau}$ gives

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

Assumption 1 represents the pooled sample as independent draws from the mixture. If n and N are instead set by design, as in the simulation, the leading term is the sum of two independent sample averages. Its limiting variance is then $\eta\{\eta\mathbb{V}\mathrm{ar}(\phi(O)\mid S=1) + (1-\eta)\mathbb{V}\mathrm{ar}(\phi(O)\mid S=0)\}$, which is smaller than $\eta\mathbb{V}\mathrm{ar}\{\phi(O)\}$ by $\eta^2(1-\eta)(\kappa_1-\kappa_0)^2$, with κ_1 and κ_0 as in Appendix B. The variance estimator in Eq. (5) still converges to $\eta\mathbb{V}\mathrm{ar}\{\phi(O)\}$, so the resulting intervals are conservative, and the same holds for $\widetilde{\tau}$.

Assumption 4 enters only through the second term of the bias expansion, and the variance argument above does not use it. Suppose instead that $\mu_1-\mu_0 = \sum_{a\in\mathcal{A}} c_a^\top g_a$, the case $r=0$ discussed after Assumption 4, which includes the simulation of Section 4, and that the calibrated pooled sample matches the summaries exactly, $\eta n^{-1}\sum_{i=1}^n g_a(X_i) + (1-\eta)N^{-1}\sum_{j=1}^N \widehat{\rho}(\widetilde{X}_j)g_a(\widetilde{X}_j) = m_a$ for every $a\in\mathcal{A}$, with $\eta = n/(n+N)$. Since ρ^* also matches the summaries, $\tau = \sum_{a\in\mathcal{A}} c_a^\top m_a$, so the average of $\{S_i + (1-S_i)\widehat{\rho}(X_i)\}\{\mu_1(X_i)-\mu_0(X_i)\}$ over the pooled sample equals τ. The expansion above then gives $\widehat{\tau}-\tau = (n+N)^{-1}\sum_{i=1}^{n+N}\phi_{(\mathrm{I})}(O_i) + o_p(n^{-1/2})$, with $\phi_{(\mathrm{I})}$ as in Appendix B. The other components of ϕ are functions of (S,X) and are uncorrelated with $\phi_{(\mathrm{I})}$, so the limiting variance of $\sqrt{n}(\widehat{\tau}-\tau)$ is at most $\eta\mathbb{V}\mathrm{ar}\{\phi(O)\}$, the limit of $n\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\tau})$, whether n and N are random or set by design. The intervals are therefore conservative, and the same holds for $\widetilde{\tau}$, whose additional terms have mean zero given (S,X). □

D Computational details for calibration

This appendix provides implementation details for the empirical calibration problem used to estimate the DT calibration weight. The goal is to approximate the Wasserstein discrepancy in Eq. (2), impose the external summary constraints, and stabilize the resulting DT weights.

Recall that the target covariate distribution is represented as

$$dP_{\mathtt{X}}(x) = \eta\, dP_{\mathtt{X},1}(x) + (1-\eta)\rho(x)\, dP_{\mathtt{X},0}(x),$$

where $P_{\mathtt{X},1}$ is the RCT covariate distribution, $P_{\mathtt{X},0}$ is the DT covariate distribution, and ρ is a normalized calibration weight for the DT component. We estimate ρ within the normalized exponential-tilt class

$$\rho_\theta(x) = \frac{\exp\{h_\theta(x)\}}{\mathbb{E}_{P_{\mathtt{X},0}}[\exp\{h_\theta(X)\}]},$$

where $h_\theta : \mathcal{X} \to \mathbb{R}$ is a bounded function. In the sample, we use the empirical normalization

$$\widehat{\rho}_\theta(\widetilde{X}_j) = \frac{\exp\{h_\theta(\widetilde{X}_j)\}}{N^{-1}\sum_{k=1}^N \exp\{h_\theta(\widetilde{X}_k)\}}, \quad j = 1,\dots,N.$$

This parameterization imposes empirical normalization, $N^{-1}\sum_{j=1}^N \widehat{\rho}_\theta(\widetilde{X}_j) = 1$, and positivity of the DT weights.

Empirical calibration objective. The empirical calibration objective has three components,

$$\widehat{\Phi}(\widehat{\rho}_\theta) + \beta\widehat{\mathcal{S}}(\widehat{\rho}_\theta) + \gamma\widehat{\mathcal{R}}(\widehat{\rho}_\theta).$$

The first component, $\widehat{\Phi}$, measures the distributional discrepancy between the RCT covariate distribution and the reweighted DT covariate distribution. We use the Wasserstein-1 discrepancy. By Kantorovich–Rubinstein duality [Villani, 2009], its population version can be written as

$$\Phi(\rho_\theta) = \sup_{\|f\|_L \le 1}\left\{\mathbb{E}_{P_{\mathtt{X},1}}[f(X)] - \mathbb{E}_{P_{\mathtt{X},0}}[\rho_\theta(X)f(X)]\right\},$$

where the supremum is over all 1-Lipschitz functions. Because the RCT component of the mixture cancels in this dual form, Φ equals the objective of Problem (2) divided by $1-\eta$, a constant factor absorbed into β and γ. The corresponding empirical version is

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

The second component, $\widehat{\mathcal{S}}$, penalizes violations of the available external target summaries. Suppose the external information contains summary statistics m_a for transformations $g_a(X)$, $a \in \mathcal{A}$. Under the mixture representation,

$$m_a = \mathbb{E}_{P_{\mathrm{X}}}[g_a(X)] = \eta \mathbb{E}_{P_{\mathrm{X},1}}[g_a(X)] + (1-\eta)\mathbb{E}_{P_{\mathrm{X},0}}[\rho(X) g_a(X)].$$

Thus the DT component should target

$$m_{a,\eta} = \frac{m_a - \eta \mathbb{E}_{P_{\mathrm{X},1}}[g_a(X)]}{1-\eta}.$$

In implementation, we use the sample analogue

$$\widehat{m}_{a,\eta} = \frac{m_a - \eta n^{-1}\sum_{i=1}^{n} g_a(X_i)}{1-\eta},$$

and define

$$\widehat{\mathcal{S}}(\widehat{\rho}_\theta) = \sum_{a \in \mathcal{A}} \left\| \frac{1}{N}\sum_{j=1}^{N} \widehat{\rho}_\theta(\widetilde{X}_j) g_a(\widetilde{X}_j) - \widehat{m}_{a,\eta} \right\|_2^2.$$

This term encourages the calibrated mixture distribution to match the available target summaries.

The third component, $\widehat{\mathcal{R}}$, controls weight dispersion,

$$\widehat{\mathcal{R}}(\widehat{\rho}_\theta) = \frac{1}{N}\sum_{j=1}^{N} \{\widehat{\rho}_\theta(\widetilde{X}_j) - 1\}^2.$$

Because the weights are empirically normalized, this is equivalent to controlling their empirical second moment up to an additive constant. This term stabilizes finite-sample estimation and mirrors the minimum-dispersion oracle selection rule in Eq. (6) when the population calibration problem has multiple solutions.

Combining these components, we estimate

$$\widehat{\theta} \in \arg\min_{\theta} \left\{ \widehat{\Phi}(\widehat{\rho}_\theta) + \beta \widehat{\mathcal{S}}(\widehat{\rho}_\theta) + \gamma \widehat{\mathcal{R}}(\widehat{\rho}_\theta) \right\}, \quad \widehat{\rho}(x) = \widehat{\rho}_{\widehat{\theta}}(x).$$

The tuning parameters β and γ control the strength of summary balance and weight stabilization, respectively.

Neural approximation of the Wasserstein discrepancy. To approximate the supremum in $\widehat{\Phi}$, we use a neural critic f_w. The empirical discrepancy is then approximated by

$$\max_{w} \left[\frac{1}{n}\sum_{i=1}^{n} f_w(X_i) - \frac{1}{N}\sum_{j=1}^{N} \widehat{\rho}_\theta(\widetilde{X}_j) f_w(\widetilde{X}_j) - \lambda \widehat{\mathcal{Q}}(f_w) \right],$$

where $\widehat{\mathcal{Q}}(f_w)$ imposes approximate Lipschitz continuity of the critic and $\lambda \ge 0$ controls the strength of this regularization. Thus the implemented full-sample objective is

$$\min_{\theta} \left\{ \max_{w} \left[\frac{1}{n}\sum_{i=1}^{n} f_w(X_i) - \frac{1}{N}\sum_{j=1}^{N} \widehat{\rho}_\theta(\widetilde{X}_j) f_w(\widetilde{X}_j) - \lambda \widehat{\mathcal{Q}}(f_w) \right] + \beta \widehat{\mathcal{S}}(\widehat{\rho}_\theta) + \gamma \widehat{\mathcal{R}}(\widehat{\rho}_\theta) \right\}.$$

The critic parameters w maximize the discrepancy between the RCT sample and the reweighted DT sample, whereas the reweighting parameters θ reduce this discrepancy while matching the external summaries and discouraging weight concentration.

Algorithm 1 Penalized Wasserstein calibration with two-time-scale updates

Require: Penalty coefficients $\lambda, \beta, \gamma \geq 0$; number of critic updates per reweighting update $n_{\texttt{critic}}$; batch sizes B_r, B_o; Adam stepsizes $(\alpha_w, \alpha_\theta)$ with $\alpha_w > \alpha_\theta$.

Require: Initial critic parameters w_0 and initial reweighting parameters θ_0.

while θ has not converged **do**

  for $k = 1$ to $n_{\texttt{critic}}$ **do**

    Sample $\{x_i\}_{i=1}^{B_r}$ from the RCT covariates and $\{\widetilde{x}_j\}_{j=1}^{B_o}$ from the DT covariates.

    Compute minibatch-normalized weights

$$\widehat{\rho}_\theta(\widetilde{x}_j) = \frac{\exp\{h_\theta(\widetilde{x}_j)\}}{B_o^{-1}\sum_{\ell=1}^{B_o}\exp\{h_\theta(\widetilde{x}_\ell)\}}, \quad j = 1, \ldots, B_o.$$

    Compute the critic objective

$$\widehat{\mathcal{L}}_f = \frac{1}{B_r}\sum_{i=1}^{B_r} f_w(x_i) - \frac{1}{B_o}\sum_{j=1}^{B_o}\widehat{\rho}_\theta(\widetilde{x}_j) f_w(\widetilde{x}_j) - \lambda\widehat{\mathcal{Q}}(f_w).$$

    Update $w \leftarrow w + \alpha_w \texttt{AdamStep}(\nabla_w \widehat{\mathcal{L}}_f)$ and keep θ fixed.

  end for

  Compute the reweighting objective

$$\widehat{\mathcal{L}}_\rho = -\frac{1}{B_o}\sum_{j=1}^{B_o}\widehat{\rho}_\theta(\widetilde{x}_j) f_w(\widetilde{x}_j) + \beta\widehat{\mathcal{S}}(\widehat{\rho}_\theta) + \gamma\widehat{\mathcal{R}}(\widehat{\rho}_\theta).$$

  Update $\theta \leftarrow \theta - \alpha_\theta \texttt{AdamStep}(\nabla_\theta \widehat{\mathcal{L}}_\rho)$ and keep w fixed.

end while

Recompute $\widehat{\rho}_{\widehat{\theta}}(\widetilde{X}_j)$ using the full DT sample normalization, $j = 1, \ldots, N$.

Lipschitz regularization. We consider three standard ways to control the Lipschitz constant of the critic. The first is the gradient penalty of WGAN-GP [Gulrajani et al., 2017],

$$\widehat{\mathcal{Q}}_{\mathrm{GP}}(f_w) = \mathbb{E}_{\widehat{X}}\left(\|\nabla_{\widehat{X}} f_w(\widehat{X})\|_2 - 1\right)^2,$$

where $\widehat{X} = \epsilon X^{(r)} + (1-\epsilon)X^{(o)}$, $X^{(r)} \sim P_{\mathtt{X},1}$, $X^{(o)} \sim P_{\mathtt{X},0}$, and $\epsilon \sim \mathrm{Unif}(0,1)$. The second is the one-sided penalty of WGAN-LP [Petzka et al., 2018],

$$\widehat{\mathcal{Q}}_{\mathrm{LP}}(f_w) = \mathbb{E}_{\widehat{X}}\left(\max\{0, \|\nabla_{\widehat{X}} f_w(\widehat{X})\|_2 - 1\}\right)^2,$$

which penalizes only gradients larger than one. The third is spectral normalization [Miyato et al., 2018], which rescales each weight matrix W by its spectral norm $\sigma(W)$, i.e., $\overline{W} = W/\sigma(W)$. Under spectral normalization, we set the explicit critic penalty to zero.

Optimization. We optimize the objective by alternating critic updates and reweighting updates under a two-time-scale update rule [Heusel et al., 2017]. With θ fixed, the critic is updated by ascending in

$$\widehat{\mathcal{L}}_f = \frac{1}{n}\sum_{i=1}^{n} f_w(X_i) - \frac{1}{N}\sum_{j=1}^{N}\widehat{\rho}_\theta(\widetilde{X}_j) f_w(\widetilde{X}_j) - \lambda\widehat{\mathcal{Q}}(f_w).$$

With w fixed, the reweighting parameter is updated by descending in

$$\widehat{\mathcal{L}}_\rho = -\frac{1}{N}\sum_{j=1}^{N}\widehat{\rho}_\theta(\widetilde{X}_j) f_w(\widetilde{X}_j) + \beta\widehat{\mathcal{S}}(\widehat{\rho}_\theta) + \gamma\widehat{\mathcal{R}}(\widehat{\rho}_\theta).$$

The first term updates the DT weights to reduce the critic discrepancy, while the remaining two terms impose external summary balance and discourage weight concentration.

Full-sample objective and minibatch implementation. The theoretical analysis corresponds to the full-sample normalized weights and the full-sample objective above. Minibatch approximations can

reduce the computation for large samples. When using minibatches, we normalize the DT weights within the DT minibatch,

$$\widehat{\rho}_\theta(\widetilde{x}_j) = \frac{\exp\{h_\theta(\widetilde{x}_j)\}}{B_o^{-1}\sum_{\ell=1}^{B_o}\exp\{h_\theta(\widetilde{x}_\ell)\}}, \quad j = 1, \ldots, B_o.$$

This minibatch normalization approximates the full-sample normalization and keeps the average minibatch DT weight equal to one during training. In minibatch training, $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{R}}$ are computed on the current DT minibatch. After training, we recompute the final weights using the full DT sample normalization and use these full-sample weights in the downstream EIF-based estimators.

Settings in the simulation. In Section 4, the inputs to h_θ are the age-band indicators, so the weights are constant within age bands. The four age bands are 18 to 29, 30 to 44, 45 to 59, and 60 to 80 years. The tilt network has one hidden layer with 32 units and its output is truncated to $[-1.5, 1.5]$. The critic has two hidden layers with 64 units, and we use the gradient penalty with $\lambda = 20$. Each of the 1,000 reweighting updates follows eight critic updates, with Adam step sizes 5×10^{-4} for the critic and 10^{-4} for the reweighting network. All updates use the full RCT and DT samples, and the gradient penalty is evaluated at up to 2,048 random interpolates. The tolerance $\delta = 0.02$ only flags replications whose summary residuals exceed it. In the estimators, the treatment probability is the known value $1/2$, and the nuisance functions fitted on the RCT are cross-fitted over two folds. The source propensity score π is estimated by the age-band frequencies of the pooled sample, and the DT auxiliary means $\mu^\dagger_{t,0}$ are fitted on the DT units of the fold in which they are evaluated. Both fits take at most four distinct values, one per age band, so they lie in a finite-dimensional class for which sample splitting is not needed. We use $\beta = 500$ and $\gamma = 0$, because the age-band constraints determine a band-constant weight and leave no competing solutions for the dispersion penalty to select among.

E Sensitivity to DT sample size

We repeat the overall target ATE experiment with the RCT sample size fixed at $n = 2{,}000$ and the DT sample size N varying from 1,000 to 4,000.

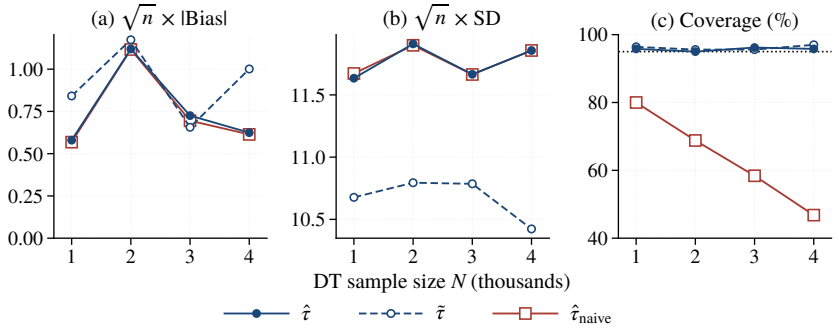


Figure 3: Overall target ATE with fixed RCT sample size $n = 2{,}000$ over 500 replications, with (a) $\sqrt{n}$-scaled absolute bias, (b) $\sqrt{n}$-scaled SD, and (c) coverage of 95% intervals as the DT sample size varies.

Here $\eta = n/(n+N)$ ranges from $2/3$ to $1/3$, while the target population remains the one constructed in Section 4. The age-band reference value does not depend on η either, because it averages the age-band effects over the target age-band proportions. Figure 3 shows that the target-population estimators remain stable as the DT sample size changes. The scaled biases are small relative to the dispersion scale. The main difference appears in the scaled standard deviation. The auxiliary-enhanced estimator $\widetilde{\tau}$ is consistently less variable than the non-auxiliary EIF estimator $\widehat{\tau}$, while $\widehat{\tau}_{\texttt{naive}}$ has a dispersion similar to $\widehat{\tau}$. The intervals of $\widehat{\tau}$ and $\widetilde{\tau}$ cover the target ATE in 95% to 97% of replications, whereas the coverage of the naive intervals falls from 80% to 47% as N increases.

Figure 4 reports the corresponding efficiency gain from auxiliary adjustment. The gain is positive for all DT sample sizes and is consistently larger for the proposed EIF-based estimator than for the naive

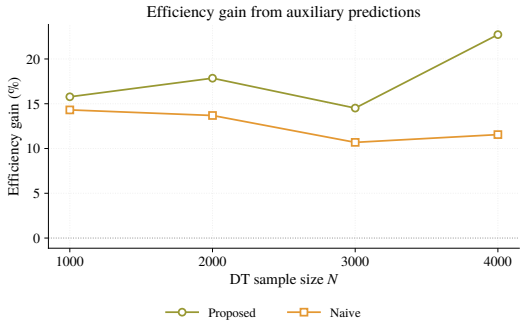


Figure 4: Efficiency gain from auxiliary adjustment for the overall target ATE with fixed RCT sample size $n = 2{,}000$ across 500 Monte Carlo replications. The proposed curve compares $\widetilde{\tau}$ with $\widehat{\tau}$, and the naive curve compares the auxiliary-enhanced naive estimator with the non-auxiliary naive estimator.

benchmark. The gain of the proposed estimator ranges from about 15% to 23%, so the precision improvement from auxiliary predictions persists across these DT sample sizes.

F Coverage diagnostics for EIF-based estimators

The variance of the naive plug-in estimator treats $\widehat{\mu}_t$ as fixed, and Figure 1(c) shows that the resulting intervals under-cover the overall target ATE. We therefore report detailed inferential diagnostics only for the proposed EIF-based estimators $\widehat{\tau}$ and $\widetilde{\tau}$. For the overall ATE, their variance estimators follow Theorem 2. For an age-band subgroup j, the estimator divides the subgroup sum of the same score contributions by the subgroup's calibrated target mass $\sum_i \mathbb{1}\{X_i \in \mathcal{X}_j\}\{S_i + (1 - S_i)\widehat{\rho}(X_i)\}$, the sample analogue of the ratio in Section 2, and its variance estimator linearizes this ratio with $\widehat{\rho}$ treated as fixed. The theorem does not cover this extension, so the subgroup coverage is an empirical check. Tables 1 and 2 report empirical coverage of nominal 95% confidence intervals constructed as $\widehat{\tau}^* \pm 1.96\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\tau}^*)}$ across $B = 500$ Monte Carlo replications with estimated calibration weights and estimated nuisance functions. Entries are percentages. The tables also include $n = N = 1{,}000$, which Figures 1 and 2 omit.

Table 1: Empirical 95% coverage for the overall target ATE across sample-size configurations.

n	N	$\widehat{\tau}$	$\widetilde{\tau}$
1,000	1,000	95.2	94.2
2,000	1,000	95.8	96.4
2,000	2,000	95.0	95.6
2,000	3,000	96.2	95.6
2,000	4,000	95.8	97.0
3,000	3,000	95.2	96.4
4,000	4,000	96.2	95.2
5,000	5,000	96.6	95.0

Table 2: Empirical 95% coverage for age-defined subgroup target ATEs.

n	N	Subgroup	$\widehat{\tau}$	$\widetilde{\tau}$
1,000	1,000	Age 18–29	95.8	94.0
		Age 30–44	96.2	95.0
		Age 45–59	93.8	94.6
		Age 60–80	95.2	94.6
2,000	1,000	Age 18–29	96.0	95.8
		Age 30–44	93.6	93.6
		Age 45–59	96.2	95.2
		Age 60–80	95.8	95.2
2,000	2,000	Age 18–29	96.0	96.0
		Age 30–44	95.0	95.4
		Age 45–59	95.2	95.2
		Age 60–80	95.8	94.8
2,000	3,000	Age 18–29	94.0	95.0
		Age 30–44	96.0	96.0
		Age 45–59	93.0	93.8
		Age 60–80	96.0	95.4
2,000	4,000	Age 18–29	94.6	95.4
		Age 30–44	94.0	95.2
		Age 45–59	96.8	95.6
		Age 60–80	96.8	96.8
3,000	3,000	Age 18–29	95.6	96.4
		Age 30–44	96.0	94.8
		Age 45–59	94.8	95.6
		Age 60–80	97.0	97.0
4,000	4,000	Age 18–29	96.0	96.0
		Age 30–44	95.4	95.4
		Age 45–59	95.2	95.2
		Age 60–80	96.4	96.0
5,000	5,000	Age 18–29	95.4	96.0
		Age 30–44	95.8	95.0
		Age 45–59	96.8	95.2
		Age 60–80	95.0	96.4

G Comparison with direct calibration of the trial

Direct calibration reweights the RCT alone so that its covariate distribution matches the external summaries, as in calibration weighting and entropy balancing [Deville and Särndal, 1992, Hainmueller, 2012]. It can only redistribute mass over trial covariate profiles. To isolate what the DT profiles add, we compare it with the calibrated mixture while holding fixed the sampled units, the external summaries, and the outcome learners. Both estimators learn the outcome models from the RCT alone and differ in how they complete the target covariate distribution.

We use the simulation of Section 4 with $n = N = 5{,}000$ and 200 replications. The outcome adjustment set is age band alone or age band with the demographics gender, education, employment, marital status, and race. The external summaries are the target age-band proportions and, with demographics in the adjustment set, also these proportions with the race proportions or with the employment, marital-status, and race proportions. To keep the comparison within the support of the trial, we retain the 81.3% of the DTs whose combination of age band, demographics, and baseline PHQ-9 category, with the categories 0–4, 5–9, 10–14, 15–19, and 20–27, occurs in the RCT superpopulation and whose baseline PHQ-9 score lies within the RCT range. We construct the target from the RCT superpopulation and the retained DTs as in Section 4. In each replication, we also discard the DTs whose covariate combination in the adjustment set does not occur in the RCT sample

and draw the N DTs from the remaining ones. Direct calibration uses entropy-balancing weights, and its standard error accounts for the estimated weights through the residuals of the weighted regression of the scores $\widehat{\psi}_i$ on the summary indicators. The mixture uses the calibration weights of Section 2 with the settings of Section 4 and the summary indicators as inputs of h_θ, estimates the source propensity score by the cell frequencies of the adjustment set, and forms its intervals with the variance estimator in Eq. (5). For the population limits, direct calibration uses the entropy-balancing weights, and the mixture uses its calibration weight fitted as in the estimator, with $3{,}000$ updates, on the superpopulations. With the joint summaries and in panel (c) of Figure 5, the mixture instead uses the weight that matches the summaries exactly. Because this comparison concerns target structure that the summaries leave unspecified, we evaluate both estimators against the ATE over the full target covariate distribution, which equals 0.75, whereas the age-band reference value of this target equals 1.02. This ATE averages potential outcomes that random forests fitted on the RCT superpopulation generate for the DT profiles, which no estimator uses.

Figure 5 shows that the DT profiles reduce the bias when the outcome adjustment includes covariates whose target distribution the summaries leave unspecified. With age band alone in the adjustment set and the age-band summaries, both estimators average the same age-band conditional effect over the same age-band distribution, up to the calibration tolerance, so their population limits nearly coincide, and at $n = N = 5{,}000$ both have an absolute bias of 0.28, with coverage of 66% for direct calibration and 67% for the mixture. With demographics in the adjustment set, the DT profiles supply the demographic composition of the target within age bands. Across the three sets of summaries, the absolute population-limit bias of the mixture is between 0.02 and 0.08, against 0.27 to 0.31 for direct calibration, and its signed values are 0.08, 0.02, and -0.02. At $n = N = 5{,}000$, the absolute bias of the mixture is 0.21, 0.17, and 0.15 with the age-band summaries alone, with the race proportions, and with the employment, marital-status, and race proportions, against 0.28, 0.31, and 0.33 for direct calibration. The intervals of the mixture cover this ATE in 88% to 89% of replications, against 62% to 66% for direct calibration. The bias of the mixture exceeds its population limit in finite samples mainly because discarding the DTs outside the support of the RCT sample removes part of the demographic structure that distinguishes the target from the trial. The standard deviation of the mixture is 2.0 times that of direct calibration, because the mixture reweights trial units toward demographic profiles that are common among the DTs and rare in the trial.

Direct calibration matches the mixture when the summaries give the joint target distribution of the adjustment covariates. With the joint target proportions of the 717 cells of age band and demographics as summaries, both population limits equal 0.71, a bias of -0.04, because both estimators then average the same cell-level effects over the same distribution of cells. Marginal summaries do not close the gap. Adding the race proportions, or the employment, marital-status, and race proportions, changes the population-limit bias of direct calibration from 0.27 to 0.28 and 0.31. Such joint proportions are rarely reported by external sources, and exact calibration of the trial to them is infeasible at $n = N = 5{,}000$, because on average 100 of the 717 cells contain no RCT unit. The DT profiles supply this joint structure without requiring it from the external source.

The advantage depends on how the DT sample represents this composition. Within each age band, let $p_{\mathrm{DT}}(c)$ and $p_{\mathrm{RCT}}(c)$ denote the proportions of a cell c of the adjustment set in the DT and RCT superpopulations. We replace the DT proportions by $q_\kappa(c) \propto p_{\mathrm{DT}}(c)^{1-\kappa}\, p_{\mathrm{RCT}}(c)^{\kappa}$, which keeps the age-band masses, the summaries, and the target fixed. At $\kappa = 1$, the DT composition equals that of the trial within age bands on the cells that contain DTs, and at $\kappa = 2$ it lies beyond the trial, away from the target. For the adjustment set with demographics and the age-band summaries, the absolute population-limit bias of the mixture is 0.08 at $\kappa = 0$, 0.28 at $\kappa = 0.5$, and 0.26 at $\kappa = 1$, where the mixture nearly coincides with direct calibration, whose bias is 0.27. At $\kappa = 1.5$ and $\kappa = 2$ it equals 0.16 and 0.11, so the bias is not monotone in the misrepresentation. DT profiles therefore add value when the summaries omit outcome-relevant target structure that the DT sample represents, and a DT sample that misrepresents this structure can remove the advantage.

H Diagnostics for the digital-twin covariates

These diagnostics separate the realism of the raw DT distribution from the stability of the calibrated representation that the method requires. We compare observed BRIGHTEN covariates, the CTGAN-generated RCT covariates, the raw Qwen3-8B DT covariates, and the calibrated RCT–DT mixture, using marginal distributions, the joint distributions of all pairs of the six categorical covariates, a

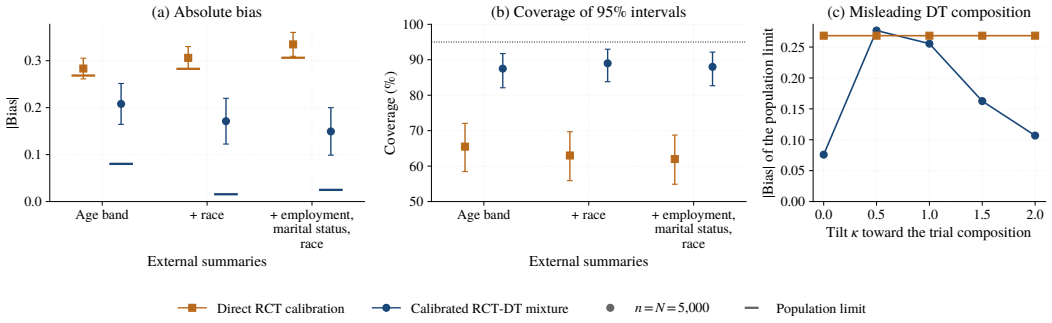


Figure 5: Direct calibration of the RCT versus the calibrated RCT–DT mixture, with the DTs restricted to the support of the trial and the outcome adjustment set of age band and demographics. Panels (a) and (b) report the absolute bias, with normal 95% intervals, and the coverage of 95% intervals, with Clopper–Pearson 95% intervals, across 200 replications for three sets of external summaries. Markers show Monte Carlo estimates at $n = N = 5{,}000$, and horizontal bars in panel (a) show the population limits. Panel (c) reports the absolute bias of the population limits when the DT composition within age bands moves toward and beyond the trial composition, with the age-band summaries.

support check, calibration residuals, and the effective sample size and maximum of the DT weights. We compute them on the RCT and DT superpopulations of size 20,000, so $\eta = 1/2$.

Every category of each covariate in the DT superpopulation also occurs in the RCT superpopulation, with an RCT share of at least 6.8%, and 99.6% of the DTs have a combination of the six categorical covariates that occurs in the RCT.

Calibration matches the age-band proportions with a maximum residual of 5.5×10^{-4} and nearly uniform weights, with an effective sample size ratio of 0.9997 and a maximum normalized weight of 1.04. Figure 6(a) shows that calibration brings the mixture close to the target in the calibrated age bands. Because the target is a reweighted mixture of the same superpopulations, these distances describe how calibration moves the mixture toward the target rather than the realism of the DTs. Figure 6(b) shows that the raw DT joint distribution differs from observed BRIGHTEN more than the CTGAN-generated RCT does, with mean two-way total variation distances of 0.35 and 0.20. A stable calibrated representation therefore does not imply that the raw DT sample reproduces the full joint distribution of observed covariates. These diagnostics assess calibration feasibility and weight stability, but they do not establish that the calibrated mixture represents the target beyond the supplied summaries.

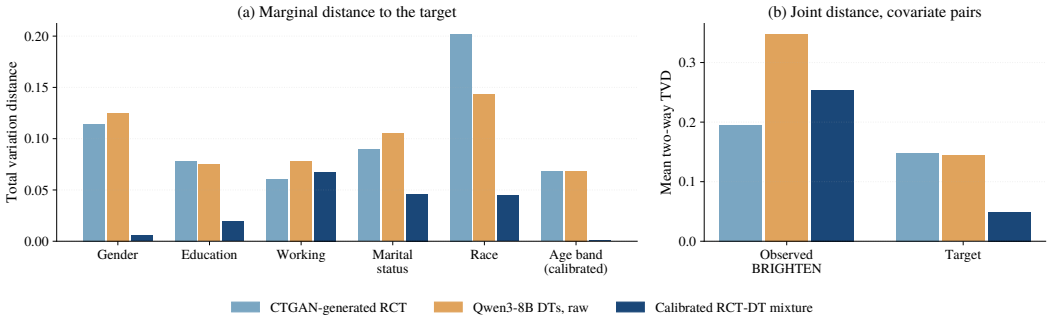


Figure 6: DT diagnostics. Panel (a) reports the total variation distance between each covariate's marginal distribution and the target. Panel (b) reports the mean two-way total variation distance over pairs of the six categorical covariates, relative to observed BRIGHTEN and to the target.

I Sensitivity to the auxiliary predictions

We vary the source of the auxiliary predictions while holding fixed the sampled units, treatment assignments, outcomes, calibration inputs, and estimators, using the first 200 replications of the simulation with $n = N = 2{,}000$. Besides the Qwen3-8B zero-shot predictions of Section 4, we use Qwen3-8B few-shot predictions and GPT-5-mini zero-shot predictions with the same prompt, given in Appendix J. A no-signal control jointly permutes the two arm-specific Qwen3-8B zero-shot predictions within each source with a fixed seed. The permutation preserves the marginal and cross-arm distributions of the predictions but breaks their record-level association with the outcomes.

The correlations between the observed-arm predictions and the RCT outcomes are 0.42, 0.44, and 0.42 for the three prediction sources. Figure 7 shows efficiency gains of 19.5%, 19.4%, and 18.7%, with coverage between 94.0% and 96.0% and higher rejection rates of $H_0 : \tau = 0$ for $\widetilde{\tau}$ than for $\widehat{\tau}$. Under the permutation, the gain falls to -1.7% and $\widetilde{\tau}$ behaves like $\widehat{\tau}$, as expected when the predictions carry no information about the outcomes. We interpret the three model and prompt conditions as a sensitivity analysis, not as an exhaustive evaluation or a ranking of models or prompts. The DT covariates come from Qwen3-8B in all conditions.

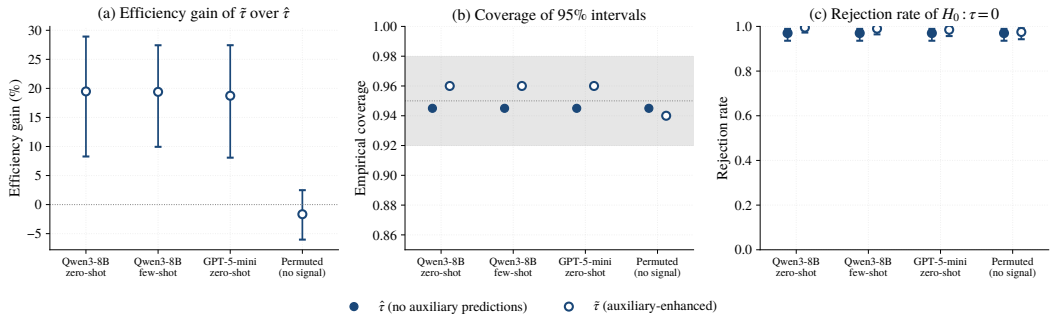


Figure 7: Sensitivity to the auxiliary predictions across 200 replications with $n = N = 2{,}000$. Panel (a) reports the efficiency gain of $\widetilde{\tau}$ over $\widehat{\tau}$ with paired bootstrap 95% intervals, panel (b) the coverage of nominal 95% intervals, with the shaded band marking ± 1.96 binomial Monte Carlo standard errors around 95%, and panel (c) the rejection rate of $H_0 : \tau = 0$, which measures power because the reference value is nonzero, with exact 95% intervals.

J Prompt templates for digital-twin and auxiliary prediction generation

We use Qwen3-8B for two distinct generation tasks. In both tasks the model is called locally with thinking mode disabled. DT covariate generation uses temperature 0.85, and auxiliary outcome prediction uses greedy decoding at temperature 0. The prompts below are representative templates extracted from the supplementary code, which contains the full implementations, including decoding parameters, output parsing, and validation.

DT covariate generation. The system message instructs the model to return JSON for realistic structured baseline variables. The user prompt requests a batch of profiles with an aligned output schema:

```
System: Return JSON only. Generate realistic structured baseline
variables for adults with depressive symptoms. Keep outputs diverse,
plausible, and valid.

User: Generate {batch_size} distinct adult baseline profiles.
Return JSON array only.
Allowed gender: [Female, Male]
Allowed race: [Non-Hispanic White, African American/Black,
Hispanic/Latino, Asian, Other]
Allowed education: [High School or less, Some College, College+]
Allowed working: [Yes, No]
```

```
Allowed marital_status: [Single, Married/Partner,
Separated/Widowed/Divorced]
age must be integer from 18 to 80
baseline_phq9_sum must be integer from 0 to 27
screen_sum must be integer from 0 to 4
Each item must contain exactly these keys: ‘‘gender’’, ‘‘race’’,
‘‘education’’, ‘‘age’’, ‘‘working’’, ‘‘marital_status’’, ‘‘baseline_phq9_sum’’,
‘‘screen_sum’’.
```

Periodic diversity steering appends soft preferences for underrepresented categories without changing the output schema. A response is used only if it parses as a JSON array, and a profile is kept only if it has exactly the eight keys, allowed categories, and integers in the stated ranges, with a component of a merged category, such as Widowed, coded to that category. No value is clipped or filled in.

Auxiliary outcome prediction. The system message instructs the model to act as a clinical outcomes classifier. For each RCT or DT unit, the user prompt provides structured baseline covariates and requests treatment-specific PHQ-9 predictions:

```
System: You are a clinical outcomes classifier. Return STRICT JSON
only.

User: You are a licensed clinical psychologist with expertise
in depression outcomes. Using ONLY the participant's structured
baseline characteristics, predict the expected end-of-study PHQ-9
total score under two treatment conditions. [...]

Participant baseline structured context:
- Age: 45 - Gender: Female - Education: College+
- Working: Yes - Marital Status: Married/Partner
- Race: Non-Hispanic White
- Baseline PHQ-9 total score: 14.0
- Mania/Psychosis screening total score: 0.0

Return STRICT JSON with exactly these keys:
{ ‘‘arm_0_phq9’’: <integer 0-27>, ‘‘arm_1_phq9’’: <integer 0-27> }
```

Few-shot and GPT-5-mini predictions. The few-shot variant adds six demonstrations to the same prompt. They are three treated and three control RCT participants drawn with a fixed seed from a reserved pool of 500 RCT records, each shown with its baseline profile and its observed end-of-study PHQ-9 under the assigned arm. The six demonstration records themselves receive predictions from a second, disjoint set of six demonstrations, so no record serves as its own demonstration. The GPT-5-mini predictions use the zero-shot system and user messages through the OpenAI Responses API with model `gpt-5-mini-2025-08-07`, minimal reasoning effort, at most 256 output tokens, and a structured output with the two integer keys.

Role of auxiliary predictions. The auxiliary predictions $Y^{\dagger}(0)$ and $Y^{\dagger}(1)$ are used only as auxiliary measurements for efficiency improvement through the calibrated outcome model of Section 2. They do not serve as observed outcomes, do not enter the causal identification, and do not change the target estimand.

K Computational resources

The experiments were run on a server with two NVIDIA RTX 6000 Ada Generation GPUs, a 32-core AMD Ryzen Threadripper PRO 7975WX CPU, and 251 GiB of memory. Qwen3-8B generated the DT covariates and the auxiliary outcome predictions on the GPUs, and the GPT-5-mini predictions of Appendix I were obtained through the OpenAI Batch API. With two CPU threads per replication and the calibration weights fitted on the CPU or on a GPU, one Monte Carlo replication took about 1 to 2 minutes for $n = N \leq 4{,}000$ and about 4 minutes for $n = N = 5{,}000$ in the main simulation, about 4 minutes in Appendix G, and about 1 to 2 minutes in Appendix I, so the reported replications used roughly 300 CPU-core hours. Generating the $20{,}000$ DT profiles took about 1.6 GPU hours. The full project also used compute for preliminary runs and for exploring the designs of Appendix G.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions, namely the calibrated mixture representation, the separation of RCT outcomes from DT covariates, and EIF-based estimators with theoretical guarantees. Section 3 establishes consistency, asymptotic normality, and variance consistency for the proposed EIF-based estimators $\widehat{\tau}$ and $\widetilde{\tau}$. The naive plug-in estimators serve as baselines. Limitations are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 3 states when the estimand is the summary-constrained target ATE. Section 5 discusses the transportability condition, the lack of identification outside the support of the trial, when DTs help, the semisynthetic scope of the evaluation, and directions for future work using limited target outcomes. Section 2.3 explains that low-quality DTs can appear as weights concentrated on a few records, and Appendix G shows that a misleading DT sample can remove the advantage over direct calibration of the trial.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions 1–4 are stated in Section 3. Full proofs of Theorem 1, Lemma 1, and Theorem 2 appear in the appendix. Each theorem references the assumptions it requires.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 describes the simulation design including superpopulation construction, DT generation, outcome model specification, treatment randomization probability, cross-fitting, and number of replications. Code for all pipeline steps is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is included in the supplementary material with a README. The BRIGHTEN data are available from the study portal under its data use terms, as described in the cited reference [Pratap et al., 2022].

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 reports superpopulation sizes, sample sizes, number of replications, cross-fitting folds, and treatment randomization probability. Appendix D provides computational details for calibration. Appendix J provides prompt templates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Bias and standard deviation are reported across $B = 500$ replications. Figure 1 reports $\sqrt{n}$-scaled bias, standard deviation, and empirical 95% coverage, and Figure 2 reports subgroup efficiency gains. Appendix F reports empirical 95% coverage diagnostics for the proposed EIF-based estimators. The variance estimators of $\widehat{\tau}$ and $\widetilde{\tau}$ follow Theorem 2. Figure 1(c) shows that the plug-in intervals of the naive estimator under-cover and that the coverage of the RCT-only intervals falls as the sample size grows. Appendices G and I report Monte Carlo and bootstrap intervals for the additional analyses.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix K reports the hardware, two NVIDIA RTX 6000 Ada Generation GPUs, a 32-core CPU, and 251 GiB of memory, the time per Monte Carlo replication, the total CPU and GPU time of the reported runs, the additional compute of preliminary runs, and the use of the OpenAI API for the GPT-5-mini predictions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses de-identified clinical trial data collected under IRB approval as described in the original study. No new human subjects data was collected.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The method aims to improve generalizability of clinical trial results, which could benefit underrepresented populations. A potential risk is that poorly calibrated DTs could lead to incorrect treatment effect estimates. The external summary constraints and the diagnostics of Appendix H can reveal some of these failures, but, as Section 5 notes, not errors in target structure that the summaries leave unspecified.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained language models or scraped datasets. The code generates synthetic clinical data that does not contain identifiable patient information.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The BRIGHTEN dataset is cited with the original publication [Pratap et al., 2022]. The supplementary code README documents external components and license terms, including the BRIGHTEN data under the data provider's terms of use, Qwen3-8B under Apache-2.0, and SDV/CTGAN under BSL-1.1. GPT-5-mini was accessed through the OpenAI API under its terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The submitted code is documented with a README describing the pipeline steps and dependencies.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or new human subjects research was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work uses only de-identified clinical trial data released by the original study. No new human subjects research was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Qwen3-8B is used as a core component for generating digital twin covariates and auxiliary outcome predictions. Qwen3-8B few-shot and GPT-5-mini zero-shot predictions are used in the sensitivity analysis of Appendix I. Appendix J provides representative prompt templates, and the supplementary code contains the exact prompts, decoding parameters, parsing, and validation logic.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.