BEYOND IGNORABLE SELECTION: CAUSAL FUSION WITH SHADOW CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized controlled trials (RCTs) provide credible causal evidence but are often limited in size and representativeness, whereas observational studies (OBS) offer broader coverage but may suffer from hidden treatment confounding. Many data-fusion approaches that combine randomized and observational data rely on sample-selection ignorability or require treatment effects to be transportable across sources. We study data fusion when source selection is nonignorable, so source membership may remain associated with potential outcomes after conditioning on observed covariates. We introduce **Causal Fusion with Shadow Calibration (CFSC)**, which uses shadow-variable restrictions to recover target-population outcome laws and calibrate randomized supervision for estimating treatment effects in the OBS target population. CFSC combines the observational treatment–outcome structure with a structured correction supervised by the recovered target-specific RCT signal. Under the stated conditions, we establish identification for target conditional and average treatment effects. In controlled simulations, CFSC remains stable as source selection strengthens and outperforms conventional fusion methods when both biases coexist. Constructed real-data benchmarks also show remarkable mean ATE-reference agreement and treatment-benefit ranking.

1 INTRODUCTION

Estimating treatment effects in a target population requires both credible causal comparisons and sufficient population coverage. Randomized controlled trials (RCTs) provide strong causal evidence through randomized treatment assignment, but they are often costly, small, and selectively recruited. Observational studies (OBS), in contrast, are typically larger and better represent the population of interest, but may suffer from hidden treatment confounding. Combining the two sources therefore offers a natural way to leverage the internal validity of randomized evidence together with the scale and coverage of observational data (Colnet et al., 2020). In this work, we take the OBS population as the target and study how randomized evidence can improve treatment-effect estimation in this population.

Existing data fusion methods mainly use randomized data to address biases in observational estimation. Trial generalization methods adjust randomized estimates for observed population differences (Lee et al., 2020), while methods such as experimental grounding (RHC) and the integrative R-learner use the RCT as a causal anchor to correct observational confounding or regularize treatment-effect estimation (Kallus et al., 2018; Wu & Yang, 2022; Fawkes et al., 2025; Dimitriou et al., 2026). Their common premise is that randomized information can be transferred across sources after conditioning on observed covariates.

However, this premise may fail when trial participation itself is selective. Take a job-training trial as an example: even among people with the same recorded education and employment history, those who are more motivated to find work may be more likely to volunteer for the trial. If they also benefit more from training, the trial may overestimate the benefit for the broader population. Randomization ensures an unbiased comparison within the trial, but the trial may still give a biased effect for the target population. Figure 1 illustrates our setting, in which the outcome Y and source membership S are dependent even after conditioning on the observed covariates X.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

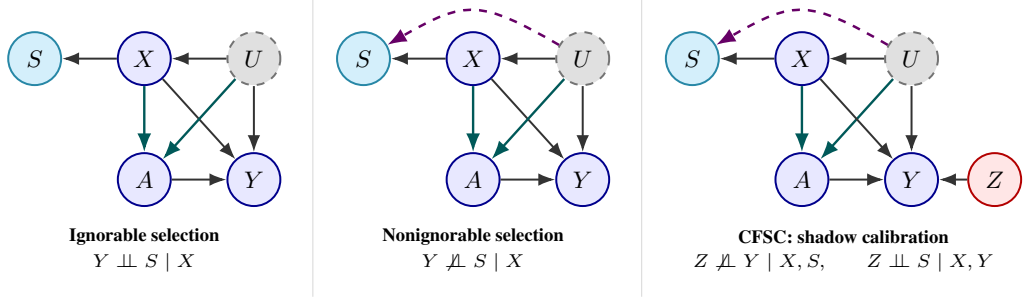

Figure 1: **Data fusion beyond ignorable selection.** (a) Under ignorable selection, Y is independent of source membership S given X. (b) Under nonignorable selection, Y and S remain dependent given X. (c) CFSC addresses this dependence using an outcome-informative shadow variable Z. Formal conditions on the potential outcomes $Y(a)$ are stated in Section 3.2 and are not implied by the arrows alone.

To address this problem, we introduce **Causal Fusion with Shadow Calibration (CFSC)**, which uses cross-source differences in a shadow variable Z to recover target outcome information and calibrate a treatment-effect model learned from the larger OBS sample. Our approach is motivated by the similarity between nonignorable source selection and missing-not-at-random (MNAR) mechanisms: just as missingness under MNAR depends on the outcome Y even after conditioning on X, source membership S may remain associated with outcome-related information beyond X. Under MNAR, a shadow variable can help recover the outcome distribution by providing outcome information while being conditionally independent of missingness given the outcome and covariates (Miao et al., 2015). The similarity between MNAR and nonignorable source selection prompts us to explore a similar strategy for addressing selection bias, leading to CFSC.

Our contributions are threefold:

- We study data fusion when hidden treatment confounding in the OBS coexists with nonignorable trial selection, a setting in which, to the best of our knowledge, no prior work has identified target treatment effects.
- We establish identification of target treatment effects under nonignorable trial selection and develop CFSC to estimate these effects through shadow calibration.
- We evaluate CFSC in controlled simulations and real-data benchmarks, demonstrating its superior performance under selection bias.

2 Related Work

2.1 Data Fusion

Research on combining randomized and observational data considers transporting trial effects or using randomized evidence to improve observational estimation (Colnet et al., 2020). For transport, outcome regression, sampling weights, and calibration weighting adjust for observed population differences (Dahabreh et al., 2018; Lee et al., 2020). More closely related methods correct confounded observational estimates: experimental grounding learns a low-complexity correction (Kallus et al., 2018), but Fawkes et al. (2025) show that experiments can falsify biased estimates yet cannot generally validate or remove bias without regularity assumptions, and introduce a pseudo-outcome GP approach under smoothness. Confounding-function and representation approaches include CorNet and the integrative R-learner (Hatt et al., 2022; Yang et al., 2025; Wu & Yang, 2022). Adaptive borrowing appears in elastic integration (Yang et al., 2020) and Causal-ICM, which uses multi-task GPs and a data-adaptive parameter to keep potentially confounded observational data from overwhelming randomized evidence (Dimitriou et al., 2026). Other recent methods study hidden-confounder alignment, power-likelihood borrowing, and partial identification under failures of standard fusion assumptions (Aloui et al., 2025; Lin et al., 2025; Lanners et al., 2025). While these works address observational confounding, correction modeling, or sensitivity, CFSC targets a distinct failure:

source membership may depend on outcome-related information. It uses shadow-variable structure to recover target-specific supervision when standard source-selection or transportability conditions fail.

2.2 MNAR IDENTIFICATION

Under missing not at random (MNAR), missingness may depend on outcomes even after conditioning on observed covariates, so identification generally requires additional structure. A broad class of approaches exploits auxiliary variables whose relationships with the outcome and missingness mechanism provide identifying information. Shadow-variable methods use variables associated with the outcome but conditionally independent of missingness given the outcome and covariates, yielding identification through parametric or completeness conditions (d'Haultfoeuille, 2010; Wang et al., 2014), odds-ratio formulations, and doubly robust estimation (Miao et al., 2015; Miao & Tchetgen Tchetgen, 2016); some functionals can also be identified without recovering the full-data law (Li et al., 2023). A complementary structure uses instrumental variables that predict missingness while being excluded from the outcome model (Sun et al., 2018), with related proxy-variable methods extending auxiliary-variable identification to latent-variable settings (Miao et al., 2018; Nilsson et al., 2023; Su et al., 2026). These results show how auxiliary information can restore identification under otherwise nonignorable mechanisms. CFSC draws on this perspective by treating nonignorable source membership in data fusion as an MNAR-type selection problem and using recovered target outcome laws to construct supervision for experimental grounding.

3 THEORY

3.1 PROBLEM SETUP

We study identification of treatment effects in an observational study (OBS) population using causal information from a randomized controlled trial (RCT). In this information transfer, the RCT population is the source and the OBS population is the target. Let $S \in \{\mathrm{RCT}, \mathrm{OBS}\}$ indicate study membership, X denote the observed pre-treatment covariates on which effects are reported, Z an additional observed pre-treatment shadow candidate, $A \in \{0, 1\}$ treatment, Y the observed outcome, and $Y(a)$ the potential outcome under treatment a. We observe independent samples $\mathcal{D}_s = \{(X_i, Z_i, A_i, Y_i) : S_i = s\}$ of size n_s from $P(X, Z, A, Y \mid S = s)$.

Definition 1 (Target treatment effects)**.** *For population s, let $\mu_a^s(x) = \mathbb{E}[Y(a) \mid X = x, S = s]$ be the response surface under treatment a. Define its conditional average treatment effect (CATE) and average treatment effect (ATE) as*

$$\tau_s(x) = \mu_1^s(x) - \mu_0^s(x), \qquad \psi_s = \mathbb{E}[\tau_s(X) \mid S = s]. \tag{1}$$

In this paper, the target population is $s = \mathrm{OBS}$, so the estimands are τ_{OBS} and ψ_{OBS}.

Assumption 1 (Basic causal and support conditions)**.** *These conditions govern the observed outcome, treatment assignment within the RCT, and coverage of the target X support.*

(i) Consistency. *Treatments are well defined, with no interference; the observed outcome equals the potential outcome under the received treatment in both sources:*

$$Y = AY(1) + (1 - A)Y(0).$$

(ii) RCT randomization. *Conditional on X, treatment in the RCT is independent of the potential outcomes and the pre-treatment shadow candidate:*

$$A \perp\!\!\!\perp \{Y(0), Y(1), Z\} \mid X, S = \mathrm{RCT}.$$

(iii) Trial positivity and source overlap. *With $e_{\mathrm{RCT}}(x) = P(A = 1 \mid X = x, S = \mathrm{RCT})$, for some $\epsilon > 0$ both trial arms have positive probability and the trial covers the target covariate support:*

$$\epsilon \le e_{\mathrm{RCT}}(x) \le 1 - \epsilon,$$
$$P(X \mid S = \mathrm{OBS}) \ll P(X \mid S = \mathrm{RCT}).$$

Here $\ll$ denotes absolute continuity.

These basic conditions support internal validity in the RCT and cover the target covariate support, but do not restrict treatment assignment in the OBS or differences in potential outcomes across sources. In the OBS, treatment may depend on potential outcomes even after conditioning on X.

Definition 2 (Observational treatment confounding)**.** *Treatment is confounded in the OBS when observational treatment exchangeability fails:*

$$\{Y(0), Y(1)\} \not\perp\!\!\!\perp A \mid X, S = \text{OBS}.$$

Randomized treatment in the RCT avoids this assignment bias among trial participants. To use trial evidence for the OBS target, however, a conventional data-fusion route additionally assumes source exchangeability (Colnet et al., 2020; Kallus et al., 2018):

$$\{Y(0), Y(1)\} \perp\!\!\!\perp S \mid X.$$

This condition requires the potential-outcome distributions to agree across sources at the same X. RCT randomization concerns treatment assignment within the trial; it does not imply random trial participation. If unobserved prognostic characteristics also influence participation, the distributions may differ even after conditioning on X. We do not impose source exchangeability.

Definition 3 (Nonignorable source selection)**.** *Source selection is nonignorable given X if, for at least one treatment arm a,*

$$Y(a) \not\perp\!\!\!\perp S \mid X, \tag{2}$$

with dependence on a set of covariate values of positive target probability.

We study target-effect identification while allowing both observational treatment confounding (Definition 2) and nonignorable source selection (Definition 3); the next section states the identifying conditions.

3.2 Identification

We now show that target treatment effects remain identifiable without source exchangeability by using a shadow variable to recover the outcome-dependent distortion induced by source selection. The key is an observed covariate that is informative about the potential outcome but conditionally excluded from the selection mechanism.

We use Z as an auxiliary variable while keeping the target CATE conditional on X. Outcome relevance is needed so that changes in the distribution of $Y(a)$ leave a detectable trace in Z. Selection exclusion is equally necessary: if Z still carried information about source membership given $X, Y(a)$, the same trace could instead arise from direct selection on Z. We formalize these requirements as follows.

Assumption 2 (Shadow-variable conditions)**.** *For each arm $a \in \{0, 1\}$,*

$$\begin{aligned} Z \not\perp\!\!\!\perp Y(a) \mid X, S = \text{RCT} \qquad & \textit{(outcome relevance)}, \\ Z \perp\!\!\!\perp S \mid X, Y(a) \qquad & \textit{(selection exclusion)}. \end{aligned} \tag{3}$$

We call a candidate satisfying these restrictions a shadow variable. We do not assume that one is known a priori; the Shadow Recovery component of Section 4 describes how candidates are screened from the observed covariates, and the results below are conditional on the selected candidate satisfying Assumption 2.

To represent the source-to-target distortion by a density ratio, we assume that, for each arm and $P(X \mid S = \text{OBS})$-almost every x, the target law of $Y(a) \mid X = x$ is absolutely continuous with respect to the corresponding source law. Let $f_a(y \mid x, S = s)$ denote these outcome densities, with sums replacing integrals for discrete variables. The unknown outcome-selection tilt r_a compares the target and source laws, whereas d is the observed source difference in Z:

$$r_a(x, y) = \frac{f_a(y \mid x, S = \text{OBS})}{f_a(y \mid x, S = \text{RCT})}, \qquad d(x, z) = \frac{p(z \mid x, S = \text{OBS})}{p(z \mid x, S = \text{RCT})}. \tag{4}$$

The shadow restrictions connect these two shifts. Consistency and randomization identify $f_a(y \mid x, z, S = \text{RCT}) = p(y \mid A = a, X = x, Z = z, S = \text{RCT})$; together with the observed source

distribution of Z, this identifies $q_a(z \mid x, y) = p(z \mid X = x, Y(a) = y, S = \mathrm{RCT})$. Selection exclusion makes this $Z \mid X, Y(a)$ kernel common across populations, giving the mixture relation (Miao et al., 2015)

$$p(z \mid x, S = \mathrm{OBS}) = \int q_a(z \mid x, y) f_a(y \mid x, S = \mathrm{OBS})\, dy. \tag{5}$$

Thus the observed target distribution of Z, combined with the source-identified kernel, restricts the unknown target outcome law. Writing the corresponding source-side conditional averaging operator as

$$\begin{aligned}(K_{a,x}h)(z) &= \mathbb{E}[h(Y(a)) \mid X = x, Z = z, S = \mathrm{RCT}] \\ &= \int h(y) f_a(y \mid x, z, S = \mathrm{RCT})\, dy.\end{aligned} \tag{6}$$

turns equation 5 into the central recovery equation

$$K_{a,x} r_a(x, \cdot) = d(x, \cdot). \tag{7}$$

Here the observable shift in Z equals a source-side conditional average of the unknown tilt. This observed equation need not have a unique solution: distinct admissible tilts may satisfy $K_{a,x} r_a = K_{a,x} r'_a = d$. Identification therefore requires $K_{a,x}$ to be injective on the relevant ratio class.

Assumption 3 (Completeness on the ratio class)**.** *For each arm and almost every target x, let $\mathcal{G}_{a,x}$ be a specified class of nonnegative measurable functions g with $\int g(y) f_a(y \mid x, S = \mathrm{RCT})\, dy = 1$, containing $r_a(x, \cdot)$. For every difference $h = g - g'$, $g, g' \in \mathcal{G}_{a,x}$,*

$$K_{a,x}h = 0 \quad P(Z \mid x, S = \mathrm{RCT})\text{-a.s.} \quad \Longrightarrow \quad h = 0 \quad P(Y(a) \mid x, S = \mathrm{RCT})\text{-a.s.} \tag{8}$$

Completeness adds no observed information; it ensures that the recovery equation has a unique admissible solution.

Theorem 1 (Identification of target treatment effects)**.** *Under the preceding outcome-support condition and Assumptions 1(i), 1(ii), 1(iii), 2, and 3, equation 7 uniquely identifies r_a within $\mathcal{G}_{a,x}$, up to source outcome-null sets. This tilt identifies the target conditional outcome density:*

$$f_a^{\mathrm{SV}}(y \mid x, z, S = \mathrm{OBS}) = \frac{r_a(x, y) f_a(y \mid x, z, S = \mathrm{RCT})}{d(x, z)}. \tag{9}$$

Here SV *refers to the Shadow Recovery procedure, and the denominator is positive target-almost surely. Integrating over the observed target distribution of $Z \mid X = x$ gives $f_a^{\mathrm{SV}}(y \mid x, S = \mathrm{OBS}) = r_a(x, y) f_a(y \mid x, S = \mathrm{RCT})$. If $\mathbb{E}[|Y(a)| \mid S = \mathrm{OBS}] < \infty$ for both arms, its first moment identifies the target response surface conditional on $X = x$:*

$$\mu_a^{\mathrm{OBS}}(x) = \int y f_a^{\mathrm{SV}}(y \mid x, S = \mathrm{OBS})\, dy. \tag{10}$$

Consequently, τ_{OBS} and ψ_{OBS} are identified.

The recovered law supplies the causal signal used in Section 4. Integrating out the auxiliary Z keeps the identified CATE aligned with the target estimand $\tau_{\mathrm{OBS}}(x)$, which is conditional on X alone. Proof is in Appendix A.5.

4 METHODOLOGY

The identification result in Section 3.2 shows that the shadow restriction can recover a target-population treatment-effect signal despite nonignorable source selection. Using this signal directly as the final CATE estimator, however, would require the comparatively small RCT to learn a potentially complex heterogeneous effect through an inverse recovery problem. The large OBS, by contrast, provides the scale and covariate coverage needed to learn rich structure in observed treatment–outcome relationships, although the resulting estimator may be biased by unmeasured treatment confounding. We therefore use the two sources asymmetrically: the OBS captures this rich structure, while the shadow-recovered RCT signal provides causal calibration. We call the resulting estimator **Causal Fusion with Shadow Calibration (CFSC)**. CFSC has three stages: *Observational Outcome Modeling* fits a potentially biased observational estimator $\widehat{w}_{\mathrm{OBS}}$; *Shadow Recovery* screens the shadow candidates and recovers target-effect supervision $\widetilde{\tau}_{\mathrm{SV}}$; and *Structured Correction* learns the discrepancy $\widehat{b}$ needed to form the final estimator $\widehat{\tau}_{\mathrm{OBS}} = \widehat{w}_{\mathrm{OBS}} - \widehat{b}$. Figure 2 summarizes this workflow.

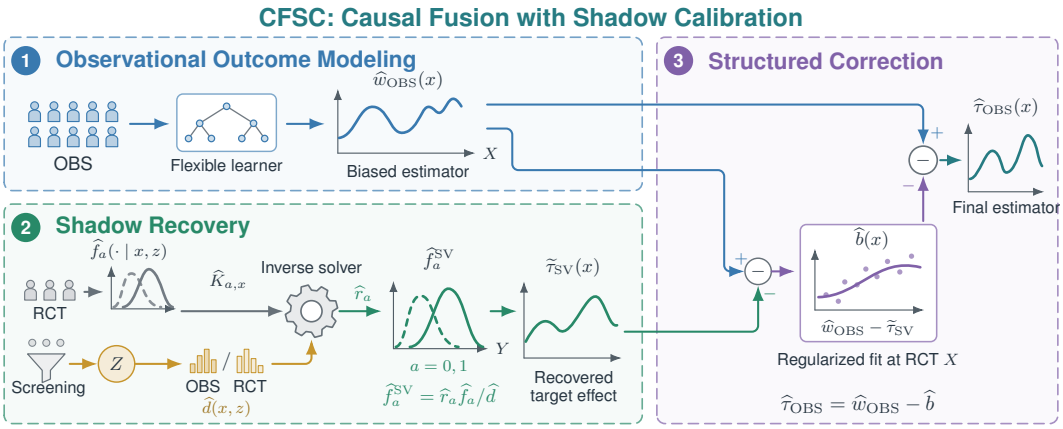

Figure 2: Overview of Causal Fusion with Shadow Calibration (CFSC). (1) Observational Outcome Modeling fits the biased OBS estimator $\widehat{w}_{\mathrm{OBS}}$. (2) Shadow Recovery screens Z and uses the shadow restriction to recover target-effect supervision $\widetilde{\tau}_{\mathrm{SV}}$ from the RCT. (3) Structured Correction fits $\widehat{b}$ at RCT covariate locations to produce final estimator $\widehat{\tau}_{\mathrm{OBS}} = \widehat{w}_{\mathrm{OBS}} - \widehat{b}$.

4.1 Causal Fusion with Shadow Calibration

Observational Outcome Modeling. For each treatment arm, define the factual OBS mean, its contrast, and the correction by

$$\begin{aligned} m_a^{\mathrm{OBS}}(x) &= \mathbb{E}[Y \mid A = a, X = x, S = \mathrm{OBS}], \\ w_{\mathrm{OBS}}(x) &= m_1^{\mathrm{OBS}}(x) - m_0^{\mathrm{OBS}}(x), \qquad b_{\mathrm{OBS}}(x) = w_{\mathrm{OBS}}(x) - \tau_{\mathrm{OBS}}(x). \end{aligned} \tag{11}$$

Thus $\tau_{\mathrm{OBS}}(x) = w_{\mathrm{OBS}}(x) - b_{\mathrm{OBS}}(x)$. This decomposition turns data fusion into a correction problem and follows a common experimental-grounding strategy (Kallus et al., 2018): the large OBS learns the rich treatment–outcome structure in w_{OBS}, while randomized evidence is used only to learn the discrepancy b_{OBS} between this observational contrast and the target causal effect. Conventional implementations use an RCT-population effect directly for this correction and therefore retain RCT-to-target selection bias when trial effects do not transport. Under the shadow restrictions, CFSC instead first recovers an effect signal for the OBS target population, removing this source-selection distortion from the randomized supervision.

We fit the observational contrast with a flexible learner $\mathcal{Q}$ and write $\widehat{w}_{\mathrm{OBS}} = \mathcal{Q}(\mathcal{D}_{\mathrm{OBS}})$. Because $\widehat{w}_{\mathrm{OBS}}$ captures observed treatment–outcome structure and may remain confounded, it is the biased estimator in Figure 2; the subsequent stages supply its causal calibration.

Shadow Recovery. Shadow Recovery begins by selecting a shadow candidate from a fixed finite set $\mathcal{Z} = \{Z_1, \ldots, Z_M\}$ of auxiliary variables observed in both sources but excluded from the target CATE covariates X. On a separate screening split, we remove candidates that add little outcome information beyond X within either randomized arm. For the remaining candidates, we fit the recovery equation and assess held-out moment residuals and tilt stability, selecting the final candidate by pre-specified fit and stability criteria. We denote the selected candidate by Z.

We apply the identification result in Section 3.2 to the selected shadow variable, recovering the OBS target outcome law and its treatment-effect signal from the RCT. We call this stage *Shadow Recovery* because the shadow restriction enables recovery under nonignorable source selection, not because the procedure recovers Z itself. For each arm a, we fit the RCT outcome kernel $\widehat{f}_a(y \mid x, z, S = \mathrm{RCT})$ and estimate the observable source shift $\widehat{d}(x, z)$. We then estimate a positive tilt $\widehat{r}_a$ from the empirical analogue of equation 7,

$$(\widehat{K}_{a,x}\widehat{r}_a)(z) \approx \widehat{d}(x, z). \tag{12}$$

Figure 2 calls the numerical solution of equation 12 the *Inverse Solver*; it is a computational step, not another model or assumption. We parameterize its positive tilt through a reference-normalized

odds ratio, with details in Appendix A.1. The fitted tilt and RCT outcome law give $\widehat{f}_a^{\mathrm{SV}}(y \mid x, S = \mathrm{OBS}) = \widehat{r}_a(x, y)\widehat{f}_a(y \mid x, S = \mathrm{RCT})$, which integrates out the auxiliary Z and yields

$$\widehat{\mu}_a^{\mathrm{SV}}(x) = \int y\, \widehat{f}_a^{\mathrm{SV}}(y \mid x, S = \mathrm{OBS})\, dy, \tag{13}$$

$$\widetilde{\tau}_{\mathrm{SV}}(x) = \widehat{\mu}_1^{\mathrm{SV}}(x) - \widehat{\mu}_0^{\mathrm{SV}}(x). \tag{14}$$

The output $\widetilde{\tau}_{\mathrm{SV}}$ is therefore a recovered target CATE function that supplies target-effect supervision, not a recovered shadow variable or an individual pseudo-outcome. The screening diagnostics assess empirical compatibility, not validity of the identifying restriction itself: they do not prove $Z \perp\!\!\!\perp S \mid X, Y(a)$ or completeness. The theoretical guarantee with screening therefore remains conditional on selecting a valid candidate with probability tending to one.

Structured Correction. The recovered target-effect supervision identifies how the observational surface must be corrected at the RCT covariate locations, without asking the comparatively small trial to relearn the full heterogeneous CATE. For each randomized unit, define

$$\widetilde{b}_i = \widehat{w}_{\mathrm{OBS}}(X_i) - \widetilde{\tau}_{\mathrm{SV}}(X_i). \tag{15}$$

We fit the structured correction by

$$\widehat{b} = \arg\min_{b \in \mathcal{H}} \left[\frac{1}{n_{\mathrm{RCT}}} \sum_{i:S_i = \mathrm{RCT}} \{\widetilde{b}_i - b(X_i)\}^2 + \lambda \mathcal{R}(b) \right], \tag{16}$$

where $\mathcal{H}$ is the correction function class, $\mathcal{R} : \mathcal{H} \to [0, \infty)$ is a complexity penalty, and $\lambda \geq 0$ controls regularization. The class encodes a simpler or smoother bias surface than the target CATE and extends the fitted correction over the common covariate support. Thus the object learned from the recovered RCT signal is a structured correction rather than the full target response surface.

The final CATE and ATE estimators are

$$\widehat{\tau}_{\mathrm{OBS}}(x) = \widehat{w}_{\mathrm{OBS}}(x) - \widehat{b}(x), \tag{17}$$

$$\widehat{\psi}_{\mathrm{OBS}} = \frac{1}{n_{\mathrm{OBS}}} \sum_{i:S_i = \mathrm{OBS}} \widehat{\tau}_{\mathrm{OBS}}(X_i). \tag{18}$$

Unlike direct experimental grounding, which calibrates an observational learner using an RCT-population effect signal, CFSC first recovers the calibration target for the OBS population through the shadow restriction. Algorithm 1 in Appendix A.2 summarizes the complete estimation procedure.

Consistency. Under consistent nuisance estimation, stable Shadow Recovery, and correct specification of the correction class, CFSC consistently estimates the target response surfaces, CATE, and ATE on the common support. Appendix A.4 states the formal results and their conditions.

5 Experiments

5.1 Synthetic Data

Data and evaluation. Each replicate contains $n = 6{,}000$ observations with three independent standard-normal covariates X, an unobserved variable U, and a noisy shadow $Z = 0.65U + \sqrt{1 - 0.65^2}\,\epsilon_Z$. We generate the potential outcomes as

$$\begin{aligned} Y(0) &= 0.5X_1 + 0.3X_2 - 0.2X_3 + U, \\ Y(1) &= Y(0) + \tau_X(X) + 0.5U, \\ \tau_X(X) &= 1 + 0.5X_1 - 0.3X_2 + 0.20(X_2^2 - 1) + 0.15X_1X_3. \end{aligned} \tag{19}$$

Logistic source sampling assigns approximately 40% of observations to the RCT, where treatment is randomized. In the source-assignment score, α_{sel} is the coefficient on the latent effect modifier U; equivalently, writing the remaining score as $\eta(X)$,

$$\mathrm{logit}\, P(S = \mathrm{RCT} \mid X, U) = \eta(X) + \alpha_{\mathrm{sel}} U.$$

Thus, $\alpha_{\text{sel}} = 0$ makes source assignment independent of U, whereas larger values increasingly separate the RCT and OBS populations along an unobserved effect modifier. Crossing RCT source selection with OBS treatment confounding yields four settings: neither bias, OBS confounding only, source selection only, and both biases simultaneously. On a held-out 30% of the OBS target sample, we evaluate 20 paired replicates using $\text{PEHE} = \{n_{\text{test}}^{-1} \sum_{i \in \text{test}} [\widehat{\tau}(X_i) - \tau_{\text{OBS}}(X_i)]^2\}^{1/2}$; lower values are better. Estimators receive neither U nor the true effects. Details are in Appendix A.6.

Results across bias regimes. Table 1 compares the methods across the four bias regimes, including PO-GP (Fawkes et al., 2025) and Causal-ICM (Dimitriou et al., 2026).

Table 1: Synthetic target-OBS PEHE (mean ± SD; lower is better). Bold marks the lowest mean.

	Data setting			
Method	Neither	OBS conf.	Source sel.	Both
OBS only				
T-learner (Künzel et al., 2019)	0.162 ± 0.048	1.074 ± 0.057	0.137 ± 0.044	0.848 ± 0.053
R-learner (Nie & Wager, 2021)	0.168 ± 0.051	1.076 ± 0.054	0.149 ± 0.034	0.854 ± 0.056
Causal Forest (Athey et al., 2019)	0.297 ± 0.040	1.099 ± 0.056	0.270 ± 0.024	0.886 ± 0.056
RCT + OBS				
IPSW (Dahabreh et al., 2018)	0.344 ± 0.031	0.344 ± 0.031	0.621 ± 0.036	0.621 ± 0.036
CW (Lee et al., 2020)	0.344 ± 0.031	0.344 ± 0.031	0.621 ± 0.036	0.621 ± 0.036
RHC (Kallus et al., 2018)	0.164 ± 0.051	0.177 ± 0.046	0.561 ± 0.048	0.577 ± 0.052
Integrative-R (Wu & Yang, 2022)	0.345 ± 0.031	0.345 ± 0.031	0.619 ± 0.037	0.619 ± 0.037
PO-GP (Fawkes et al., 2025)	0.371 ± 0.032	0.429 ± 0.044	0.690 ± 0.045	0.729 ± 0.062
Causal-ICM (Dimitriou et al., 2026)	0.153 ± 0.021	0.154 ± 0.022	0.559 ± 0.059	0.560 ± 0.058
CFSC (ours)	$\mathbf{0.131 \pm 0.019}$	$\mathbf{0.145 \pm 0.020}$	$\mathbf{0.130 \pm 0.018}$	$\mathbf{0.146 \pm 0.024}$

With neither bias, direct OBS learning and RCT–OBS methods are competitive because the OBS contrast is causally valid and the RCT effect transports; CFSC retains a modest advantage. Under OBS confounding, OBS-only estimators deteriorate because they cannot remove latent treatment-assignment bias, whereas randomized supervision substantially improves accuracy. Under source selection alone, the OBS contrast remains valid, so incorporating an unrepresentative trial can be harmful. Accordingly, PO-GP and Causal-ICM deteriorate even though direct OBS estimators remain accurate. When both biases coexist, neither the raw OBS contrast nor direct randomized correction is reliable; CFSC remains accurate by calibrating randomized supervision for the target population.

Component and sensitivity analyses. Table 2 gives a cumulative component analysis under joint OBS treatment confounding and RCT source selection. From the same OBS backbone, direct RCT correction reduces confounding but assumes trial representativeness; shadow correction then calibrates this supervision to the target and yields full CFSC. Figure 3 holds OBS confounding fixed and increases α_{sel}. Comparison methods deteriorate as the RCT becomes less representative, whereas CFSC remains stable.

Table 2: Component analysis: OBS conf. + source sel.

Specification	PEHE
OBS only	0.854
+ Direct RCT corr.	0.577
+ Shadow corr. (CFSC)	**0.146**

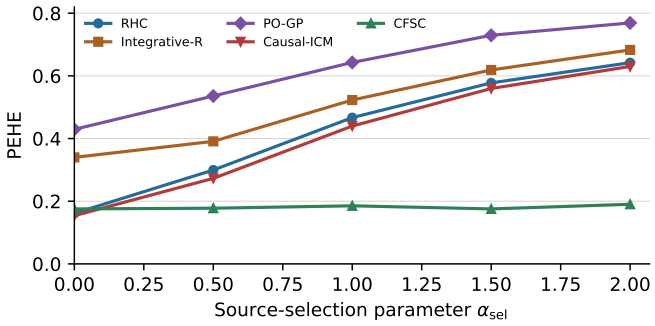

Figure 3: Selection sensitivity analysis.

5.2 REAL-WORLD DATA

Data and evaluation. We construct selection benchmarks from ACTG (ZDV+ddI versus ddI; follow-up CD4) and JTPA (randomized program assignment; earnings). After reserving 20% of records for design, we sample RCT membership with $P(S = \mathrm{RCT}) = \mathrm{expit}(a_0 + V_{\mathrm{std}})$, targeting 30% for ACTG and 20% for JTPA; V is baseline stratification in ACTG and age in JTPA. Within each of five target-population folds, we retain 70% of OBS training observations in expectation using $P(\mathrm{retain} \mid A, Y) = \mathrm{expit}\{b_0 + (0.5 - A)Y_{\mathrm{std}}\}$. This outcome-dependent thinning induces selection bias; evaluation folds remain intact. All methods exclude V and Z from X; CFSC additionally uses baseline CD4 as Z in ACTG and prior earnings in JTPA. Treatment, outcome, and shadow values remain unchanged.

Because individual effects are unavailable, we report $\epsilon_{\mathrm{ATE}} = |\widehat{\tau}_{\mathrm{ATE}} - \widehat{\tau}^{\mathrm{AIPW}}_{\mathrm{ATE}}|$ using the evaluation-fold mean prediction and its randomized AIPW reference. AUTOC measures how well estimated CATE ranks treatment benefit across targeting fractions; larger is better (Yadlowsky et al., 2025).

Table 3: Real-data evaluation (mean ± SD). Arrows indicate the preferred direction; bold marks the best mean.

	ACTG		JTPA	
Method	$\epsilon_{\mathrm{ATE}} \downarrow$	AUTOC↑	$\epsilon_{\mathrm{ATE}} \downarrow$	AUTOC↑
OBS only				
T-learner (Künzel et al., 2019)	210.9 ± 481.7	11.50 ± 10.30	3572.3 ± 127.59	173.7 ± 220.30
R-learner (Nie & Wager, 2021)	38.54 ± 17.28	11.58 ± 8.47	3571.3 ± 130.34	71.14 ± 265.66
Causal Forest (Athey et al., 2019)	36.57 ± 5.53	10.78 ± 7.92	3574.1 ± 139.88	-171.63 ± 222.49
RCT + OBS				
IPSW (Dahabreh et al., 2018)	28.84 ± 21.88	11.05 ± 6.83	835.96 ± 672.09	-9.09 ± 250.88
CW (Lee et al., 2020)	28.69 ± 21.77	10.81 ± 6.79	835.97 ± 672.04	-13.15 ± 249.46
RHC (Kallus et al., 2018)	23.62 ± 15.27	11.44 ± 8.26	785.06 ± 582.74	119.46 ± 208.95
Integrative-R (Wu & Yang, 2022)	27.08 ± 22.27	13.71 ± 3.71	838.62 ± 669.85	-10.80 ± 309.64
PO-GP (Fawkes et al., 2025)	23.92 ± 19.02	7.92 ± 10.20	743.23 ± 629.82	-277.63 ± 487.35
Causal-ICM (Dimitriou et al., 2026)	21.84 ± 16.33	15.82 ± 4.42	787.42 ± 614.61	-454.25 ± 450.44
CFSC (ours)	$\mathbf{21.37 \pm 12.11}$	$\mathbf{16.43 \pm 6.29}$	$\mathbf{704.92 \pm 435.67}$	$\mathbf{250.95 \pm 162.15}$

On ACTG, Causal-ICM and CFSC perform similarly in both ATE agreement and treatment-benefit ranking, with a small mean advantage for CFSC. Their behavior separates on JTPA: CFSC achieves the strongest mean ATE agreement and prioritization, whereas PO-GP and Causal-ICM have negative mean AUTOC. Together with the synthetic results, this pattern suggests that CFSC is more stable when source selection is pronounced. Because variability is substantial, we interpret these results as qualitative evidence from constructed selection benchmarks rather than claims about hidden confounding in the original studies.

6 CONCLUSION

Randomization removes within-trial treatment-assignment bias but does not make trial participants representative of the target population. Under nonignorable trial participation, direct randomized correction can transmit source-selection bias. CFSC instead recovers target-specific RCT supervision through shadow calibration and corrects the OBS treatment contrast. Under the stated conditions, this identifies and consistently estimates the target effect on the common support. Simulations improve CATE accuracy, while real-data-based benchmarks show better agreement with a randomized ATE reference and higher treatment-benefit ranking scores.

More broadly, our results highlight that randomized evidence may itself require population calibration before serving as causal supervision for data fusion. The current guarantees remain conditional on a valid and sufficiently informative shadow variable, stable recovery, and adequate support overlap. A natural direction is to extend this target-specific supervision principle beyond shadow variables, for example using instruments for source selection under suitable relevance and exclusion conditions. This principle may also support reverse-direction fusion, using observational data to improve effect estimation for the RCT population.

AI USE STATEMENT

Generative AI tools were used to assist with the initial paper scaffold, organization, language editing, and drafting the discussion, conclusion, and submission statements based on the existing manuscript. The authors are responsible for the final content, including all AI-assisted text, mathematical claims, citations, code, and empirical results.

ETHICS STATEMENT

This work develops methods for causal estimation and evaluates them using controlled simulations and benchmarks based on existing ACTG and JTPA data. Such estimates may inform healthcare or policy decisions, where errors in transporting effects can have consequences for individuals and underrepresented groups. Accounting for nonignorable selection does not by itself guarantee equitable or reliable decisions: invalid shadow restrictions, weak overlap, or model misspecification may produce misleading estimates. The reported benchmark results should therefore be interpreted as methodological evaluations, rather than evidence supporting a particular treatment or policy. Any application requires domain-specific assessment of the identifying assumptions, subgroup coverage, and uncertainty.

REPRODUCIBILITY STATEMENT

Section 4 specifies the recovery and correction estimators, Section 3.1 states the identification conditions, and Section A.4 establishes consistency, including a conditional extension to screened candidates. Appendix A.5 provides the proofs, with recovery regularity conditions in Section A.4. Section 5 describes the comparisons and evaluation metrics; Sections 5.1 and 5.2 give the synthetic and real-data-based results. Appendix A.6 records the supplied design and implementation settings. An anonymous code example is available at `https://anonymous.4open.science/r/CFSC-55F3/`.

REFERENCES

Ahmed Aloui, Juncheng Dong, Ali Hasan, and Vahid Tarokh. Conditional average treatment effect estimation under hidden confounders. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 91–110. PMLR, 2025. URL https://proceedings.mlr.press/v286/aloui25b.html.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: A review. *arXiv preprint arXiv:2011.08047*, 2020. URL https://arxiv.org/abs/2011.08047.

Issa J. Dahabreh, Sarah E. Robertson, Jon A. Steingrimsson, Elizabeth A. Stuart, and Miguel A. Hernan. Extending inferences from a randomized trial to a new target population. *arXiv preprint arXiv:1805.00550*, 2018. URL https://arxiv.org/abs/1805.00550.

Xavier d'Haultfoeuille. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15, 2010. doi: 10.1016/j.jeconom.2009.06.005. URL https://doi.org/10.1016/j.jeconom.2009.06.005.

Evangelos Dimitriou, Edwin Fong, Jens Magelund Tarp, Karla DiazOrdaz, and Brieuc Lehmann. Causal-ICM: A data fusion framework for heterogeneous treatment effect estimation with multi-task gaussian processes. In Bijan Mazaheri and Niels Richard Hanson (eds.), *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 248–276. PMLR, 2026. URL https://proceedings.mlr.press/v323/dimitriou26a.html.

Jake Fawkes, Michael O'Riordan, Athanasios Vlontzos, Oriol Corcoll, and Ciarán Mark Gilligan-Lee. The hardness of validating observational studies with experimental data. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1819–1827. PMLR, 2025. URL https://proceedings.mlr.press/v258/fawkes25b.html.

Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022. URL https://arxiv.org/abs/2202.12891.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *arXiv preprint arXiv:1810.11646*, 2018. URL https://arxiv.org/abs/1810.11646.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116. URL https://doi.org/10.1073/pnas.1804597116.

Quinn Lanners, Cynthia Rudin, Alexander Volfovsky, and Harsh Parikh. Data fusion for partial identification of causal effects, 2025. URL https://arxiv.org/abs/2505.24296.

Dasom Lee, Shu Yang, Lin Dong, Xiaofei Wang, Donglin Zeng, and Jianwen Cai. Improving trial generalizability using observational studies. *arXiv preprint arXiv:2003.01242*, 2020. doi: 10.48550/arXiv.2003.01242. URL https://arxiv.org/abs/2003.01242.

Wei Li, Wang Miao, and Eric Tchetgen Tchetgen. Non-parametric inference about mean functionals of non-ignorable non-response data without identifying the joint distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):913–935, 2023. doi: 10.1093/jrsssb/qkad047. URL https://arxiv.org/abs/2110.05776.

Xi Lin, Jens Magelund Tarp, and Robin J. Evans. Combining experimental and observational data through a power likelihood. *Biometrics*, 81(1):ujaf008, 2025. doi: 10.1093/biomtc/ujaf008. URL https://doi.org/10.1093/biomtc/ujaf008.

Wang Miao and Eric J. Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016. doi: 10.1093/biomet/asw016. URL https://arxiv.org/abs/1509.04508.

Wang Miao, Lan Liu, Eric Tchetgen Tchetgen, and Zhi Geng. Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*, 2015. doi: 10.48550/arXiv.1509.02556. URL https://arxiv.org/abs/1509.02556.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018. doi: 10.1093/biomet/asy038. URL https://arxiv.org/abs/1609.08816.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. doi: 10.1093/biomet/asaa076. URL https://doi.org/10.1093/biomet/asaa076.

Anton Nilsson, Jonas Björk, and Carl Bonander. Proxy variables and the generalizability of study results. *American Journal of Epidemiology*, 192(3):448–454, 2023. doi: 10.1093/aje/kwac200. URL https://doi.org/10.1093/aje/kwac200.

Zehao Su, Helene C. Rytgaard, Henrik Ravn, and Frank Eriksson. Proximal indirect comparison. *Biometrika*, 113(1):asaf044, 2026. doi: 10.1093/biomet/asaf044. URL https://arxiv.org/abs/2405.10773.

BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric Tchetgen Tchetgen. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28(4):1965–1983, 2018. doi: 10.5705/ss.202016.0324. URL https://arxiv.org/abs/1607.03197.

Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 24(3):1097–1116, 2014. doi: 10.5705/ss.2012.074. URL https://doi.org/10.5705/ss.2012.074.

Lili Wu and Shu Yang. Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 904–926. PMLR, 2022. URL https://proceedings.mlr.press/v177/wu22a.html.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51, 2025. doi: 10.1080/01621459.2024.2393466. URL https://doi.org/10.1080/01621459.2024.2393466.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv preprint arXiv:2005.10579*, 2020. URL https://arxiv.org/abs/2005.10579.

Shu Yang, Siyi Liu, Donglin Zeng, and Xiaofei Wang. Data fusion methods for the heterogeneity of treatment effect and confounding function. *Bernoulli*, 31(4):2987–3012, 2025. doi: 10.3150/24-BEJ1835. URL https://doi.org/10.3150/24-BEJ1835.

A Additional Details

A.1 Shadow Recovery Equation

The selection density ratio r_a in equation 4 and the reference-normalized odds ratio OR_a in Section 4 describe the same outcome-law tilt up to a factor depending on x. Choose y_0 with $0 < r_a(x, y_0) < \infty$; for continuous outcomes use a version of $r_a(x, \cdot)$ continuous at a support point y_0, and for discrete outcomes use a point of positive probability. The reference may depend on a, x. Constrain $\widehat{\mathrm{OR}}_a(x, y_0) = 1$ and set

$$\widehat{r}_a(x, y) = \frac{\widehat{\mathrm{OR}}_a(x, y)}{\int \widehat{\mathrm{OR}}_a(x, u)\widehat{f}_a(u \mid x, S = \mathrm{RCT})\, du}. \tag{20}$$

The corresponding normalized target outcome law is

$$f_a^{\mathrm{SV}}(y \mid x, z, S = \mathrm{OBS}) = \frac{\mathrm{OR}_a(x, y) f_a(y \mid x, z, S = \mathrm{RCT})}{\int \mathrm{OR}_a(x, u) f_a(u \mid x, z, S = \mathrm{RCT})\, du}. \tag{21}$$

These conditions serve only the odds-ratio parameterization: multiplication of $\mathrm{OR}_a(x, \cdot)$ by a positive function of x cancels from equation 21. The additive correction b_{OBS} instead concerns confounding within the target population and enters only after recovery of the target effect.

A.2 CFSC Estimation Algorithm

Algorithm 1 presents the density-based formulation of Shadow Recovery and Structured Correction. Here n_s denotes the size of the estimation sample from source s. The experiments instantiate the recovery step through the conditional-mean model and explicit moment inversion described in Appendix A.6.1; the corresponding fitting splits and correction convention are specified there. Screening diagnostics assess empirical relevance and recovery stability, rather than verifying the shadow restriction or completeness.

Algorithm 1 CFSC: density-based formulation

Input: RCT and OBS samples $\mathcal{D}_{\mathrm{RCT}}$ and $\mathcal{D}_{\mathrm{OBS}}$; candidate set $\mathcal{Z} = \{Z_1, \ldots, Z_M\}$; observational learner $\mathcal{Q}$; positive tilt classes; correction class $\mathcal{H}$, penalty $\mathcal{R}$, and coefficient λ; prespecified screening and stability criteria.

Output: Target CATE $\widehat{\tau}_{\mathrm{OBS}}(x)$ and ATE $\widehat{\psi}_{\mathrm{OBS}}$.

1. Partition both source samples into independent screening and estimation splits; subdivide the screening data into fitting and validation portions.
2. On the screening split, discard each Z_j that adds insufficient outcome information beyond X in either randomized arm. For every retained candidate, fit the arm-specific RCT outcome kernels and the observable source shift, solve the empirical recovery equation, and evaluate its held-out moment residual and tilt stability. Select the candidate $\widehat{Z}$ using the prespecified criteria.
3. Using the estimation split, fit the observational learner and form $\widehat{w}_{\mathrm{OBS}} = \mathcal{Q}(\mathcal{D}_{\mathrm{OBS}})$.
4. For each arm $a \in \{0, 1\}$, using $Z = \widehat{Z}$:
   a. fit $\widehat{f}_a(y \mid x, z, S = \mathrm{RCT})$ and the marginal $\widehat{f}_a(y \mid x, S = \mathrm{RCT})$;
   b. estimate $\widehat{d}(x, z) = \widehat{p}(z \mid x, S = \mathrm{OBS})/\widehat{p}(z \mid x, S = \mathrm{RCT})$;
   c. solve $(\widehat{K}_{a,x}\widehat{r}_a)(z) \approx \widehat{d}(x, z)$ over the positive tilt class, using the reference-normalized parameterization in equation 20;
   d. set
$$\widehat{\mu}_a^{\mathrm{SV}}(x) = \int y\, \widehat{r}_a(x, y)\widehat{f}_a(y \mid x, S = \mathrm{RCT})\, dy.$$
5. Form the recovered target-effect supervision $\widetilde{\tau}_{\mathrm{SV}}(x) = \widehat{\mu}_1^{\mathrm{SV}}(x) - \widehat{\mu}_0^{\mathrm{SV}}(x)$ and, for every unit in the RCT estimation split, the correction label $\widetilde{b}_i = \widehat{w}_{\mathrm{OBS}}(X_i) - \widetilde{\tau}_{\mathrm{SV}}(X_i)$.
6. Fit the structured correction
$$\widehat{b} = \arg\min_{b \in \mathcal{H}} \left[\frac{1}{n_{\mathrm{RCT}}} \sum_{i: S_i = \mathrm{RCT}} \{\widetilde{b}_i - b(X_i)\}^2 + \lambda \mathcal{R}(b) \right].$$
7. Return $\widehat{\tau}_{\mathrm{OBS}}(x) = \widehat{w}_{\mathrm{OBS}}(x) - \widehat{b}(x)$ and $\widehat{\psi}_{\mathrm{OBS}} = n_{\mathrm{OBS}}^{-1} \sum_{i: S_i = \mathrm{OBS}} \widehat{\tau}_{\mathrm{OBS}}(X_i)$.

A.3 Estimation regularity conditions

We collect sufficient conditions for the consistency results. Within each source s, observations are assumed independent and identically distributed (i.i.d.) from $P(X, Z, A, Y \mid S = s)$; the two source samples are independent. Both sample sizes diverge, and their X laws have the same compact support $\mathcal{X}$. Recovery conditions below apply on the relevant joint support of (X, Z), while $\|\cdot\|_\infty$ denotes the supremum norm on the stated support.

Recovery conditions. Define the global operator by $(K_a r)(x, z) = (K_{a,x} r(x, \cdot))(z)$. Choose versions for which the population recovery identities hold throughout these supports. For each arm, the true and fitted ratios belong with probability tending to one to a fixed class $\mathcal{C}_a$ of functions of (x, y), compact in the uniform norm, with $0 < c \le r \le C < \infty$. On these versions, K_a is injective on $\mathcal{C}_a$ in the uniform norm. This stability requirement strengthens the almost-sure identification condition. If the odds-ratio parameterization is used, the reference condition above also holds.

The fitted RCT conditional outcome densities are proper densities and satisfy

$$\sup_{x,z} \int (1 + |y|) \left| \widehat{f}_a(y \mid x, z, S = \mathrm{RCT}) - f_a(y \mid x, z, S = \mathrm{RCT}) \right| dy = o_p(1), \tag{22}$$

$$\sup_{x,z} \int (1 + |y|) f_a(y \mid x, z, S = \mathrm{RCT})\, dy \le M < \infty. \tag{23}$$

The fitted marginal RCT outcome density, obtained by integrating its conditional counterpart over the RCT distribution of $Z \mid X$, is proper and satisfies

$$\sup_{x} \int (1 + |y|) \left| \widehat{f}_a(y \mid x, S = \mathrm{RCT}) - f_a(y \mid x, S = \mathrm{RCT}) \right| dy = o_p(1).$$

Let $(\widehat{K}_a r)(x,z) = \int r(x,y)\widehat{f}_a(y \mid x,z,S=\mathrm{RCT})\,dy$, and let $\widehat{d}$ estimate the observable shadow density ratio. We require

$$\|\widehat{d} - d\|_\infty = o_p(1), \qquad \|\widehat{K}_a\widehat{r}_a - \widehat{d}\|_\infty = o_p(1), \tag{24}$$

with numerical integration error vanishing uniformly over the full covariate support.

Correction conditions. OBS treatment overlap holds on $\mathcal{X}$, and $\|\widehat{w}_{\mathrm{OBS}} - w_{\mathrm{OBS}}\|_\infty = o_p(1)$. The correction class $\mathcal{H} \subset C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of continuous real-valued functions, is compact in the uniform norm and contains b_{OBS}. The RCT covariate law has full support on $\mathcal{X}$. With the nonnegative penalty and coefficient specified in Section 4, the fitted correction approximately minimizes equation 16 with $o_p(1)$ optimization error and $\lambda\mathcal{R}(b_{\mathrm{OBS}}) = o_p(1)$.

A.4 Consistency Guarantees

Identification alone does not guarantee stable estimation because Shadow Recovery is an inverse problem. Let $\mathcal{X}$ be the common covariate support and $\|g\|_\infty = \sup_{x\in\mathcal{X}} |g(x)|$. We require consistent nuisance estimation, a stable tilt class, vanishing recovery and numerical errors, and a correctly specified correction class. Appendix A.3 gives sufficient conditions.

Proposition 1 (Consistency of Shadow Recovery)**.** *Under the conditions of Theorem 1, together with the recovery regularity conditions in Appendix A.3, the recovered response surfaces and target-effect signal are uniformly consistent:*

$$\|\widehat{\mu}_a^{\mathrm{SV}} - \mu_a^{\mathrm{OBS}}\|_\infty = o_p(1), \qquad \|\widetilde{\tau}_{\mathrm{SV}} - \tau_{\mathrm{OBS}}\|_\infty = o_p(1). \tag{25}$$

Theorem 2 (Consistency of the calibrated CATE)**.** *Under Proposition 1 and the correction regularity conditions in Appendix A.3, including $\|\widehat{w}_{\mathrm{OBS}} - w_{\mathrm{OBS}}\|_\infty = o_p(1)$ and $b_{\mathrm{OBS}} \in \mathcal{H}$, the fitted correction and final CATE satisfy*

$$\|\widehat{b} - b_{\mathrm{OBS}}\|_\infty = o_p(1), \qquad \|\widehat{\tau}_{\mathrm{OBS}} - \tau_{\mathrm{OBS}}\|_\infty = o_p(1). \tag{26}$$

Corollary 1 (Plug-in ATE consistency)**.** *Under Theorem 2, the plug-in target ATE in equation 18 satisfies $\widehat{\psi}_{\mathrm{OBS}} \xrightarrow{p} \psi_{\mathrm{OBS}}$.*

The result follows from two error reductions. First, Proposition 1 makes the raw correction labels in equation 15 converge to $b_{\mathrm{OBS}}(X)$ on RCT covariate values. Second, uniform consistency of the correction regression propagates this discrepancy over the common support. The final error is therefore bounded by

$$\|\widehat{\tau}_{\mathrm{OBS}} - \tau_{\mathrm{OBS}}\|_\infty \le \|\widehat{w}_{\mathrm{OBS}} - w_{\mathrm{OBS}}\|_\infty + \|\widehat{b} - b_{\mathrm{OBS}}\|_\infty.$$

Full proofs are given in Appendix A.5. The theorem establishes consistency on common support; it does not by itself provide rates, confidence intervals, efficiency guarantees, or extrapolation beyond RCT support.

For the fixed finite candidate set $\mathcal{Z}$ screened during Shadow Recovery in Section 4, let $\widehat{Z}$ be the selected candidate and $\mathcal{Z}_{\mathrm{valid}} \subseteq \mathcal{Z}$ those satisfying Assumptions 2 and 3, the outcome-support condition in Section 3.2, and the stated regularity conditions. With independent screening and estimation splits, the same guarantees hold provided

$$\Pr(\widehat{Z} \in \mathcal{Z}_{\mathrm{valid}}) \longrightarrow 1. \tag{27}$$

This condition clarifies the role of screening: empirical relevance and recovery diagnostics can prioritize plausible candidates, but the consistency claim remains conditional on asymptotically valid selection.

Corollary 2 (Consistency with screened shadow selection)**.** *If equation 27 holds and the fixed-candidate regularity conditions hold for each valid candidate, the screened CATE and ATE estimators inherit Theorem 2 and Corollary 1.*

A.5 PROOFS

A.5.1 PROOF OF THEOREM 1

Fix a shadow candidate Z satisfying the shadow assumptions, an arm a, and a value x for which the conditions hold. Write $f_a(y, z \mid x, S = s)$ for the joint density of $(Y(a), Z)$ conditional on $X = x, S = s$. Consistency and joint randomization of treatment and Z given X imply

$$f_a(y, z \mid x, S = \mathrm{RCT}) = p(y \mid A = a, X = x, Z = z, S = \mathrm{RCT})\, p(z \mid x, S = \mathrm{RCT}).$$

Both factors are observed, so the joint potential-outcome law, its conditional kernel $f_a(y \mid x, z, S = \mathrm{RCT})$, and $K_{a,x}$ are identified. The second factor uses all trial units, rather than the distribution of Z within arm a; using it here relies on the joint randomization condition in Assumption 1(ii). Marginalizing over z identifies $f_a(y \mid x, S = \mathrm{RCT})$, and Bayes' rule identifies $q_a(z \mid x, y)$.

Let $q_a(z \mid x, y)$ denote the common conditional law of Z given $X = x, Y(a) = y$ in the two sources, as implied by equation 3. Factoring the joint densities gives

$$\begin{aligned} f_a(y, z \mid x, S = \mathrm{OBS}) &= q_a(z \mid x, y) f_a(y \mid x, S = \mathrm{OBS}) \\ &= r_a(x, y) f_a(y, z \mid x, S = \mathrm{RCT}). \end{aligned} \tag{28}$$

The one-sided outcome-support condition and the common kernel imply that the target law of $Z \mid X = x$ is absolutely continuous with respect to its source law. Thus $d(x, z)$ exists and is finite source-almost surely, and it is positive target-almost surely; no reverse absolute-continuity condition is needed for identification. Integration over y and division by $p(z \mid x, S = \mathrm{RCT})$ yield equation 7. The ratio d is observable because Z is observed in both sources. If r'_a is another admissible solution, then $K_{a,x}(r_a - r'_a) = 0$, so Assumption 3 implies equality almost surely. Existence follows from the true full-data law. Integrating the moment equation over the RCT distribution of Z also shows

$$\int r_a(x, y) f_a(y \mid x, S = \mathrm{RCT})\, dy = 1.$$

Bayes' rule gives the selection odds

$$\frac{\Pr(S = \mathrm{OBS} \mid X = x, Y(a) = y)}{\Pr(S = \mathrm{RCT} \mid X = x, Y(a) = y)} = r_a(x, y) \frac{\Pr(S = \mathrm{OBS} \mid X = x)}{\Pr(S = \mathrm{RCT} \mid X = x)}.$$

When the reference condition stated above is used, taking the ratio at y and y_0 cancels the source-marginal odds and gives OR_a. For continuous outcomes, continuity at the support point y_0 fixes its value from the almost-everywhere identified ratio. This reference condition is unnecessary for identification of the recovered law, since the normalizing factor cancels. The source sampling fractions need not identify absolute population selection probabilities to identify the normalized odds ratio.

Finally, divide equation 28 by $p(z \mid x, S = \mathrm{OBS})$ to obtain equation 9. Since

$$\int \mathrm{OR}_a(x, u) f_a(u \mid x, z, S = \mathrm{RCT})\, du = \frac{d(x, z)}{r_a(x, y_0)},$$

this is precisely equation 21. This proves both uniqueness of the selection tilt and identification of the recovered target outcome distribution.

The recovered conditional distribution equals the law of $Y(a) \mid X, Z, S = \mathrm{OBS}$. Integrating it over $p(z \mid x, S = \mathrm{OBS})$ gives $f_a^{\mathrm{SV}}(y \mid x, S = \mathrm{OBS}) = r_a(x, y) f_a(y \mid x, S = \mathrm{RCT})$, whose first moment is exactly $\mu_a^{\mathrm{OBS}}(x)$ in equation 10. Subtraction across arms gives $\tau_{\mathrm{OBS}}(x)$ in equation 1; averaging over the observed target law of X gives ψ_{OBS}. The first-moment condition in Theorem 1 ensures these means exist.

A.5.2 PROOF OF PROPOSITION 1

Suppress the arm subscript where unambiguous. For every $r \in \mathcal{C}_a$, equation 22 and the uniform bound C give

$$\|\widehat{K}_a r - K_a r\|_\infty \le C \sup_{x,z} \int |\widehat{f}_a - f_a|(y \mid x, z, S = \mathrm{RCT})\, dy = o_p(1),$$

uniformly over $\mathcal{C}_a$. Hence, by equation 24,

$$\|K_a\widehat{r}_a - d\|_\infty \le \sup_{r\in\mathcal{C}_a} \|(K_a - \widehat{K}_a)r\|_\infty + \|\widehat{K}_a\widehat{r}_a - \widehat{d}\|_\infty + \|\widehat{d} - d\|_\infty = o_p(1).$$

The map K_a is continuous in the uniform norm because its kernels are probability densities. For every $\eta > 0$, compactness and injectivity therefore imply

$$\inf_{r\in\mathcal{C}_a:\, \|r-r_a\|_\infty \ge \eta} \|K_a r - d\|_\infty > 0$$

whenever the set is nonempty: otherwise a convergent subsequence would give a second solution to $K_a r = d$. It follows that $\|\widehat{r}_a - r_a\|_\infty = o_p(1)$. The reference normalization gives $\widehat{\mathrm{OR}}_a(x,y) = \widehat{r}_a(x,y)/\widehat{r}_a(x,y_0)$ exactly, so the lower bound c proves uniform consistency of the fitted odds ratio as well.

To propagate this result through the actual normalized tilt, put

$$g_a(y \mid x,z) = r_a(x,y) f_a(y \mid x,z,S=\mathrm{RCT}),$$
$$\widehat{g}_a(y \mid x,z) = \widehat{r}_a(x,y)\widehat{f}_a(y \mid x,z,S=\mathrm{RCT}).$$

Writing $\Delta_a = \sup_{x,z}\int(1+|y|)|\widehat{g}_a - g_a|\,dy$, the preceding bounds yield

$$\Delta_a \le C \sup_{x,z}\int (1+|y|)|\widehat{f}_a - f_a|\,dy + M\|\widehat{r}_a - r_a\|_\infty = o_p(1).$$

The normalizers $D_a = \int g_a\,dy = d$ and $\widehat{D}_a = \int \widehat{g}_a\,dy$ are at least c, and $\|\widehat{D}_a - D_a\|_\infty \le \Delta_a$. Since $\int(1+|y|)g_a\,dy \le CM$, normalization gives

$$\sup_{x,z}\int (1+|y|)|\widehat{f}_a^{\mathrm{SV}} - f_a^{\mathrm{SV}}|(y \mid x,z,S=\mathrm{OBS})\,dy \le (c^{-1} + CMc^{-2})\Delta_a = o_p(1). \tag{29}$$

The cancellation of the normalization in equation 20 shows that $\widehat{g}_a/\widehat{D}_a$ is exactly the fitted law in Section 4.

For the target outcome law conditional on $X = x$, equation 10 and equation 13 give

$$\|\widehat{\mu}_a^{\mathrm{SV}} - \mu_a^{\mathrm{OBS}}\|_\infty \le \sup_x \int |y|\,|\widehat{f}_a^{\mathrm{SV}} - f_a^{\mathrm{SV}}|(y \mid x, S=\mathrm{OBS})\,dy + o_p(1) = o_p(1),$$

by uniform convergence of the tilt and marginal RCT density, where the remainder is numerical integration error. Subtracting across arms proves equation 25 for the CATE conditional on X. These arguments propagate the assumed nuisance convergence to the recovered tilt, outcome law, and target-effect signal.

A.5.3 Proof of Theorem 2

Write $b_0 = b_{\mathrm{OBS}}$ and $\delta_n = \max_{i:S_i=\mathrm{RCT}} |\widetilde{b}_i - b_0(X_i)|$. The definition equation 15, consistency of the base learner, and Proposition 1 imply $\delta_n = o_p(1)$. This is a bound on the actual fitted labels, so independence of those labels from their training observations is not required.

Compactness of $\mathcal{H}$ implies $B = \sup_{b\in\mathcal{H}} \|b\|_\infty < \infty$. Define the empirical loss with true labels and its population counterpart by

$$Q_n(b) = \frac{1}{n_{\mathrm{RCT}}} \sum_{i:S_i=\mathrm{RCT}} \{b_0(X_i) - b(X_i)\}^2, \qquad Q(b) = \mathbb{E}[\{b_0(X) - b(X)\}^2 \mid S=\mathrm{RCT}].$$

Let $\widetilde{Q}_n(b)$ replace $b_0(X_i)$ by the fitted labels. Expanding the squares yields

$$\sup_{b\in\mathcal{H}} |\widetilde{Q}_n(b) - Q_n(b)| \le 4B\delta_n + \delta_n^2 = o_p(1).$$

To verify uniform convergence, $\sup_{b\in\mathcal{H}} |Q_n(b) - Q(b)| = o_p(1)$ follows from a finite uniform-norm cover of the compact class: apply the LLN at its finitely many centers and use the bound

$4B\|b-b'\|_\infty$ for the difference between the corresponding squared losses. Let the cover radius tend to zero. Consequently $\sup_b |\widetilde{Q}_n(b) - Q(b)| = o_p(1)$.

Approximate minimization of the actual penalized objective, nonnegative penalty, and $b_0 \in \mathcal{H}$ now give

$$\widetilde{Q}_n(\widehat{b}) \le \widetilde{Q}_n(b_0) + \lambda\mathcal{R}(b_0) + o_p(1) = o_p(1),$$

and therefore $Q(\widehat{b}) = o_p(1)$. For each $\eta > 0$, the compact subset $\{b \in \mathcal{H} : \|b - b_0\|_\infty \ge \eta\}$ has strictly positive minimum Q. Indeed, any zero-risk continuous function equals b_0 everywhere on $\mathcal{X}$ by full support of the RCT covariate law. This proves $\|\widehat{b} - b_0\|_\infty = o_p(1)$, a stronger statement than mere integrated regression consistency. In particular, it justifies the claimed pointwise convergence; integrated convergence alone would not do so.

Finally, using precisely the subtraction in equation 17,

$$\widehat{\tau}_{\mathrm{OBS}}(x) - \tau_{\mathrm{OBS}}(x) = \{\widehat{w}_{\mathrm{OBS}}(x) - w_{\mathrm{OBS}}(x)\} - \{\widehat{b}(x) - b_{\mathrm{OBS}}(x)\}.$$

The triangle inequality proves equation 26. The result applies on the shared support only. It does not justify extrapolation to covariate values absent from the RCT or convergence to the true correction when $b_{\mathrm{OBS}} \notin \mathcal{H}$.

A.5.4 Proof of Corollary 1

For the empirical target average already defined in equation 18,

$$|\widehat{\psi}_{\mathrm{OBS}} - \psi_{\mathrm{OBS}}| \le \frac{1}{n_{\mathrm{OBS}}} \sum_{i:S_i=\mathrm{OBS}} |\widehat{\tau}_{\mathrm{OBS}}(X_i) - \tau_{\mathrm{OBS}}(X_i)| + \left| \frac{1}{n_{\mathrm{OBS}}} \sum_{i:S_i=\mathrm{OBS}} \tau_{\mathrm{OBS}}(X_i) - \mathbb{E}[\tau_{\mathrm{OBS}}(X) \mid S = \mathrm{OBS}] \right|.$$

The first term is at most the uniform CATE error and tends to zero in probability. The second tends to zero by the LLN and integrability of the target potential outcomes. No independence between the fitted CATE function and these OBS observations is needed for the first bound.

A.5.5 Proof of Corollary 2

For each valid candidate Z_j, let $E_{j,n}$ denote the uniform CATE error of its estimator fitted on the estimation split. By Theorem 2, $E_{j,n} = o_p(1)$. On the event of valid selection, the screened error is bounded by $\max_{j:Z_j \in \mathcal{Z}_{\mathrm{valid}}} E_{j,n}$. For every $\varepsilon > 0$,

$$\Pr(E_{\widehat{Z},n} > \varepsilon) \le \Pr(\widehat{Z} \notin \mathcal{Z}_{\mathrm{valid}}) + \sum_{j:Z_j \in \mathcal{Z}_{\mathrm{valid}}} \Pr(E_{j,n} > \varepsilon) \longrightarrow 0.$$

The first term vanishes by equation 27, and the finite sum vanishes by candidate-wise consistency. The same argument applied to the absolute ATE errors and Corollary 1 proves screened ATE consistency. The independent split allows the fixed-candidate estimation assumptions to be applied without conditioning the fitting data on the screening decision.

A.6 Experimental details

The four synthetic regimes set $(\beta_u, \alpha_{\mathrm{sel}})$ to $(0,0)$, $(1,0)$, $(0,1.5)$, and $(1,1.5)$, respectively. Here β_u controls dependence of OBS treatment on U, while α_{sel} is the coefficient on U in the logistic source-assignment score, $\mathrm{logit}\, P(S = \mathrm{RCT} \mid X, U) = \eta(X) + \alpha_{\mathrm{sel}} U$. The specified target treatment-effect component is

$$\tau_X(X) = 1 + 0.5X_1 - 0.3X_2 + 0.20(X_2^2 - 1) + 0.15X_1X_3,$$

with additional latent effect modification $\tau_u = 0.5$. The shadow proxy is $Z = \rho U + \sqrt{1-\rho^2}\epsilon_Z$ with $\rho = 0.65$. Each dataset has $n = 6{,}000$ and a population RCT fraction of 0.4. Within a synthetic replicate, methods use the same held-out 30% of OBS target rows. The T-learner fits separate quadratic outcome regressions. The R-learner uses quadratic R-loss, three-fold nuisance cross-fitting, and slope penalty 0.03. Causal Forest uses a fixed `CausalForestDML` implementation.

RHC and CFSC share the fitted observational R-learner and use a linear ridge correction. Integrative-R uses its repository implementation with linear target and bias bases. CFSC uses linear-X shadow screening, quadratic-X conditional-mean recovery, and $\lambda_c = 5$ for the correction, as detailed below.

For completeness, the absolute ATE errors (mean $\pm$ SD) in Neither, OBS confounding, Source selection, and Both are 0.054 ± 0.038, 0.089 ± 0.055, 0.587 ± 0.053, and 0.611 ± 0.062 for PO-GP, and 0.061 ± 0.040, 0.061 ± 0.040, 0.547 ± 0.058, and 0.547 ± 0.058 for Causal-ICM.

For ACTG and JTPA, 20% of original records are first reserved for choosing the sampling design. Among the remaining records, source membership is sampled with $P(S = \mathrm{RCT}) = \mathrm{expit}(a_0 + V_{\mathrm{std}})$, where the intercept sets the expected RCT share to 30% for ACTG and 20% for JTPA. The source-selection variable V is baseline stratification for ACTG and age for JTPA. In each of five target-population folds, the OBS training portion is retained with probability $\mathrm{expit}\{b_0 + (0.5 - A)Y_{\mathrm{std}}\}$, with the intercept chosen for an expected 70% retention. Evaluation folds are left intact and use randomized treatment and observed outcomes to form the AIPW ATE reference. The shadow Z is baseline CD4 (`cd40`) for ACTG and prior earnings (`prevearn`) for JTPA. V and Z are excluded from the common predictor set X, and only CFSC uses Z in addition to X. No treatment, outcome, or shadow value is regenerated or changed.

A.6.1 CFSC IMPLEMENTATION THROUGH CONDITIONAL MEANS

Recovery model and identification relation. Let $\phi_d(x)$ contain all nonconstant monomials of total degree at most d, including pairwise interactions when $d = 2$. The experimental recovery model is

$$E[Z \mid X = x, Y(a) = y, S = \mathrm{RCT}] = \ell_a(x) + \beta_a y, \qquad \ell_a(x) = \eta_a + B_a \phi_d(x). \tag{30}$$

Here Z may be vector-valued and β_a is constant in x. Under the shadow exclusion condition, this conditional mean also applies in the target source on the relevant support. With finite first moments, iterated expectation gives

$$h_O(x) := E[Z \mid X = x, S = \mathrm{OBS}] = \ell_a(x) + \beta_a \mu_a^O(x), \qquad \mu_a^O(x) = E[Y(a) \mid X = x, S = \mathrm{OBS}]. \tag{31}$$

This is the first-moment implication of the shadow mixture identity underlying the recovery equation. Under model equation 30 and $\beta_a \neq 0$, it identifies the target mean directly. The experimental procedure estimates this mean functional; it does not fit an outcome density or an outcome-tilt function. The density-based formulation above describes recovery of the full outcome law, whereas the direct calculation here additionally specifies its reverse conditional mean.

Regression fits and numerical inversion. For each randomized arm, ordinary least squares regresses Z on $[1, \phi_d(X), Y]$. A separate quadratic least-squares regression of Z on X in the OBS recovery sample, pooling both treatment arms, estimates h_O. Let $\widehat{\Sigma}_a$ be the sample covariance of the arm-specific reverse-regression residuals plus $10^{-8} I$. The implemented calculation is

$$\widehat{v}_a = \widehat{\Sigma}_a^{-1} \widehat{\beta}_a, \qquad \widehat{\mu}_a^{\mathrm{SV}}(x) = \frac{\widehat{v}_a^\top \{\widehat{h}_O(x) - \widehat{\ell}_a(x)\}}{\widehat{\beta}_a^\top \widehat{v}_a}, \qquad \widetilde{\tau}_{\mathrm{SV}}(x) = \widehat{\mu}_1^{\mathrm{SV}}(x) - \widehat{\mu}_0^{\mathrm{SV}}(x). \tag{32}$$

The code computes $\widehat{v}_a$ by a linear solve and rejects an arm with $\widehat{\beta}_a^\top \widehat{v}_a < 10^{-6}$. For the scalar shadows used here, the recovered mean simplifies to $(\widehat{h}_O - \widehat{\ell}_a)/\widehat{\beta}_a$. No clipping is applied to the recovered means. The denominator normalizes the inverse coefficient, rather than a probability density.

Synthetic screening and fitting splits. The screened specification considers Z and two independent noise candidates. Each source training sample is partitioned into 30% screening and 70% recovery subsets, stratifying the RCT split by treatment. Screening subsets are further divided into 60% fitting and 40% validation portions. A candidate must have absolute residual outcome correlation at least 0.1 in both randomized arms after linear adjustment for X. Screening uses linear-X reverse regressions and quadratic OBS shadow-mean regressions. Write M for the mean squared difference between the implied shadow means and a separately fitted OBS validation mean, divided by the OBS validation variance (floored at 10^{-8}). Write T for the maximum, over arms, of the mean squared difference between the recovered mean and a linear RCT outcome regression on validation

covariates, divided by the arm's training residual mean square (also floored at 10^{-8}). Candidates satisfying $M \leq 1$ and $T \leq 25$ are ranked by $M + 0.01T$, breaking ties by column name; an empty eligible set raises an error. These are empirical mean-prediction and stability diagnostics. Final synthetic recovery uses $d = 2$ on the recovery subsets. The source-selection sweep supplies the prespecified shadow Z directly and retains the same recovery split.

Observational backbone and structured correction. CFSC uses the quadratic OBS R-learner described below, with three-fold nuisance cross-fitting, propensity clipping to $[0.05, 0.95]$, and mean R-loss slope penalty 0.03. Its population R-loss target, with the population nuisance functions and unrestricted effect regression, is the factual OBS outcome contrast $E[Y \mid X, A = 1, S = \mathrm{OBS}] - E[Y \mid X, A = 0, S = \mathrm{OBS}]$. The fitted backbone uses the full OBS training sample. Recovery predictions are evaluated at all RCT training covariates, where the code fits

$$(\widehat{c}_0, \widehat{c}) = \arg\min_{c_0, c} \left[\frac{1}{n_R} \sum_{i \in \mathcal{D}_R} \{\widetilde{\tau}_{\mathrm{SV}}(X_i) - \widehat{w}_O(X_i) - c_0 - X_i^\top c\}^2 + 5\|c\|_2^2 \right]. \tag{33}$$

The intercept is unpenalized; the summed-loss ridge parameter is $5n_R$, and the solver uses an SVD. Prediction is $\widehat{\tau}_O(x) = \widehat{w}_O(x) + \widehat{c}_0 + x^\top \widehat{c}$, equivalent to the manuscript's subtraction convention with $\widehat{b}(x) = -(\widehat{c}_0 + x^\top \widehat{c})$. Thus the screening/recovery partition applies to the recovery fits; the backbone and correction use their full respective training samples. The correction labels are fitted recovery predictions, without an additional cross-fitting stage.

ACTG and JTPA specifications. The design-stage shadows are held fixed across the evaluation folds: baseline CD4 for ACTG and prior earnings for JTPA. Reverse regression uses $d = 1$ for ACTG and $d = 2$ for JTPA; both use quadratic OBS shadow-mean regression and unpenalized reverse-regression coefficients. Within each fold, recovery uses 70% of the RCT sample and 70% of the retained OBS training sample. In particular, the fitted shadow mean uses the post-thinning OBS distribution, while evaluation uses the intact held-out target fold. No additional unthinned covariate sample enters this recovery regression. Covariates and shadows are standardized using pooled RCT and retained OBS training observations, and outcomes are centered and divided by the pooled sample standard deviation. Correction uses these standardized covariates with $\lambda_c = 5$; predictions are multiplied by the outcome standard deviation to restore original units. Evaluation-fold outcomes enter the evaluation scores, not CFSC fitting.

A.7 Baseline methods and implementation details

We describe the estimators in Tables 1 and 3, distinguishing their statistical motivation from the implementations evaluated here. Write $G = \mathbf{1}\{S = \mathrm{RCT}\}$, and let $\mathcal{D}_R$ and $\mathcal{D}_O$ denote the RCT and OBS training samples. Baseline predictions use the common covariates X; neither latent variables nor the auxiliary shadow Z enter their fitting procedures. Target ATE predictions average the fitted effect over evaluation covariates.

A.7.1 Observational learners

T-learner. The T-learner (Künzel et al., 2019) fits a separate outcome regression $\widehat{\mu}_a^O(x)$ in each OBS treatment arm and returns $\widehat{\mu}_1^O(x) - \widehat{\mu}_0^O(x)$. Our OBS-only specification uses unpenalized quadratic regressions with pairwise interactions and intercepts. It learns the observational conditional outcome contrast; interpreting this contrast causally requires treatment exchangeability given X and overlap in OBS.

R-learner. The R-learner (Nie & Wager, 2021) residualizes outcome and treatment before fitting effect heterogeneity. Let $\widehat{m}^{(-k)}$ and $\widehat{e}^{(-k)}$ estimate the OBS conditional outcome mean and treatment propensity without fold k. With standardized quadratic features $\phi_2(x)$, our objective is

$$\frac{1}{n_O} \sum_{i \in \mathcal{D}_O} \left[Y_i - \widehat{m}^{(-k(i))}(X_i) - \{A_i - \widehat{e}^{(-k(i))}(X_i)\}\{\theta_0 + \phi_2(X_i)^\top \theta\} \right]^2 + 0.03\|\theta\|_2^2. \tag{34}$$

We use three treatment-stratified folds, quadratic ridge outcome nuisances (summed-loss penalty 1), and standardized logistic propensity models with $C = 1$. Propensities are clipped to $[0.05, 0.95]$.

The effect intercept is unpenalized; the corresponding summed-loss ridge coefficient is $0.03n_O$. Residualization reduces sensitivity to nuisance estimation error under the R-learner assumptions, but does not remove latent treatment confounding.

Causal Forest. We use `econml.CausalForestDML`, which combines nuisance residualization with adaptive forest neighborhoods for heterogeneous-effect estimation (Athey et al., 2019). The fixed specification uses 100 trees, minimum leaf size 10, discrete treatment, and three-fold nuisance fitting. Outcome and treatment nuisances are linear regression and standardized logistic regression, respectively. Only OBS observations enter the fit. Its causal interpretation also relies on measured-covariate treatment exchangeability and overlap.

A.7.2 TRIAL WEIGHTING

IPSW. Inverse probability of sampling weighting adjusts observed source imbalance (Dahabreh et al., 2018). A logistic classifier $\widehat{\pi}(x) = \widehat{P}(G = 1 \mid X = x)$ is fitted to pooled training covariates. For the OBS target, trial observations receive inverse-odds weights

$$\widetilde{w}_i = \frac{1 - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)}, \qquad i \in \mathcal{D}_R.$$

Probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$, raw weights to $[0.05, 20]$, and weights are then normalized to mean one. They approximate the OBS-to-RCT covariate density ratio up to a constant. This addresses measured covariate shift under source overlap and conditional transportability, but cannot balance an unobserved effect modifier.

CW. Calibration weighting matches observed target covariate moments (Lee et al., 2020). Let $h(x)$ contain an intercept and pooled-standardized first-order covariates, and let $\bar{h}_O$ be their OBS mean. Our entropy-calibration implementation solves

$$\widehat{\lambda} = \arg\min_{\lambda} \left\{ \log\left(\frac{1}{n_R} \sum_{i \in \mathcal{D}_R} e^{\lambda^\top h(X_i)} \right) - \lambda^\top \bar{h}_O \right\}, \qquad w_i \propto e^{\widehat{\lambda}^\top h(X_i)}.$$

We use BFGS and first-order balance constraints, without quadratic or interaction terms. The mean-one weights are clipped to $[10^{-6}, 10^6]$ and renormalized; clipping and numerical tolerance can leave residual imbalance.

CATE extension. For both IPSW and CW, we fit a weighted *linear* outcome regression within each RCT treatment arm and subtract the fitted means. OBS outcomes do not enter these regressions. These are regression-based CATE extensions of sampling/calibration weighting, rather than the original marginal weighted-ATE estimators. Their outcome basis is linear, whereas the OBS-only T-learner is quadratic, so the comparison also reflects outcome-model specification.

A.7.3 RANDOMIZED CORRECTION AND INTEGRATIVE LEARNING

RHC. Experimental grounding (Kallus et al., 2018) uses randomized evidence to correct a potentially biased observational effect estimate. Our RHC starts from the same fitted OBS R-learner $\widehat{w}_O$ as CFSC. Separate linear RCT arm regressions give $\widehat{\mu}_a^R$, and the clipped empirical trial treatment fraction gives $\widehat{p}_R$. The supervision signal is

$$\widetilde{\tau}_i^R = \widehat{\mu}_1^R(X_i) - \widehat{\mu}_0^R(X_i) + \frac{A_i\{Y_i - \widehat{\mu}_1^R(X_i)\}}{\widehat{p}_R} - \frac{(1 - A_i)\{Y_i - \widehat{\mu}_0^R(X_i)\}}{1 - \widehat{p}_R}.$$

The arm regressions use the trial training sample without additional cross-fitting. A linear correction minimizes

$$\frac{1}{n_R} \sum_{i \in \mathcal{D}_R} \{\widetilde{\tau}_i^R - \widehat{w}_O(X_i) - c_0 - X_i^\top c\}^2 + 5\|c\|_2^2.$$

The fitted effect is

$$\widehat{\tau}_{\mathrm{RHC}}(x) = \widehat{w}_O(x) + \widehat{c}_0 + x^\top \widehat{c}.$$

The intercept is unpenalized. This additive convention is equivalent to subtracting a bias with the opposite sign. The approach relies on a sufficiently simple correction and trial supervision appropriate for the target. Under latent source-dependent effect modification, it can correct toward the trial effect rather than the OBS effect. CFSC replaces this direct trial signal with shadow-recovered target supervision while retaining the backbone and correction class.

Integrative-R. The integrative R-learner (Wu & Yang, 2022) jointly models a causal effect and an OBS confounding function. We use the repository implementation without a selection-correction plugin. Pooled linear outcome and logistic treatment models on (X, G) give residuals $r_{Yi} = Y_i - \widehat{m}(X_i, G_i)$ and $r_{Ai} = A_i - \widehat{e}(X_i, G_i)$, with propensity clipping to $[0.05, 0.95]$. For $\phi_1(x) = (1, x^\top)^\top$, it solves

$$\min_{\theta,\gamma} \sum_{i\in\mathcal{D}_R\cup\mathcal{D}_O} \left[r_{Yi} - r_{Ai}\{\phi_1(X_i)^\top\theta + (1-G_i)\phi_1(X_i)^\top\gamma\}\right]^2 + 10^{-4}\|\theta_{-0}\|_2^2 + 10^{-2}\|\gamma_{-0}\|_2^2. \tag{35}$$

Intercepts are excluded from the penalties. The prediction is $\phi_1(x)^\top\widehat{\theta}$; the second component represents OBS treatment-confounding bias. The adapter uses in-sample pooled nuisance fits and linear effect and bias bases. The shared causal component assumes conditional transportability of trial effects.

A.7.4 Gaussian-process fusion

Pseudo-outcome GP (PO-GP). The approach of Fawkes et al. (2025) learns a smooth correction to an observational effect estimate from randomized pseudo-outcomes. Following the author implementation, the initial estimate is $\widehat{w}_O(x) = x^\top\widehat{\beta}_1$, obtained by OBS ordinary least squares on $[X, AX]$ without an intercept. This differs from the RHC/CFSC R-learner backbone. With known randomized treatment probability p, trial labels are

$$D_i = \frac{A_i - p}{p(1-p)}Y_i - \widehat{w}_O(X_i).$$

The fixed three-task linear model of coregionalization has covariance

$$K\{(x,a),(x',a')\} = k_\delta(x,x') + v_a v_{a'} k_\nu(x,x'), \qquad (v_0, v_1, v_2) = \left(-\frac{1}{1-p}, \frac{1}{p}, 0\right).$$

Tasks 0 and 1 represent control and treated pseudo-outcomes; task 2 extracts the common correction without the treatment-dependent nuisance term. Both spatial kernels are ARD RBF kernels. A constant mean and shared Gaussian likelihood are fitted by exact marginal likelihood using Adam (learning rate 0.1, 2,000 iterations), double precision, and exact Cholesky factorization. Prediction adds the task-2 posterior mean to $\widehat{w}_O(x)$. We use $p = 0.5$ for simulations and the ACTG contrast, and $p = 2/3$ for JTPA. Smoothness regularizes the correction but does not by itself identify an OBS-specific effect under nonignorable source selection.

Causal-ICM. Causal-ICM (Dimitriou et al., 2026) borrows information between randomized and observational outcome surfaces using multi-task GPs. Each treatment arm has a separate GP with kernel

$$K_a\{(x,s),(x',s')\} = k_a^{\mathrm{RBF}}(x,x')B_{ss'}, \qquad B = WW^\top + 10^{-9}I_2, \qquad W = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}.$$

Within each arm, task 0 is RCT and task 1 is OBS. The cross-source correlation controls borrowing from the observational surface. We use the authors' rank-2 construction, isotropic RBF kernels, and optimizer, fixing $\rho = 0.5$ without cross-validation and optimizing the kernel and likelihood parameters. Both arm predictions query RCT task 0, so $\widehat{\tau}(x) = \widehat{\mu}_{1,R}(x) - \widehat{\mu}_{0,R}(x)$. This is the fixed-borrowing specification of Causal-ICM.

Provenance and evaluation. The GP adapters load pinned author code: PO-GP from the PO-GP repository (revision `660cbf6`), and Causal-ICM from the Causal-ICM repository (revision

2554b02). The latter contains committed merge-conflict markers; the loader retains the HEAD branch consistently and records original and loaded source hashes. For PO-GP, a least-squares solver replaces explicit matrix inversion. Fitting uses training samples and prediction uses held-out target covariates. Real-data covariate/outcome scaling is fitted on pooled RCT and retained OBS training observations, and effects are returned to original outcome units. The five real-data outer folds evaluate predictions and are distinct from nuisance cross-fitting or hyperparameter selection. Their AIPW references support ATE-level error and AUTOC, not individual-effect PEHE.

A.8 SOURCE-SELECTION SENSITIVITY

Table 4: Source-selection sensitivity: target-OBS PEHE (mean $\pm$ SD over 20 repetitions; lower is better). OBS confounding is fixed at $\beta_u = 1$, and CFSC uses the known valid shadow variable throughout. This separate sweep should not be combined numerically with the four-regime comparison in Table 1.

	α_{sel}				
Method	0	0.5	1.0	1.5	2.0
RHC	0.161 ± 0.030	0.300 ± 0.050	0.467 ± 0.049	0.578 ± 0.044	0.642 ± 0.042
Integrative-R	0.340 ± 0.020	0.391 ± 0.034	0.523 ± 0.039	0.619 ± 0.037	0.683 ± 0.035
PO-GP	0.429 ± 0.044	0.536 ± 0.059	0.643 ± 0.060	0.729 ± 0.062	0.769 ± 0.047
Causal-ICM	0.154 ± 0.022	0.273 ± 0.060	0.440 ± 0.060	0.560 ± 0.058	0.630 ± 0.051
CFSC (ours)	0.175 ± 0.036	0.178 ± 0.050	0.185 ± 0.042	0.175 ± 0.040	0.190 ± 0.054