BEYOND EXCHANGEABILITY: DISTRIBUTION-SHIFT-AWARE INTEGRATION OF EXTERNAL CONTROL DATA IN RANDOMIZED TRIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized controlled trials (RCTs) are the gold standard for evaluating causal effects but are often costly and difficult to scale; consequently, they are frequently augmented with auxiliary external controls in many applications. Prior approaches to borrowing such data typically rely on exchangeability, under which the external controls are readily usable for inference in the trial population. In practice, however, differences in eligibility criteria, standard of care, and data collection procedures may induce distribution shifts between the RCT and the external controls, making exchangeability implausible. In this paper, we propose a general framework for incorporating external controls that explicitly accounts for distribution shifts through an augmented estimating function, allowing information about the shifts to be leveraged for efficiency gains even when exchangeability does not hold. To guard against potential bias and efficiency loss due to misspecification of the distribution shift, we further develop an adaptive shrinkage estimator that preserves consistency while guaranteeing efficiency dominance over the trial-only benchmark. Synthetic experiments and a real data application demonstrate the practical advantages of the proposed approaches.

Key Words: Adaptive shrinkage; causal inference; distribution shift; external control data; exchangeability.

1 INTRODUCTION

Randomized controlled trials (RCTs) are the gold standard for evaluating treatment effects because randomization removes confounding bias by design (Imbens & Rubin, 2015). In practice, however, RCTs often produce estimates that are not sufficiently precise, due to optimistic assumptions about event rates, follow-up, and accrual, as well as limited power for subgroup analyses. This has motivated growing interest in augmenting RCTs with auxiliary external data, such as historical trials, hospital registries, electronic health records, and claims (Pocock, 1976; Colnet et al., 2024). In many studies of experimental interventions, these auxiliary data contain only untreated subjects and are therefore used as external controls (ECs). A straightforward way to incorporate ECs is to assume *exchangeability*: conditional on observed covariates, the untreated potential outcome distribution, or some more essential characteristics such as its conditional mean, is comparable between the trial and EC populations. Under this assumption, ECs are readily usable for trial inference, and many methods have been developed to borrow such external information (Dahabreh et al., 2019; Li et al., 2023; Valancius et al., 2024).

In practice, however, exchangeability is often tenuous. Differences in eligibility criteria, standards of care, calendar time, recruitment mechanisms, and measurement protocols can induce distribution shifts that make ECs systematically different from trial controls. Recent work has therefore developed guarded approaches aiming to reduce the risk of biased borrowing by selecting compatible external subjects or by downweighting, discounting, or shrinking external information (Rosenman et al., 2023; Yang et al., 2023; Gao et al., 2025). These strategies improve robustness but still rely on the compatibility of the available EC data, and may discard or attenuate useful information. Moreover, existing methods typically borrow EC information only for the control arm, overlooking the

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.

possibility that external covariate information, when properly calibrated to the trial population, can improve inference for both the treatment and control arms.

To tackle these challenges, we develop a general framework for integrating ECs into trial inference that explicitly accounts for distribution shifts, thereby allowing available information about discrepancies between the trial and EC populations to be incorporated for potential efficiency gains without requiring exchangeability. Specifically, we consider both *covariate* shift, which captures differences in covariate distributions, and *concept* shift, which captures differences in conditional outcome distributions. We derive calibration equations that explicitly involve these shifts and use them to augment trial-only efficient influence functions, yielding EC-augmented estimating equations for both the treatment and control arms that can draw on shift information from sources such as design knowledge, previously linked studies, or validation data.

Importantly, EC data alone do not necessarily improve efficiency. When no additional information about the distribution shifts is available, estimating unrestricted shift models solely from the analysis data need not yield efficiency gains. Moreover, misspecification of the shift models and variability in their estimation may introduce bias or render the augmented estimator less efficient than trial-only inference. To guard against these risks, we further develop a data-adaptive shrinkage estimator that combines the EC-augmented estimator with a trial-only benchmark. The shrinkage estimator preserves trial-based consistency and guarantees asymptotic efficiency dominance over the trial-only benchmark, while retaining the potential efficiency gains when the available shift information is sufficiently informative and reliable.

To summarize, we make the following novel contributions in this work:

- We introduce a new framework for integrating EC data into RCT analysis without assuming exchangeability, by explicitly exploiting distribution shifts between the trial and EC populations.
- We propose calibration-based augmented estimating functions from trial-only efficient influence functions, exploiting both covariate and concept shifts so that EC data can improve inference for both treated and control arms.
- We further propose a data-adaptive shrinkage estimator that remains consistent under shift-model misspecification while guaranteeing efficiency dominance over the trial-only benchmark.

We have conducted extensive synthetic experiments and a real-data application to demonstrate the finite-sample advantages of the proposed approaches.

2 PRELIMINARY

Let $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ denote the pre-treatment covariates and $Y \in \mathcal{Y} \subset \mathbb{R}$ the outcome of interest. Let $T \in \{0, 1\}$ denote the treatment assignment, with $T = 1$ for treatment and $T = 0$ for control. We adopt the potential outcome framework $Y = TY_1 + (1 - T)Y_0$, where Y_1 and Y_0 denote the potential outcomes under treatment and control, respectively.

Suppose we have access to data from an RCT and an external control cohort that contains only untreated subjects. Let R indicate the data source, where $R = 1$ for RCT and $R = 0$ for EC. The observed data $\mathcal{O}$ consist of N independent and identically distributed (i.i.d.) samples of $(R, T, \mathbf{X}, Y)$ from the combined population, partitioned into the RCT sample, $\mathcal{O}_1 = \{(r_i = 1, t_i, \mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$, and the EC sample, $\mathcal{O}_2 = \{(r_i = 0, t_i = 0, \mathbf{x}_i, y_i),\ i = n + 1, \ldots, N\}$. Assume that $n/N \to \kappa \in (0, 1)$ as $n, N \to \infty$.

The causal estimand of interest is the average treatment effect (ATE) in the trial population, defined as $\tau = \tau_1 - \tau_0$, where $\tau_t = \mathbb{E}(Y_t \mid R = 1)$ for $t \in \{0, 1\}$. The following standard assumptions, which formalize the RCT design, ensure identification of the causal parameters (Rosenbaum & Rubin, 1983):

Assumption 1. *(i)* $T \perp\!\!\!\perp (\mathbf{X}, Y_0, Y_1) \mid R = 1$*; (ii)* $0 < \pi = \mathrm{pr}(T = 1 \mid R = 1) < 1$.

Under Assumption 1, τ_1 and τ_0 are identified from the trial data $\mathcal{O}_1$. A natural pair of gold-standard estimators is given by the treated and control sample means $\widetilde{\tau}_1^{\mathrm{gs}} = \sum_{i=1}^n t_i y_i / \sum_{i=1}^n t_i$ and $\widetilde{\tau}_0^{\mathrm{gs}} = \sum_{i=1}^n (1 - t_i) y_i / \sum_{i=1}^n (1 - t_i)$. To characterize the most efficient estimator based on the trial data,

define $\mu_t(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{x}, t, R = 1)$ for $t \in \{0, 1\}$. The efficient influence functions (EIF) (Robins et al., 1994) for estimating τ_1 and τ_0, are, respectively,

$$\widetilde{\psi}_1 = \frac{RT}{\kappa\pi}\{Y - \mu_1(\mathbf{X})\} + \frac{R}{\kappa}\{\mu_1(\mathbf{X}) - \tau_1\}, \tag{1}$$

$$\text{and } \widetilde{\psi}_0 = \frac{R(1-T)}{\kappa(1-\pi)}\{Y - \mu_0(\mathbf{X})\} + \frac{R}{\kappa}\{\mu_0(\mathbf{X}) - \tau_0\}. \tag{2}$$

Denote $\widetilde{\psi} = \widetilde{\psi}_1 - \widetilde{\psi}_0$. Thus, the semiparametric efficiency bound for estimating τ based on the trial data solely is $\widetilde{B} = \mathbb{E}(\widetilde{\psi}^2)$.

3 Exploitation of Distribution Shifts between RCT and EC

A straightforward way to integrate EC data in RCT is to assume certain similarity between the two sources, which was termed *exchangeability* in the literature. For instance, one may assume the distribution exchangeability, namely,

$$p(y_0 \mid \mathbf{x}) = q(y_0 \mid \mathbf{x}),$$

where $p(\cdot)$ and $q(\cdot)$ denote the distribution in the RCT and the EC populations, respectively, throughout the paper. This certainly implies that the conditional mean of Y_0 is shared across the trial and the external controls, namely, $\mathbb{E}(Y_0 \mid \mathbf{x}, R = 1) = \mathbb{E}(Y_0 \mid \mathbf{x}, R = 0)$, i.e., the mean exchangeablity. Accordingly, the external controls can be readily used to inform untreated outcomes in the trial population.

However, this type of strong similarity assumption may often fail in applications because it is the rule rather than the exception that the distributions between RCT and EC differ. To move beyond exchangeability, one needs to explicitly model the discrepancy between the two populations. Let $a(\mathbf{x}) := q(\mathbf{x})/p(\mathbf{x})$ represent the *covariate* shift between the two data sources. Further, let $b(\mathbf{x}, y_0) := q(y_0 \mid \mathbf{x})/p(y_0 \mid \mathbf{x})$ represent the *concept* shift in the sense of the conditional distribution of the untreated potential outcome. To ease the presentation, we define

$$\begin{aligned} k(\mathbf{x}) &:= \text{pr}(R = 1 \mid \mathbf{x}) = \frac{\kappa}{\kappa + (1-\kappa)a(\mathbf{x})}, \\ \text{and } \rho(\mathbf{x}, y_0) &:= \text{pr}(R = 1 \mid \mathbf{x}, y_0, T = 0) = \frac{\kappa(1-\pi)}{\kappa(1-\pi) + (1-\kappa)a(\mathbf{x})b(\mathbf{x}, y_0)}. \end{aligned} \tag{3}$$

The following lemma shows how $k(\mathbf{x})$ and $\rho(\mathbf{x}, y_0)$ can be used to balance distributions across the trial, external controls, and the combined population.

Lemma 1. *Under Assumption 1, for any integrable functions $g_1(\mathbf{x})$ and $g_2(\mathbf{x}, y_0)$, it holds that*

$$\mathbb{E}\{g_1(\mathbf{X}) \mid R = 1\} = \mathbb{E}\left\{\frac{k(\mathbf{X})}{\kappa} g_1(\mathbf{X})\right\}, \text{ and}$$

$$\mathbb{E}\{g_2(\mathbf{X}, Y_0) \mid R = 1\} = \mathbb{E}\left\{\frac{1-T}{\kappa(1-\pi)}\rho(\mathbf{X}, Y_0) g_2(\mathbf{X}, Y_0)\right\}.$$

Lemma 1 motivates the idea of augmenting the trial-only estimating functions, $\widetilde{\psi}_1$ and $\widetilde{\psi}_0$, to incorporate external controls by exploiting and modeling the distributional shifts. For τ_1, the EC sample contributes only covariate information through $\mathbf{X}$, yielding the following EC-augmented estimating function:

$$\psi_1 = \frac{RT}{\kappa\pi}\{Y - \mu_1(\mathbf{X})\} + \frac{k(\mathbf{X})}{\kappa}\{\mu_1(\mathbf{X}) - \tau_1\}. \tag{4}$$

For τ_0, the EC sample provides information on both covariates and untreated outcomes, leading to the following augmented estimating function:

$$\psi_0 = \frac{1-T}{\kappa(1-\pi)}\rho(\mathbf{X}, Y)\{Y - \mu_0(\mathbf{X})\} + \frac{k(\mathbf{X})}{\kappa}\{\mu_0(\mathbf{X}) - \tau_0\}. \tag{5}$$

Let $\Delta(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$ and let $\psi = \psi_1 - \psi_0$ be the estimating function of τ. The following proposition establishes the validity of ψ_t and characterizes its potential efficiency improvement over the trial-only estimating function $\widetilde{\psi}_t$.

Proposition 1. *Under Assumption 1, we have* $\mathbb{E}(\psi_t) = \mathbb{E}(\psi) = 0$ *for* $t \in \{0, 1\}$. *Let*

$$\mathcal{V}_\rho = \frac{1}{\kappa^2(1-\pi)^2}\mathbb{E}\left[(1-T)\{Y-\mu_0(\mathbf{X})\}^2\rho(\mathbf{X},Y)\left\{1-\rho(\mathbf{X},Y)\right\}\right] \geq 0,$$

denote the contribution of the concept shift. Then

$$\mathbb{E}(\widetilde{\psi}_1^2) - \mathbb{E}(\psi_1^2) = \frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_1(\mathbf{X})-\tau_1\}^2 k(\mathbf{X})\left\{1-k(\mathbf{X})\right\}\right] \geq 0,$$

$$\mathbb{E}(\widetilde{\psi}_0^2) - \mathbb{E}(\psi_0^2) = \frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_0(\mathbf{X})-\tau_0\}^2 k(\mathbf{X})\left\{1-k(\mathbf{X})\right\}\right] + \mathcal{V}_\rho \geq 0,$$

$$\textit{and}\quad \mathbb{E}(\widetilde{\psi}^2) - \mathbb{E}(\psi^2) = \frac{1}{\kappa^2}\mathbb{E}\left[\{\Delta(\mathbf{X})-\tau\}^2 k(\mathbf{X})\left\{1-k(\mathbf{X})\right\}\right] + \mathcal{V}_\rho \geq 0.$$

Each difference is strictly positive as long as there exists a set of $\mathbf{x}$ *with nonzero measure on which the densities overlap, i.e.,* $p(\mathbf{x})q(\mathbf{x}) > 0$, *and the relevant mean function,* $\mu_t(\mathbf{x})$ *or* $\Delta(\mathbf{x})$, *is non-constant. An analogous condition applies to* $\mathcal{V}_\rho$.

4 Proposed Augmented Estimator Incorporating EC Data

The augmented estimating functions (4)-(5) explicitly account for distribution shifts through the weights k and ρ, thereby incorporating external controls without assuming exchangeability. In this section, we focus on the corresponding augmented estimators under parametric working models $k(\mathbf{x};\boldsymbol{\alpha})$, $\rho(\mathbf{x}, y_0;\boldsymbol{\beta})$, and $\mu_t(\mathbf{x};\boldsymbol{\gamma}_t)$ for $k(\mathbf{x})$, $\rho(\mathbf{x}, y_0)$, and $\mu_t(\mathbf{x})$, respectively, and study their asymptotic properties. This also provides a concrete illustration of how estimating the distribution shifts affects the resulting treatment effect estimators.

Let $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}_0, \widehat{\boldsymbol{\gamma}}_1)$ denote $N^{1/2}$-consistent estimators of the nuisance parameters, with a probability limit $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}_0^*, \boldsymbol{\gamma}_1^*)$. Then, an estimator of τ_1 that solves (4) is

$$\widehat{\tau}_1 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{r_t t_i}{\widehat{\pi}}\{y_i - \mu_1(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_1)\} + k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}})\mu_1(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_1)\right] \Big/ \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}}),$$

and an estimator of τ_0 that solves (5) is

$$\widehat{\tau}_0 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1-t_i}{1-\widehat{\pi}}\rho(y_i,\mathbf{x}_i;\widehat{\boldsymbol{\beta}})\{y_i - \mu_0(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_0)\} + k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}})\mu_0(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_0)\right] \Big/ \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\pi} = \#\{T=1\}/n$. An estimator of the treatment effect τ is then given by $\widehat{\tau} = \widehat{\tau}_1 - \widehat{\tau}_0$.

Compared with the trial-based estimators solving (1)–(2), the augmented estimator $\widehat{\tau}_1$ entails an extra working model $k(\mathbf{x};\boldsymbol{\alpha})$, while $\widehat{\tau}_0$ entails extra working models $k(\mathbf{x};\boldsymbol{\alpha})$ and $\rho(\mathbf{x}, y_0;\boldsymbol{\beta})$ to incorporate external controls. The following Theorem 1 establishes the consistency and asymptotic normality of the augmented estimators.

Theorem 1. *Under Assumption 1 and regularity conditions in Theorems 2.6 and 3.4 of Newey & McFadden (1994), the estimator* $\widehat{\tau}_1$ *is consistent and asymptotically normal (CAN) if* $k(\mathbf{x};\boldsymbol{\alpha}^*) = k(\mathbf{x})$. *Moreover, the estimator* $\widehat{\tau}_0$ *is CAN if* $k(\mathbf{x};\boldsymbol{\alpha}^*) = k(\mathbf{x})$ *and* $\rho(y_0,\mathbf{x};\boldsymbol{\beta}^*) = \rho(y_0,\mathbf{x})$.

Theorem 1 shows that consistency of $\widehat{\tau}_1$ relies on correct model specification of $k(\mathbf{x})$, while remaining robust to potential misspecification of $\mu_1(\mathbf{x})$. Likewise, the consistency of $\widehat{\tau}_0$ relies on correct model specification of $k(\mathbf{x})$ and $\rho(y_0,\mathbf{x})$, while remaining robust to potential misspecification of $\mu_0(\mathbf{x})$. As shown in (3), correct specification of $k(\mathbf{x})$ and $\rho(y_0,\mathbf{x})$ is equivalent to a "precise" characterization of the distribution shifts $a(\mathbf{x})$ and $b(\mathbf{x}, y_0)$. Because the augmented estimating functions incorporate external data through the modeled distributional shifts, consistency naturally hinges on correct specification of those shifts. At the same time, the augmented estimator $\widehat{\tau}_t$ retains robustness to possible misspecification of the outcome models $\mu_t(\mathbf{x})$, mirroring the robustness property enjoyed by trial-only doubly robust estimators from (1)-(2). The following Theorem 2 further presents the asymptotic representation when all working models are correctly specified.

Theorem 2. *Suppose* $k(\mathbf{x};\boldsymbol{\alpha}^*) = k(\mathbf{x})$, $\rho(y_0,\mathbf{x};\boldsymbol{\beta}^*) = \rho(y_0,\mathbf{x})$, $\mu_t(\mathbf{x};\boldsymbol{\gamma}_t^*) = \mu_t(\mathbf{x})$ *for* $t \in \{0,1\}$, *and assumptions in Theorem 1 hold. Define*

$$\Gamma_t = \mathbb{E}\left[\frac{\{\mu_t(\mathbf{X}) - \tau_t\}}{\kappa}\frac{\partial k(\mathbf{X};\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right], \text{ and } \Gamma_\rho = \mathbb{E}\left[\frac{1-T}{\kappa(1-\pi)}\{Y_0 - \mu_0(\mathbf{X})\}\frac{\partial \rho(\mathbf{X},Y_0;\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right],$$

for $t \in \{0,1\}$. *Then, we have*
(i) $\sqrt{N}(\widehat{\tau}_1 - \tau_1) = N^{-1/2}\sum_{i=1}^N \psi_{1,i} + \mathcal{R}_1 + o_p(1)$, *where* $\mathcal{R}_1 = \Gamma_1 \cdot \sqrt{N}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$;
(ii) $\sqrt{N}(\widehat{\tau}_0 - \tau_0) = N^{-1/2}\sum_{i=1}^N \psi_{0,i} + \mathcal{R}_2 + \mathcal{R}_3 + o_p(1)$, *where* $\mathcal{R}_2 = \Gamma_0 \cdot \sqrt{N}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$ *and* $\mathcal{R}_3 = \Gamma_\rho \cdot \sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$.

Theorem 2 shows that, under correct model specification, estimating $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_0)$ does not jeopardize the asymptotic behavior of $(\widehat{\tau}_1, \widehat{\tau}_0)$, though estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ does. Indeed, this well aligns with the intuition that some prior knowledge about how distribution shifts helps gaining the efficiency. The more one knows, the more efficient the estimators are. An ideal scenario arises when the distribution shifts are completely known from the study design or reliable external knowledge, so that the weights k and ρ need not be estimated. In this case, the remainder terms $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ in Theorem 2 vanish, yielding $\sqrt{N}(\widehat{\tau}_1 - \tau_1) \xrightarrow{d} \mathcal{N}\{0, \mathbb{E}(\psi_1^2)\}$, and $\sqrt{N}(\widehat{\tau}_0 - \tau_0) \xrightarrow{d} \mathcal{N}\{0, \mathbb{E}(\psi_0^2)\}$. Thus, by Proposition 1, the augmented estimators guarantee an efficiency gain over trial-only inference. When the shifts are only partially known, so that the working models for k and ρ contain fewer free parameters, efficiency gains are also plausible; see Section 6 for a numerical illustration. Efficiency gains of this type have been explored in the missing data literature. For example, Kim & Yu (2011) considered an exponential tilting model for outcome-dependent selection; that is, $q(y_0 \mid \mathbf{x}) \propto \exp(\theta y_0)\, p(y_0 \mid \mathbf{x})$, with θ known a priori.

In practice, it is also of interest to consider a validation setting in which an auxiliary validation dataset is available for estimating the shifts. Suppose an independent validation sample of size m is available for estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, with $m/N \to \xi \in (0,\infty)$, and $\sqrt{m}\{(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathrm{T}}\}^{\mathrm{T}} = m^{-1/2}\sum_{j=1}^m \boldsymbol{\phi}_j + o_p(1)$, where $\boldsymbol{\phi}_j = (\boldsymbol{\phi}_{\alpha,j}^{\mathrm{T}}, \boldsymbol{\phi}_{\beta,j}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbb{E}(\boldsymbol{\phi}) = \mathbf{0}$. Then, Theorem 2 implies that $\sqrt{N}(\widehat{\tau}_1 - \tau_1) \xrightarrow{d} \mathcal{N}(0, \sigma_{1,\mathrm{val}}^2)$ and $\sqrt{N}(\widehat{\tau}_0 - \tau_0) \xrightarrow{d} \mathcal{N}(0, \sigma_{0,\mathrm{val}}^2)$, where

$$\sigma_{1,\mathrm{val}}^2 = \mathbb{E}(\psi_1^2) + \xi^{-1}\Gamma_1 \mathrm{var}(\boldsymbol{\phi}_\alpha)\Gamma_1^{\mathrm{T}}, \text{ and } \sigma_{0,\mathrm{val}}^2 = \mathbb{E}(\psi_0^2) + \xi^{-1}(\Gamma_0, \Gamma_\rho)\mathrm{var}(\boldsymbol{\phi})(\Gamma_0, \Gamma_\rho)^{\mathrm{T}}.$$

Therefore, the additional variance induced by estimating the distribution shift from validation data is of order ξ^{-1}. As the validation sample becomes large relative to the primary sample, the extra terms vanish, and the asymptotic variances converge to $\mathbb{E}(\psi_1^2)$ and $\mathbb{E}(\psi_0^2)$, which are exactly the variances obtained when the distribution shift is completely known.

5 Proposed Shrinkage Estimator with Guaranteed Consistency and Efficiency Dominance

In applications, the EC-augmented estimators proposed in Section 4 may suffer from inconsistency or loss of efficiency because of the misspecification of the shift model or the additional variability from sufficiently estimating it. To address this issue, we propose a dependable, data-adaptive shrinkage estimator that fuses $\widehat{\tau}_t$ (potentially more efficient but misspecification-sensitive) with the trial-only estimator $\widetilde{\tau}_t$ (guaranteed consistency but potentially less efficient), to guarantee consistency while also retaining potential efficiency gains, even under model misspecification.

We consider estimators of the form

$$\widehat{\tau}_t(\lambda_t) = \widetilde{\tau}_t + \lambda_t(\widehat{\tau}_t - \widetilde{\tau}_t), \quad \lambda_t \in \mathbb{R},$$

where λ_t is the shrinkage weight. Note that $\widehat{\tau}_t(0) = \widetilde{\tau}_t$ and $\widehat{\tau}_t(1) = \widehat{\tau}_t$. The key insight is as follows. If $\widehat{\tau}_t$ is inconsistent (i.e., $\widehat{\tau}_t \not\to \tau_t$), we expect λ_t to shrink toward zero, so that the combined estimator reduces to the consistent baseline $\widetilde{\tau}_t$. When $\widehat{\tau}_t$ is consistent for τ_t, $\widehat{\tau}_t(\lambda_t)$ retains consistency for any choice of λ_t. In this case, we expect λ_t to converge to an optimal λ_t^* that minimizes the asymptotic variance of $\widehat{\tau}_t(\lambda_t)$, ensuring that it is at least as efficient as either separate estimator.

We first calculate λ_t^*. Noting that $\mathrm{var}\{\widehat{\tau}_t(\lambda_t)\} = \mathrm{var}(\widetilde{\tau}_t) + 2\lambda_t \mathrm{cov}(\widehat{\tau}_t - \widetilde{\tau}_t, \widetilde{\tau}_t) + \lambda_t^2 \mathrm{var}(\widehat{\tau}_t - \widetilde{\tau}_t)$, we have

$$\lambda_t^* = \mathrm{argmin}_{\lambda_t}\mathrm{var}\{\widehat{\tau}_t(\lambda_t)\} = -\frac{\mathrm{cov}(\widehat{\tau}_t - \widetilde{\tau}_t, \widetilde{\tau}_t)}{\mathrm{var}(\widehat{\tau}_t - \widetilde{\tau}_t)}.$$

It is straightforward to verify that

$$\sigma_t^2 := \mathrm{var}\{\widehat{\tau}_t(\lambda_t^*)\} = \frac{\mathrm{var}(\widetilde{\tau}_t)\mathrm{var}(\widehat{\tau}_t) - \mathrm{cov}(\widehat{\tau}_t, \widetilde{\tau}_t)^2}{\mathrm{var}(\widehat{\tau}_t - \widetilde{\tau}_t)} \leq \min\{\mathrm{var}(\widetilde{\tau}_t), \mathrm{var}(\widehat{\tau}_t)\}. \tag{6}$$

To facilitate adaptive shrinkage driven by the consistency of $\widehat{\tau}_t$, we define, with $V_{t,N} := \mathrm{var}(\widehat{\tau}_t - \widetilde{\tau}_t)$, and a given $\gamma \in (0, 3/2)$,

$$\delta_{t,N} = \frac{V_{t,N}^2}{V_{t,N}^2 + (\widehat{\tau}_t - \widetilde{\tau}_t)^4 / N^\gamma} = \left\{ 1 + N^{-\gamma} \left(\frac{\widehat{\tau}_t - \widetilde{\tau}_t}{\sqrt{V_{t,N}}} \right)^4 \right\}^{-1}. \tag{7}$$

Noting that $V_{t,N} = O(N^{-1})$, if $\widehat{\tau}_t$ is consistent, then $\widehat{\tau}_t - \widetilde{\tau}_t = O_p(N^{-1/2})$ and hence $\delta_{t,N} = 1 + O_p(N^{-\gamma})$. Otherwise, if $\widehat{\tau}_t$ is inconsistent, i.e., $\widehat{\tau}_t - \widetilde{\tau}_t = c_t^* + O_p(N^{-1/2})$ for some $c_t^* \neq 0$, we have $\delta_{t,N} = O_p(N^{\gamma-2}) = o_p(N^{-1/2})$. Denote $\lambda_{t,N} = \delta_{t,N}\lambda_t^*$, then we have

$$\lambda_{t,N} = \begin{cases} \lambda_t^* + O_p(N^{-\gamma}), & \text{if } \widehat{\tau}_t \xrightarrow{p} \tau_t, \\ o_p(N^{-1/2}), & \text{otherwise.} \end{cases}$$

The tuning parameter γ balances protection against bias with the pursuit of efficiency. Specifically, a smaller γ shrinks $\delta_{t,N}$ to zero faster and extends protection to biases as small as $N^{\gamma/4-1/2}$, at the cost of returning to one more slowly, at rate $N^{-\gamma}$, when $\widehat{\tau}_t$ is consistent. We use $\gamma = 1/3$ in our numerical experiments, which guards against biases of order exceeding $N^{-5/12}$. It is also of interest to consider *local* model misspecification, in which the bias of $\widehat{\tau}_t$ is of the same order $N^{-1/2}$ as its own standard error. In the interest of space, we defer the detailed discussion and an alternative combination strategy to Appendix A under local misspecification.

Let $\widehat{\lambda}_{t,N}$ denote the empirical analog of $\lambda_{t,N}$ as presented in Appendix H. We define the resulting shrinkage estimator as $\widehat{\tau}_t^{(s)} = \widehat{\tau}_t(\widehat{\lambda}_{t,N})$. The following Theorem 3 establishes the asymptotic normality of $\widehat{\tau}_t^{(s)}$ and the validity of inference under arbitrary misspecification.

Theorem 3. *Suppose that, for $t \in \{0, 1\}$, (i) $\widetilde{\tau}_t$ is a $\sqrt{N}$-consistent estimator of τ_t; (ii) $(\widetilde{\tau}_t, \widehat{\tau}_t)$ are jointly asymptotically normal about their probability limits (τ_t, τ_t^*), that is, $\sqrt{N}\{(\widetilde{\tau}_t - \tau_t), (\widehat{\tau}_t - \tau_t^*)\}^{\mathrm{T}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$ with $\boldsymbol{\Sigma}_t$ positive definite; and (iii) a consistent estimator $\widehat{\boldsymbol{\Sigma}}_t \xrightarrow{p} \boldsymbol{\Sigma}_t$ is available, so that $\widehat{V}_{t,N}/V_{t,N} \xrightarrow{p} 1$. Write $\boldsymbol{c}(\lambda) = (1-\lambda, \lambda)^{\mathrm{T}}$ and let $\lambda_t^\dagger = \lambda_t^* \mathbb{1}(\tau_t^* = \tau_t)$ be the probability limit of $\widehat{\lambda}_{t,N}$. Then, we have $\sqrt{N}\{\widehat{\tau}_t^{(s)} - \tau_t\} \xrightarrow{d} \mathcal{N}(0, v_t)$, with $v_t = \boldsymbol{c}(\lambda_t^\dagger)^{\mathrm{T}} \boldsymbol{\Sigma}_t \boldsymbol{c}(\lambda_t^\dagger)$. In addition, we have $\widehat{v}_t = \boldsymbol{c}(\widehat{\lambda}_{t,N})^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_t \boldsymbol{c}(\widehat{\lambda}_{t,N}) \xrightarrow{p} v_t$.*

Theorem 3 describes the adaptive behavior of $\widehat{\tau}_t^{(s)}$. If $\widehat{\tau}_t$ is inconsistent due to model misspecification, then $\boldsymbol{c}(\lambda_t^\dagger) = (1, 0)$ and $\widehat{\tau}_t^{(s)}$ degenerates to the trial-only estimator $\widetilde{\tau}_t$. If $\widehat{\tau}_t$ is consistent for τ_t, then $\boldsymbol{c}(\lambda_t^\dagger) = (1 - \lambda_t^*, \lambda_t^*)$, and $\widehat{\tau}_t^{(s)}$ aggregates $\widehat{\tau}_t$ and $\widetilde{\tau}_t$ to incorporate information from the external control data to improve efficiency. Therefore, $\widehat{\tau}_t^{(s)}$ adaptively interpolates between robustness and efficiency, and remains consistent regardless of whether the working models are correctly specified. In practice, one may use the simple gold-standard estimators, $\widetilde{\tau}_1^{\mathrm{gs}}$ and $\widetilde{\tau}_0^{\mathrm{gs}}$, as baselines, which require no working models. Alternatively, one may use the trial-only doubly robust estimators obtained by solving estimating equations (1)-(2), which additionally involve working models $\mu_t(\mathbf{x}; \boldsymbol{\gamma}_t^*)$. It is well known that consistency of the doubly robust estimator does not rely on correct specification of $\mu_t(\mathbf{x}; \boldsymbol{\gamma}_t^*)$, although efficiency does. The following corollary summarizes the shrinkage estimator's guaranteed consistency and its potential efficiency gains.

Corollary 1. *Suppose assumptions in Theorem 1 and Theorem 3 hold. Then the shrinkage estimator $\widehat{\tau}_t^{(s)}$ is consistent and asymptotically normal, with asymptotic variance no larger than that of the trial-based estimator $\widetilde{\tau}_t$, for any choice of working models $k(\mathbf{x}; \boldsymbol{\alpha}^*)$, $\rho(y_0, \mathbf{x}; \boldsymbol{\beta}^*)$ and $\mu_t(\mathbf{x}; \boldsymbol{\gamma}_t^*)$.*

Remark 1. We construct shrinkage estimators for the treatment and control arms, respectively. The motivation is that consistency of $\widehat{\tau}_1$ only requires correct specification of $k(\mathbf{x})$ and is therefore more plausible than consistency of $\widehat{\tau}_0$, which requires correct specification of both $k(\mathbf{x})$ and $\rho(\mathbf{x}, y_0)$. In case that $k(\mathbf{x})$ is correctly specified but $\rho(\mathbf{x}, y_0)$ is not, the separate strategy can still leverage $\widehat{\tau}_1$ to

borrow information from the external data, while shrinking the control mean toward the trial-only estimator $\widetilde{\tau}_0$ to preserve consistency. It is also straightforward to adapt the construction when the estimand of interest is the treatment effect τ. In this case, one can form a single shrinkage estimator that combines $\widetilde{\tau}$ and $\widehat{\tau}$, and define an adaptive shrinkage weight via an analogous procedure. The resulting estimator remains consistent and is guaranteed to be at least as efficient as the baseline estimator $\widetilde{\tau}$.

6 Numerical Studies

In this section, we conduct three synthetic experiments and a real-data analysis to assess the finite-sample performance of the proposed estimators and compare them with the main existing approaches to external-control borrowing. The first simulation illustrates how the methods behave under increasing departures from exchangeability, the second quantifies the efficiency gains attainable from different types of information about the distributional shift, and the third assesses the robustness of the shrinkage estimator to model misspecification. Supplemental results are collected in Appendix B.

For the trial, we generate $X \sim \mathcal{N}(0,1)$ and $Y_t \mid x \sim \mathcal{N}\{\mu_t(x), \sigma^2(x)\}$ with $\mu_0(x) = 1+x+0.5x^2$, $\mu_1(x) = 2 + 1.8x + 0.5x^2$, and $\sigma^2(x) = 0.5 + 0.5x^2$. The treatment assignment probability is $\pi = 0.5$. For the external controls, we generate $X \sim \mathcal{N}(\mu_E, \sigma_E^2)$ with $(\mu_E, \sigma_E) = (0.5, 1.2)$, which controls the magnitude of the covariate shift, and we let the distribution of the untreated outcome be the exponential tilt of its trial counterpart, $q(y_0 \mid x) \propto \exp(\theta y_0)\, p(y_0 \mid x)$, so that $Y \mid x \sim \mathcal{N}\{\mu_0(x) + \theta\sigma^2(x),\ \sigma^2(x)\}$. The tilting parameter θ controls the magnitude of the concept shift, with $\theta = 0.5$ as the default and $\theta = 0$ corresponding to the scenario in which exchangeability holds. Under this data-generating mechanism, the shift functions take the logistic forms $k(x;\boldsymbol{\alpha}) = \mathrm{expit}(\alpha_0 + \alpha_1 x + \alpha_2 x^2)$ and $\rho(x, y_0; \boldsymbol{\beta}) = \mathrm{expit}(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 y_0)$ exactly, with $\beta_3 = -\theta$. Throughout, we use 2000 Monte Carlo replicates unless stated otherwise.

We compare the trial-only estimator $\widetilde{\tau}_t$ constructed from (1)-(2), the EC-augmented estimator $\widehat{\tau}_t$ and the shrinkage estimator $\widehat{\tau}_t^{(s)}$ with four alternatives: the *naive pooling* estimator, which treats the external controls as trial controls without adjustment; the *exchangeability-efficient* estimator (Li et al., 2023), which is semiparametrically efficient under conditional mean exchangeability; the *adaptive selective borrowing* approach (Gao et al., 2025), which uses an adaptive lasso to borrow only the subset of external controls that are compatible with exchangeability; and the *randomization-aware* approach (Karlsson et al., 2025), which uses the external controls solely to construct the outcome regression while relying on the known trial randomization probability for identification.

Experiment I. In this experiment, we vary θ to control the degree of violation of the exchangeability and treat its value as known a priori as in Kim & Yu (2011). The proposed method can explicitly incorporate this information to adjust the external outcomes, whereas the competing methods cannot. Figure 1 reports the bias, coverage, and relative variance of $\widehat{\tau}$ for a trial of size $n = 500$ within a total sample of $N = 1500$. Tables and additional results are deferred to Appendix B.

Naive pooling is biased throughout. When exchangeability holds, i.e. $\theta = 0$, all the remaining methods are unbiased with satisfactory coverage, and the methods of Li et al. (2023), Gao et al. (2025), and our method exploiting the knowledge of $\theta = 0$ all exhibit clear efficiency gains. As the departure from exchangeability grows, the exchangeability-efficient estimator becomes increasingly biased, as expected, and its coverage deteriorates. Adaptive selective borrowing shrinks the retained external sample to nearly an empty set, from 97.5% at $\theta = 0$ to 0.1% once $\theta \geq 0.75$; it therefore remains unbiased but achieves no efficiency gain, while exhibiting some numerical instability for intermediate values of θ. The randomization-aware estimator is unbiased across all settings, but shows no appreciable efficiency gain. In contrast, the proposed estimator uses θ directly through the working model $\rho(x, y_0; \boldsymbol{\beta})$, without any model for the conditional distribution of Y_0 beyond its mean, and remains unbiased with nominal coverage and a variance between 0.65 and 0.92 times that of the trial-only estimator across the whole range of θ.

Experiment II. This experiment investigates numerically what information the proposed estimator is able to exploit. We consider three factors, namely the availability of an external validation sample, knowledge of the shape of the shift, and the size of the external sample relative to that of the trial.

Figure 2(a) displays the variance of the proposed augmented estimator relative to that of the trial-only estimator, plotted against the size of the validation sample for three values of the trial fraction

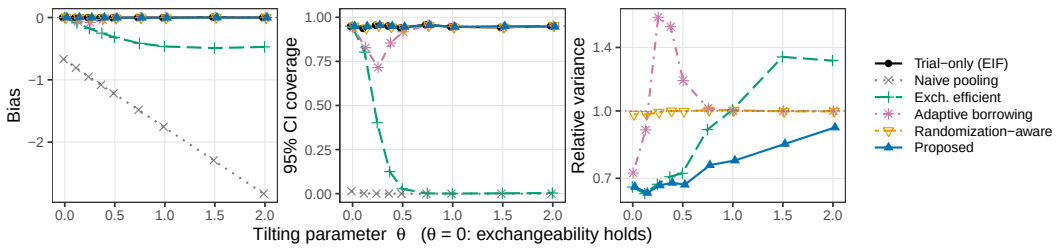

Figure 1: Estimation of τ as the tilting parameter θ grows, with $n = 500$ and $N = 1500$. The proposed estimator treats θ as known. Relative variance is on a log scale against the trial-only estimator, with smaller values indicating greater efficiency.

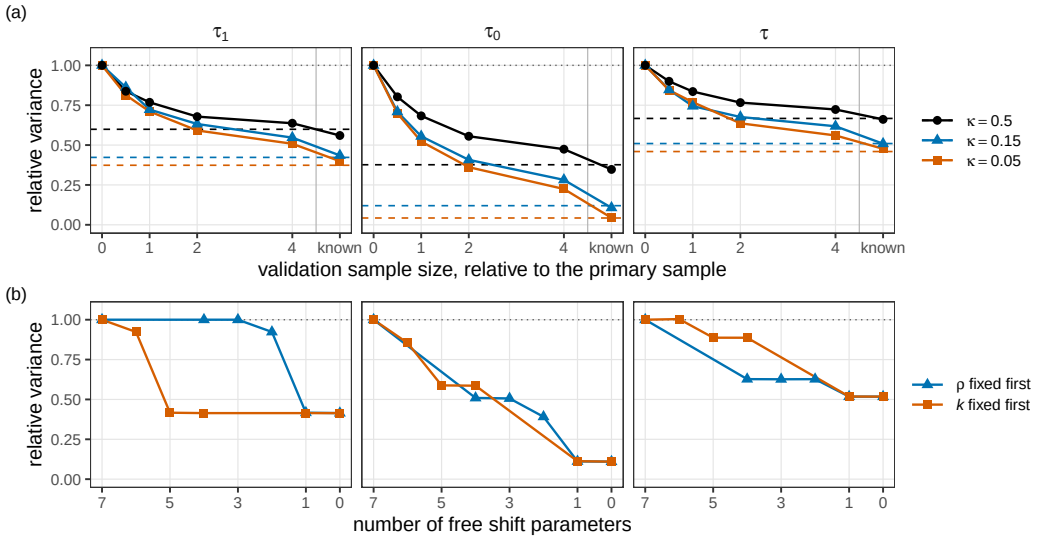

Figure 2: Relative variance of the proposed augmented estimator as information about the shift increases, through (a) a validation sample of increasing size and (b) components of k and ρ fixed at their true values. Dashed lines in (a) indicate the known-shift limits given in Proposition 1.

$\kappa = n/N$. The efficiency gain increases steadily with the validation sample size and approaches the known-shift benchmark indicated by the dashed lines. A larger external sample likewise yields a greater gain. Figure 2(b) displays the same quantity as individual components of k and ρ are successively treated as known, proceeding from all seven unknown to none along two distinct orderings. Additional knowledge of the shift model yields a further gain at every step, and the two orderings converge to the same benchmark.

Experiment III. In this experiment, we examine the robustness of the different estimators to potential model misspecification. We consider four scenarios, where S1 specifies μ_t, k and ρ correctly, S2 misspecifies μ_t alone, S3 misspecifies k and ρ alone, and S4 misspecifies all three. Each misspecified shift model omits a quadratic term present in the true model, and the misspecified outcome model uses $|x|$ in place of x as covariates. All estimators are provided with a validation sample of the same size as the primary sample.

Figure 3 presents boxplots of the trial-only estimator $\widetilde{\tau}_t$, the augmented estimator $\widehat{\tau}_t$, and the shrinkage estimator $\widehat{\tau}_t^{(s)}$ across the four scenarios. Under S1 and S2, all three estimators are centred at the true value, illustrating the robustness of $\widehat{\tau}_t$ to misspecification of μ_t, as established in Theorem 1. Under S3 and S4, in which k and ρ are misspecified, $\widehat{\tau}_t$ is visibly biased, whereas $\widehat{\tau}_t^{(s)}$ remains centred at the truth. These results indicate that the shrinkage estimator preserves consistency even when the models for the external data are misspecified.

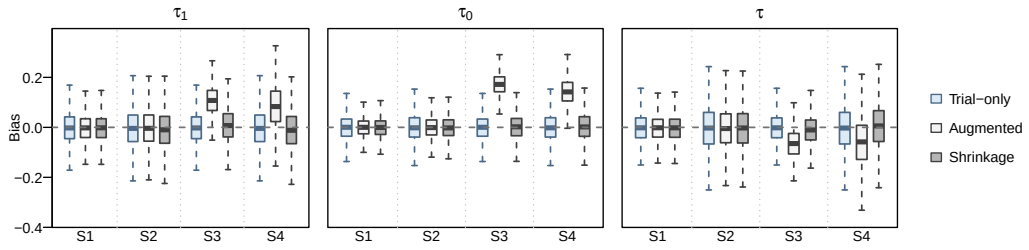

Figure 3: Empirical bias under four working-model misspecification scenarios.

Real-data application. We evaluate our method on the LaLonde training-program benchmark, combining data from the National Supported Work (NSW), a randomized job-training program, with an external control group from the Panel Study of Income Dynamics (PSID) (Lalonde, 1986; Dehejia & Wahba, 1999). In the interest of space, detailed results are deferred to Appendix C.

7 Conclusions

Related work. External controls have long been used to improve trial precision, dating back to historical-control combination designs (Pocock, 1976). Under exchangeability between trial and external controls, a range of methods have been proposed to borrow EC information for trial inference (Dahabreh et al., 2019; Li et al., 2023; Valancius et al., 2024; Wang et al., 2024b). Related historical-borrowing approaches assess compatibility or discount external information before borrowing, including test-then-pool methods (Viele et al., 2014) and Bayesian dynamic borrowing (Hobbs et al., 2012; Schmidli et al., 2014).

When exchangeability is questionable, recent work has developed more guarded strategies for incorporating external controls. These include selective borrowing (Gao et al., 2025; Yang et al., 2025), weighted or shrinkage-based combinations (Oberst et al., 2022; Rosenman et al., 2023; Yang et al., 2023), and prognostic covariate adjustment (Schuler et al., 2022; Liao et al., 2025). Randomization-aware combinations further preserve trial-based consistency and can protect against efficiency loss when EC is unreliable (Wang et al., 2024a; Karlsson et al., 2025; Bartolomeis et al., 2026). Our work differs from these lines by explicitly exploiting the distribution shifts that void exchangeability, using them to calibrate EC data toward the trial population, and combining the resulting augmented estimator with a trial-only benchmark via adaptive shrinkage to retain trial-based guarantees.

More broadly, EC borrowing is connected to the broader literature on data fusion and integration, including efficient fusion (Li & Luedtke, 2023), multi-source integration (Yang & Ding, 2020), bias correction and double machine learning (Parikh et al., 2025), long-term effects (Ghassami et al., 2022), and transfer learning (Wu & Yang, 2023); see recent reviews for broader context (Colnet et al., 2024; Lin et al., 2024).

Position of our work in the literature. Our work departs from existing approaches in three main ways. First, while most prior methods rely on some form of exchangeability between RCT and the EC population, we treat distribution shift as a structural feature of the data rather than a nuisance to be assumed away, and we directly incorporate it into estimation through calibration equations. Second, in contrast to guarded approaches that defensively attenuate EC influence via selection, down-weighting, or shrinkage, our framework aims to actively exploit EC data under a transparent shift model; in particular, the adaptive shrinkage proposal provides a principled safeguard that recovers the trial-only benchmark under shift-model misspecification. Third, whereas existing methods typically borrow EC information only for the control arm, our calibration-based augmentation leverages EC covariate information for both arms, so that, when properly aligned with the trial population, EC data can sharpen inference for treatment-arm functionals as well.

Methodologically, our framework connects to the broader literature on distribution shift and on data fusion and transportability in causal inference, specialized to the semiparametric efficiency calculus of trial inference. The combination of explicit shift modeling, dual-arm borrowing, and provable efficiency dominance distinguishes our paper from prior work in the literature.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing codes for the numerical experiments, to provide feedback on the manuscript, and to polish the paper for readability. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

REFERENCES

Timothy B. Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018. doi: https://doi.org/10.3982/ECTA14434. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14434.

Piersilvio De Bartolomeis, Javier Abad, Guanbo Wang, Konstantin Donhauser, Raymond M Duch, Fanny Yang, and Issa Dahabreh. Efficient randomized experiments using foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=WGYEOAB0Rk.

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. *Statistical Science*, 39(1):165–191, 2024. doi: 10.1214/23-STS889.

Issa J. Dahabreh, Sarah E. Robertson, Eric J. Tchetgen, Elizabeth A. Stuart, and Miguel A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, June 2019. ISSN 0006-341X. doi: 10.1111/biom.13009. URL https://doi.org/10.1111/biom.13009.

Rh Dehejia and S Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, December 1999. ISSN 0162-1459. doi: 10.2307/2669919.

David L. Donoho. Statistical Estimation and Optimal Recovery. *The Annals of Statistics*, 22(1): 238 – 270, 1994. doi: 10.1214/aos/1176325367. URL https://doi.org/10.1214/aos/1176325367.

Chenyin Gao, Shu Yang, Mingyang Shan, Wenyu Ye, Ilya Lipkovich, and Douglas Faries. Improving randomized controlled trial analysis via data-adaptive borrowing. *Biometrika*, 112(2): asae069, 2025. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asae069. URL https://doi.org/10.1093/biomet/asae069.

AmirEmad Ghassami, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv preprint arXiv:2201.10743*, April 2022. URL http://arxiv.org/abs/2201.10743.

Brian P. Hobbs, Daniel J. Sargent, and Bradley P. Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7(3):639–674, 2012. doi: 10.1214/12-BA722.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. URL https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB.

Rickard Karlsson, Guanbo Wang, Piersilvio De Bartolomeis, Jesse H. Krijthe, and Issa J. Dahabreh. Robust integration of external control data in randomized trials. *arXiv preprint arXiv:2406.17971*, September 2025. doi: 10.48550/arXiv.2406.17971. URL http://arxiv.org/abs/2406.17971.

Jae Kwang Kim and Cindy Long Yu. A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106(493):157–165, March 2011. ISSN 0162-1459, 1537-274X. doi: 10.1198/jasa.2011.tm10104. URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2011.tm10104.

Rj Lalonde. Evaluating the econometric evaluations of training-programs with experimental-data. *American Economic Review*, 76(4):604–620, September 1986. ISSN 0002-8282.

Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054, December 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad007. URL https://doi.org/10.1093/biomet/asad007.

Xinyu Li, Wang Miao, Fang Lu, and Xiao-Hua Zhou. Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403, 2023. ISSN 1541-0420. doi: 10.1111/biom.13583. URL https://doi.org/10.1111/biom.13583.

Lauren D. Liao, Emilie Højbjerre-Frandsen, Alan E. Hubbard, and Alejandro Schuler. Prognostic adjustment with efficient estimators to unbiasedly leverage historical data in randomized trials. *The International Journal of Biostatistics*, 21(1):1–15, 2025. doi: 10.1515/ijb-2024-0018.

Xi Lin, Jens Magelund Tarp, and Robin J. Evans. Data fusion for efficiency gain in ATE estimation: A practical review with simulations. *arXiv preprint arXiv:2407.01186*, 2024. doi: 10.48550/arXiv.2407.01186.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pp. 2111–2245. Elsevier, January 1994. doi: 10.1016/S1573-4412(05)80005-4. URL https://www.sciencedirect.com/science/article/pii/S1573441205800054.

Michael Oberst, Alexander D'Amour, Minmin Chen, Yuyan Wang, David Sontag, and Steve Yadlowsky. Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. 2022. doi: 10.48550/ARXIV.2205.10467.

Harsh Parikh, Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. A double machine learning approach for combining experimental and observational studies. *Observational Studies*, 11(3):249–300, 2025. doi: 10.1353/obs.2025.a973068.

Stuart J. Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188, 1976. doi: 10.1016/0021-9681(76)90044-8.

James M. Robins, Andrea Rotnitzky, and Lueping Zhao. Estimation of regression-coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 (427):846–866, September 1994. ISSN 0162-1459. doi: 10.2307/2290910.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, April 1983. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Evan T.R. Rosenman, Guillaume Basse, Art B. Owen, and Mike Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 2023. doi: 10.1111/biom.13827.

Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014. doi: 10.1111/biom.12242.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, Charles Fisher, for the Critical Path for Alzheimer's Disease, the Alzheimer's Disease Neuroimaging Initiative, and the Alzheimer's Disease Cooperative Study. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2022. doi: 10.1515/ijb-2021-0072.

Michael Valancius, Herbert Pang, Jiawen Zhu, Stephen R Cole, Michele Jonsson Funk, and Michael R Kosorok. A causal inference framework for leveraging external controls in hybrid trials. *Biometrics*, 80(4):ujae095, 2024. doi: 10.1093/biomtc/ujae095.

Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G. Ibrahim, Nelson Kinnersley, Stacy Lindborg, Sandrine Micallef, Satrajit Roychoudhury, and Laura Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2014. doi: 10.1002/pst.1589.

Guanbo Wang, Rickard Karlsson, Piersilvio De Bartolomeis, James Robins, Eric Tchetgen Tchetgen, and Issa J Dahabreh. Robust trial augmentation using external data. *arXiv preprint arXiv:2505.01201*, 2024a.

Guanbo Wang, Melanie Poulin-Costello, Herbert Pang, Jiawen Zhu, Hans-Joachim Helms, Irmarie Reyes-Rivera, Robert W Platt, Menglan Pang, and Artemis Koukounari. Evaluating hybrid controls methodology in early-phase oncology trials: A simulation study based on the morpheus-uc trial. *Pharmaceutical Statistics*, 23(1):31–45, 2024b.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 0012-9682. doi: 10.2307/1912526. URL https://www.jstor.org/stable/1912526.

Lili Wu and Shu Yang. Transfer learning of individualized treatment rules from experimental to real-world data. *Journal of Computational and Graphical Statistics*, 32(3):1036–1045, 2023. doi: 10.1080/10618600.2022.2141752.

Qinwei Yang, Jingyi Li, and Peng Wu. Adaptive data-borrowing for improving treatment effect estimation using external controls. *arXiv preprint arXiv:2508.03282*, October 2025. doi: 10.48550/arXiv.2508.03282. URL http://arxiv.org/abs/2508.03282.

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, July 2020. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2019.1609973. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2019.1609973.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596, July 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad017. URL https://doi.org/10.1093/jrsssb/qkad017.

A Local Misspecification and Alternative Bias-Aware Estimator

In the conventional data-generating process, we pre-determine either $\tau_t^* = \tau_t$ or $\tau_t^* \neq \tau_t$, with the distinction fixed as the sample size N grows. *Local misspecification* is the case in between, with $\tau_t^* - \tau_t = hN^{-1/2}$ for a fixed h, so that the discrepancy has the same order as the standard error of $\widetilde{\tau}_t$ and a larger sample never settles which of the two holds. It becomes relevant in practice when the trial is small enough that τ_t^* and τ_t differ by little relative to the sampling error. The studentized discrepancy $T_{t,N} := (\widehat{\tau}_t - \widetilde{\tau}_t)/\sqrt{V_{t,N}}$ is then $O_p(1)$, so that $\delta_{t,N} \xrightarrow{p} 1$, $\widehat{\lambda}_{t,N} \xrightarrow{p} \lambda_t^*$, and

$$\sqrt{N}\{\widehat{\tau}_t^{(s)} - \tau_t\} \xrightarrow{d} \mathcal{N}(\lambda_t^* h, v_t). \tag{8}$$

Therefore, the coverage falls below the nominal level whenever $h \neq 0$, though the point estimate remains $\sqrt{N}$-consistent.

One remedy is to bound the bias manually by the analyst. Suppose one is willing to assert $|\tau_t^* - \tau_t| \leq B$. Let

$$\lambda_B = \lambda_t^* \frac{V_{t,N}}{V_{t,N} + B^2},$$

which has the form of $\lambda_{t,N} = \delta_{t,N}\lambda_t^*$ with the data-driven factor $N^{-\gamma}T_{t,N}^4$ of (7) replaced by the deterministic $B^2/V_{t,N}$. Let $\widehat{\sigma}_t^2(\lambda) = \boldsymbol{c}(\lambda)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_t\boldsymbol{c}(\lambda)/N$ for the estimated variance of $\widehat{\tau}_t(\lambda)$. Then, a bias-aware estimator is given by

$$\widehat{\tau}_t^{(B)} = \widetilde{\tau}_t + \widehat{\lambda}_B(\widehat{\tau}_t - \widetilde{\tau}_t), \quad \text{with } \widehat{\lambda}_B = \widehat{\lambda}_t^* \frac{\widehat{V}_{t,N}}{\widehat{V}_{t,N} + B^2}.$$

Based on this estimator, a confidence interval at nominal level $1-\alpha$ that explicitly accounts for the bias is

$$\widehat{\tau}_t^{(B)} \pm \mathrm{cv}_{1-\alpha}\left\{\frac{\widehat{\lambda}_B B}{\widehat{\sigma}_t(\widehat{\lambda}_B)}\right\}\widehat{\sigma}_t(\widehat{\lambda}_B),$$

where $\mathrm{cv}_{1-\alpha}(r)$ denotes the $1-\alpha$ quantile of $|\mathcal{N}(r,1)|$, the fixed-length critical value (Donoho, 1994; Armstrong & Kolesár, 2018). By construction, it attains coverage at least $1-\alpha$ uniformly over $|\tau_t^* - \tau_t| \leq B$.

The bias-aware estimator is motivated differently from the shrinkage estimator proposed in Section 5. Specifically, the shrinkage estimator decides whether, and by how much, to borrow from the external data based on the observed discrepancy between the two estimators. This strategy cannot detect a local bias by construction, as it is indistinguishable from sampling noise at the relevant scale. In contrast, the bias-aware estimator λ_B is built on how much bias one is prepared to tolerate. It yields valid inference throughout the range at the cost of a weight fixed by assumption rather than learned, and of no protection beyond the bound. Since B traces the entire path from the oracle combination $\widehat{\tau}_t(\lambda_t^*)$ at $B = 0$ to the trial-only estimator as $B \to \infty$, it is recommended to report the estimate and the interval across a range of B for sensitivity analysis.

B Additional Numerical Results

This appendix collects the full numerical summaries underlying the figures of Section 6. Throughout, "Rel. var." denotes the ratio of the empirical variance of a given estimator to that of the trial-only estimator, which is the quantity the figures plot, "Rel. MSE" the corresponding ratio of mean squared errors, $\overline{\mathrm{SE}}$ the root mean square of the estimated standard errors. Monte Carlo standard errors for the coverage entries are at most 0.005 with 2000 replicates.

In summary, Table A.1 and Figure A.1 accompany the first experiment, reporting the values underlying Figure 1 together with results for the two arm-specific estimands. Tables A.2–A.4 accompany the second experiment, where Table A.2 records the bias and coverage, and the remaining two report the values underlying its two panels. Table A.5 accompanies the third experiment.

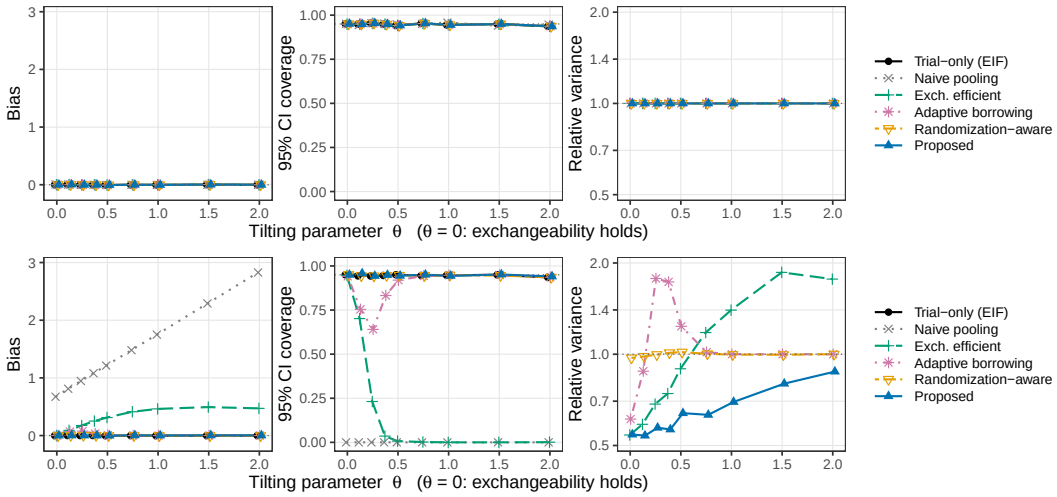

Figure A.1: Experiment I. Arm-specific estimands for τ_1 (top) and τ_0 (bottom).

Table A.1: Experiment I. Comparison for τ as the tilting parameter θ grows ($n = 500$, $N = 1500$), corresponding to Figure 1. The proposed estimator treats θ as known.

Method	Metric	0.00	0.12	0.25	0.38	0.50	0.75	1.00	1.50	2.00
Trial-only (EIF)	Bias	0.001	0.003	-0.001	-0.001	-0.002	-0.001	-0.001	0.004	-0.001
	Coverage	0.949	0.936	0.955	0.951	0.940	0.958	0.946	0.943	0.951
	Rel. var.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Naive pooling	Bias	-0.668	-0.808	-0.949	-1.081	-1.218	-1.482	-1.753	-2.292	-2.827
	Coverage	0.014	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	Rel. var.	2.64	2.52	2.42	2.52	2.52	2.87	2.89	3.01	3.59
Exch. efficient	Bias	0.000	-0.087	-0.176	-0.251	-0.317	-0.414	-0.465	-0.490	-0.472
	Coverage	0.947	0.802	0.402	0.126	0.028	0.002	0.001	0.002	0.004
	Rel. var.	0.67	0.65	0.68	0.71	0.72	0.91	1.01	1.33	1.31
Adaptive borrowing	Bias	0.000	-0.065	-0.079	-0.033	-0.012	-0.001	-0.001	0.004	-0.001
	Coverage	0.942	0.828	0.716	0.855	0.917	0.956	0.947	0.942	0.952
	Rel. var.	0.72	0.90	1.64	1.56	1.17	1.01	1.00	1.00	1.00
Randomization-aware	Bias	0.000	-0.000	-0.007	-0.007	-0.009	-0.006	-0.006	0.001	-0.003
	Coverage	0.950	0.939	0.953	0.946	0.938	0.953	0.945	0.943	0.953
	Rel. var.	0.98	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
Proposed	Bias	0.000	0.002	-0.003	-0.001	-0.004	-0.003	-0.004	-0.000	-0.003
	Coverage	0.949	0.949	0.955	0.950	0.946	0.956	0.946	0.950	0.947
	Rel. var.	0.67	0.65	0.67	0.68	0.68	0.75	0.77	0.84	0.92

Table A.2: Experiment II. Bias, dispersion and coverage as the amount of information about the shift varies at $n = 1500$ and $N = 3500$. Here ξ is the size of the validation sample relative to the primary sample. Figure 2 plots relative variance only.

Method	ξ	Bias	SD	$\overline{\text{SE}}$	SE/SD	Coverage	Rel. var.
τ_1							
Trial-only (EIF)	–	-0.0021	0.064	0.062	0.96	0.939	1.000
Proposed	0	-0.0021	0.064	0.062	0.96	0.939	1.000
	0.5	-0.0019	0.059	0.057	0.97	0.950	0.839
	1	-0.0028	0.057	0.055	0.97	0.934	0.777
	2	-0.0006	0.053	0.052	0.98	0.949	0.678
	4	-0.0014	0.051	0.050	0.97	0.944	0.635
Proposed (known shift)	–	-0.0009	0.047	0.046	0.98	0.945	0.541
τ_0							
Trial-only (EIF)	–	-0.0007	0.050	0.048	0.96	0.939	1.000
Proposed	0	-0.0007	0.050	0.048	0.96	0.939	1.000
	0.5	-0.0003	0.044	0.043	0.98	0.942	0.752
	1	-0.0011	0.040	0.039	0.98	0.948	0.631
	2	0.0002	0.035	0.036	1.01	0.956	0.494
	4	0.0001	0.033	0.033	1.00	0.951	0.424
Proposed (known shift)	–	0.0005	0.026	0.027	1.04	0.959	0.274
τ							
Trial-only (EIF)	–	-0.0014	0.057	0.056	0.97	0.949	1.000
Proposed	0	-0.0014	0.057	0.056	0.97	0.949	1.000
	0.5	-0.0016	0.053	0.052	0.98	0.946	0.858
	1	-0.0017	0.051	0.050	0.98	0.942	0.805
	2	-0.0007	0.050	0.048	0.97	0.946	0.754
	4	-0.0015	0.048	0.047	0.97	0.951	0.711
Proposed (known shift)	–	-0.0014	0.046	0.044	0.97	0.941	0.640

Table A.3: Experiment II. Variance relative to trial-only as the validation size ξ and the trial fraction κ vary, with the Proposition 1 limit in the last column.

Estimand	κ	$\xi = 0$	$\xi = 0.5$	$\xi = 1$	$\xi = 2$	$\xi = 4$	known	Prop. 1
τ_1	0.500	1.000	0.838	0.768	0.678	0.637	0.560	0.599
	0.150	1.000	0.862	0.723	0.632	0.547	0.435	0.423
	0.050	1.000	0.814	0.711	0.592	0.508	0.399	0.373
τ_0	0.500	1.000	0.802	0.683	0.555	0.474	0.347	0.377
	0.150	1.000	0.709	0.555	0.407	0.282	0.107	0.120
	0.050	1.000	0.699	0.522	0.361	0.225	0.044	0.043
τ	0.500	1.000	0.900	0.836	0.766	0.723	0.661	0.667
	0.150	1.000	0.847	0.744	0.676	0.618	0.509	0.510
	0.050	1.000	0.844	0.769	0.636	0.560	0.477	0.459

Table A.4: Experiment II. Variance relative to trial-only when part of the shift is fixed at its true value, corresponding to Figure 2(b).

Free parameters	τ_1			τ_0			τ		
in (k, ρ)	0.50	0.15	0.05	0.50	0.15	0.05	0.50	0.15	0.05
$(3, 4)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$(2, 4)$	0.931	0.923	0.919	0.916	0.854	0.884	1.001	1.004	1.024
$(1, 4)$	0.584	0.417	0.394	0.710	0.588	0.603	0.926	0.887	0.921
$(3, 1)$	1.000	1.000	1.000	0.683	0.509	0.466	0.743	0.627	0.576
$(0, 4)$	0.560	0.414	0.394	0.694	0.586	0.603	0.922	0.887	0.921
$(3, 0)$	1.000	1.000	1.000	0.654	0.506	0.465	0.729	0.626	0.580
$(2, 0)$	0.931	0.923	0.919	0.590	0.391	0.336	0.729	0.627	0.586
$(1, 0)$	0.584	0.417	0.394	0.363	0.111	0.046	0.665	0.518	0.476
$(0, 1)$	0.560	0.414	0.394	0.379	0.113	0.044	0.674	0.518	0.473
$(0, 0)$	0.560	0.414	0.394	0.347	0.110	0.046	0.661	0.517	0.476

Table A.5: Experiment III. Robustness under the four specification scenarios at $n = 1500$, $N = 3500$, corresponding to Figure 3.

Scenario	Estimator	Bias			Coverage		
		τ_1	τ_0	τ	τ_1	τ_0	τ
S1	Trial-only	-0.0021	-0.0007	-0.0014	0.939	0.939	0.949
	Augmented	-0.0023	-0.0004	-0.0018	0.943	0.952	0.950
	Shrinkage	-0.0028	-0.0012	-0.0016	0.937	0.941	0.945
S2	Trial-only	-0.0027	-0.0006	-0.0021	0.949	0.946	0.949
	Augmented	-0.0030	-0.0008	-0.0021	0.953	0.951	0.939
	Shrinkage	-0.0099	-0.0016	-0.0017	0.944	0.941	0.938
S3	Trial-only	-0.0021	-0.0007	-0.0014	0.939	0.939	0.949
	Augmented	0.1078	0.1719	-0.0642	0.558	0.032	0.785
	Shrinkage	0.0077	0.0016	-0.0100	0.914	0.939	0.938
S4	Trial-only	-0.0027	-0.0006	-0.0021	0.949	0.946	0.949
	Augmented	0.0845	0.1427	-0.0582	0.839	0.253	0.918
	Shrinkage	-0.0104	0.0029	0.0063	0.934	0.940	0.943

C Real Data Application

We evaluate our method on the LaLonde training-program benchmark, combining data from the National Supported Work (NSW), a randomized job-training program, with an external control group from the Panel Study of Income Dynamics (PSID) (Lalonde, 1986; Dehejia & Wahba, 1999). The outcome is the change in earnings between 1975 and 1978, defined as `re_dif` $=$ `re78` $-$ `re75`. We adjust for six pre-treatment covariates: `Age`, `Education`, `Black`, `Hispanic`, `Married`, and `Nodegree`. To mimic a validation setting, we randomly split the data into two equal parts; one is used for primary estimation, and the other is treated as a validation sample.

It is worth noting that the NSW trial and the PSID external controls are drawn from distinct data sources, with different sampling frames, collection periods, and measurement protocols, so that distribution shift between them is to be expected. Figure A.2 illustrates this shift. Panel (a) shows the joint covariate distribution by projecting the mixed covariates onto their leading principal components, revealing separation between the two samples in covariate space. Panels (b)-(c) further show marginal covariate shift. Panel (d) examines conditional shift by grouping observations into deciles of a pooled covariate-based prediction score and comparing the outcome distributions within each decile. Differences in the boxplots within the same score stratum suggest that the conditional outcome distributions $p(y_0 \mid \mathbf{x})$ and $q(y_0 \mid \mathbf{x})$ differ across the two samples.

Exchangeability is therefore unlikely to hold in this application, which motivates applying the proposed procedure to model the discrepancy between the two data sources explicitly. In particular, we adopt logistic working models for $k(\mathbf{x})$ and $\rho(\mathbf{x}, y_0)$, and estimate these nuisance components from the validation set. Table A.6 reports the resulting estimates, standard deviations, and 95% confidence intervals for the trial-only estimator $\widetilde{\tau}_t$, the augmented estimator $\widehat{\tau}_t$, and the shrinkage estimator $\widehat{\tau}_t^{(s)}$. Without a validation set, directly incorporating EC data does not uniformly reduce variability relative to the trial-only estimator: the SDs of $\widehat{\tau}_1$ and $\widehat{\tau}$ are slightly larger than those of the corresponding trial-only baselines $\widetilde{\tau}_1$ and $\widetilde{\tau}$, whereas the SD of $\widehat{\tau}_0$ is slightly smaller. By contrast, the shrinkage estimator retains the theoretical efficiency safeguard, with SDs no larger than those of the trial-only baselines across all three estimands. With a validation set, the EC-augmented estimators yield clearer variance reductions, especially for τ_0 and τ, and the shrinkage estimator preserves these gains while maintaining its efficiency guarantee.

Table A.6: Real-data analysis results for the NSW trial and PSID external control samples.

Estimand		Trial-only	Without validation set		With validation set	
			Augmented	Shrinkage	Augmented	Shrinkage
τ_1	Est. (SD)	2.418 (0.551)	2.399 (0.568)	2.444 (0.543)	2.416 (0.546)	2.411 (0.542)
	95% CI	[1.338, 3.498]	[1.286, 3.512]	[1.380, 3.508]	[1.346, 3.486]	[1.349, 3.473]
τ_0	Est. (SD)	1.567 (0.436)	1.686 (0.429)	1.787 (0.427)	1.647 (0.355)	1.653 (0.355)
	95% CI	[0.712, 2.422]	[0.845, 2.527]	[0.950, 2.624]	[0.951, 2.343]	[0.957, 2.349]
τ	Est. (SD)	0.851 (0.702)	0.713 (0.711)	0.907 (0.701)	0.769 (0.649)	0.759 (0.648)
	95% CI	[-0.524, 2.227]	[-0.681, 2.107]	[-0.467, 2.280]	[-0.503, 2.041]	[-0.512, 2.030]

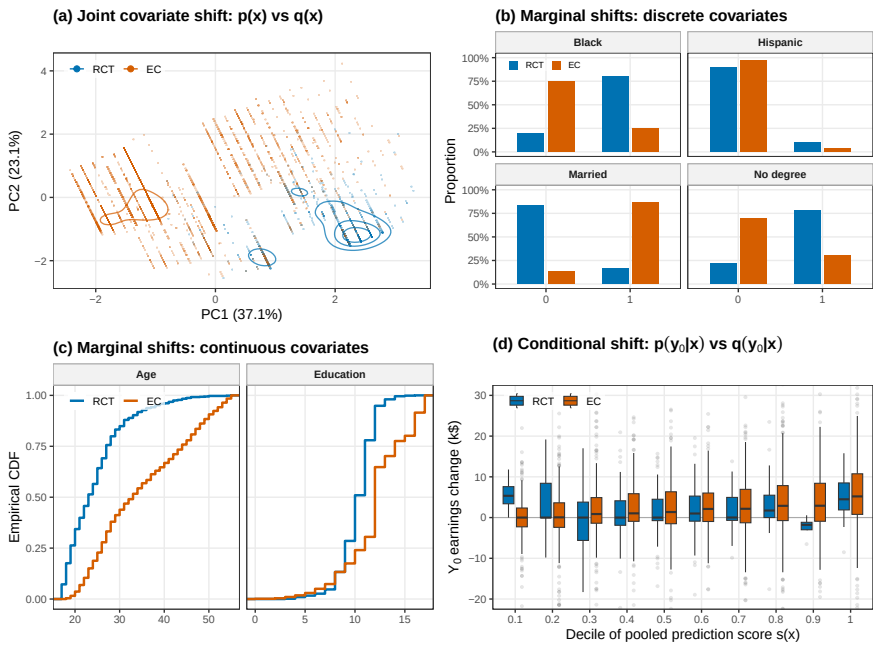

Figure A.2: Visualization of distribution shifts between the NSW RCT and PSID EC samples. (a) PCA projection of pooled covariates; percentages indicate variance explained, and contours show group-specific densities. (b)-(c) Marginal shifts in discrete covariates via group-wise proportions and in continuous covariates via Empirical CDFs. (d) Boxplots of observed Y_0 within deciles of the pooled prediction score $s(\mathbf{x}) = \widehat{\mathbb{E}}(Y_0 \mid \mathbf{x})$.

D Proof of Lemma 1

Proof. Noting that

$$f(\mathbf{x}) = \frac{\mathrm{pr}(R=1)p(\mathbf{x})}{\mathrm{pr}(R=1 \mid \mathbf{x})} = \frac{\kappa p(\mathbf{x})}{k(\mathbf{x})},$$

we have

$$\mathbb{E}\left\{g_1(\mathbf{X}) \mid R=1\right\} = \int g_1(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \int g_1(\mathbf{x})\frac{k(\mathbf{x})}{\kappa}f(\mathbf{x})d\mathbf{x} = \mathbb{E}\left\{\frac{k(\mathbf{x})}{\kappa}g_1(\mathbf{X})\right\}.$$

For the second part, noting that

$$f(\mathbf{x}, y_0, T=0) = \frac{\mathrm{pr}(R=1, T=0)p(\mathbf{x})p(y_0 \mid \mathbf{x})}{\mathrm{pr}(R=1 \mid \mathbf{x}, y_0, T=0)} = \frac{\kappa(1-\pi)}{\rho(\mathbf{x}, y_0)}p(\mathbf{x})p(y_0 \mid \mathbf{x}).$$

we have

$$\begin{aligned}\mathbb{E}\left\{g_2(\mathbf{X}, Y_0) \mid R=1\right\} &= \iint g_2(\mathbf{x}, y_0)p(y_0 \mid \mathbf{x})p(\mathbf{x})dy_0 d\mathbf{x} \\ &= \iint g_2(\mathbf{x}, y_0)\frac{\rho(\mathbf{x}, y_0)}{\kappa(1-\pi)}f(\mathbf{x}, y_0, T=0)dy_0 d\mathbf{x} = \mathbb{E}\left\{\frac{1-T}{\kappa(1-\pi)}\rho(\mathbf{X}, Y_0)g_2(\mathbf{X}, Y_0)\right\}.\end{aligned}$$

This completes the proof. □

E Proof of Proposition 1

Proof. We first consider ψ_1. Note that

$$\psi_1 - \widetilde{\psi}_1 = -\frac{1}{\kappa}\{\mu_1(\mathbf{X}) - \tau_1\}\{R - k(\mathbf{X})\}.$$

Denote

$$S_1 = \frac{RT}{\kappa\pi}\{Y_1 - \mu_1(\mathbf{X})\},\ \ S_2 = \frac{R}{\kappa}\left\{\mu_1(\mathbf{X}) - \tau_1\right\}, \text{ and } S_3 = -\frac{1}{\kappa}\{\mu_1(\mathbf{X}) - \tau_1\}\{R - k(\mathbf{X})\}.$$

Then, we have $\widetilde{\psi}_1 = S_1 + S_2$ and $\psi_1 = S_1 + S_2 + S_3$. It is easy to check $\mathbb{E}(S_1) = \mathbb{E}(S_2) = \mathbb{E}(S_3) = \mathbb{E}(S_1S_2) = \mathbb{E}(S_1S_3) = 0$. Thus, we have $\mathbb{E}(\widetilde{\psi}_1^2) = \mathbb{E}(S_1^2) + \mathbb{E}(S_2^2)$, and $\mathbb{E}(\psi_1^2) = \mathbb{E}(S_1^2) + \mathbb{E}(S_2^2) + \mathbb{E}(S_3^2) + 2\mathbb{E}(S_2S_3)$. We can calculate that

$$\mathbb{E}(S_3^2) = \frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_1(\mathbf{x}) - \tau_1\}^2 k(\mathbf{x})\{1 - k(\mathbf{x})\}\right],$$
$$\mathbb{E}(S_2S_3) = -\frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_1(\mathbf{x}) - \tau_1\}^2 k(\mathbf{x})\{1 - k(\mathbf{x})\}\right].$$

Therefore, we have

$$\mathbb{E}(\psi_1^2) - \mathbb{E}(\widetilde{\psi}_1^2) = \mathbb{E}(S_3^2) + 2\mathbb{E}(S_2S_3) = -\frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_1(\mathbf{x}) - \tau_1\}^2 k(\mathbf{x})\{1 - k(\mathbf{x})\}\right] \le 0.$$

The equality holds if and only if $k(\mathbf{X}) = 1$ almost surely, equivalently, $\kappa = 1$.

Next, we consider ψ_0. Note that

$$\psi_0 - \widetilde{\psi}_0 = -\frac{1}{\kappa}\{\mu_0(\mathbf{X}) - \tau_0\}\{R - k(\mathbf{X})\} - \frac{1 - T}{\kappa(1 - \pi)}\{Y_0 - \mu_0(\mathbf{X})\}\left\{R - \rho(\mathbf{X}, Y_0)\right\}.$$

We slightly abuse the notation and denote

$$S_1 = \frac{R(1 - T)}{\kappa(1 - \pi)}\{Y_0 - \mu_0(\mathbf{X})\},$$
$$S_2 = \frac{R}{\kappa}\left\{\mu_0(\mathbf{X}) - \tau_0\right\},$$
$$S_3 = -\frac{1}{\kappa}\{\mu_0(\mathbf{X}) - \tau_0\}\{R - k(\mathbf{X})\},$$
$$S_4 = -\frac{1 - T}{\kappa(1 - \pi)}\{Y_0 - \mu_0(\mathbf{X})\}\left\{R - \rho(\mathbf{X}, Y_0)\right\}.$$

We have $\widetilde{\psi}_0 = S_1 + S_2$ and $\psi_0 = S_1 + S_2 + S_3 + S_4$. It is easy to check $\mathbb{E}(S_1) = \mathbb{E}(S_2) = \mathbb{E}(S_3) = \mathbb{E}(S_4) = \mathbb{E}(S_1S_2) = \mathbb{E}(S_1S_3) = 0$. Thus, we have $\mathbb{E}(\widetilde{\psi}_0^2) = \mathbb{E}(S_1^2) + \mathbb{E}(S_2^2)$, and

$$\mathbb{E}(\psi_0^2) = \mathbb{E}(S_1^2) + \mathbb{E}(S_2^2) + \mathbb{E}(S_3^2) + \mathbb{E}(S_4^2) + 2\mathbb{E}(S_1S_4) + 2\mathbb{E}(S_2S_3) + 2\mathbb{E}(S_2S_4) + 2\mathbb{E}(S_3S_4).$$

We can calculate that

$$\mathbb{E}(S_3^2) = \frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_0(\mathbf{X}) - \tau_0\}^2 k(\mathbf{X})\{1 - k(\mathbf{X})\}\right],$$
$$\mathbb{E}(S_4^2) = \frac{1}{\kappa^2(1 - \pi)^2}\mathbb{E}\left[(1 - T)\{Y_0 - \mu_0(\mathbf{X})\}^2\rho(\mathbf{X}, Y_0)\{1 - \rho(\mathbf{X}, Y_0)\}\right],$$
$$\mathbb{E}(S_1S_4) = -\frac{1}{\kappa^2(1 - \pi)^2}\mathbb{E}\left[(1 - T)\{Y_0 - \mu_0(\mathbf{X})\}^2\rho(\mathbf{X}, Y_0)\{1 - \rho(\mathbf{X}, Y_0)\}\right],$$
$$\mathbb{E}(S_2S_3) = -\frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_0(\mathbf{X}) - \tau_0\}\{\mu_0(\mathbf{X}) - \tau_0\}k(\mathbf{X})\{1 - k(\mathbf{X})\}\right],$$
$$\mathbb{E}(S_2S_4) = -\frac{1}{\kappa^2(1 - \pi)}\mathbb{E}\left[(1 - T)\{Y_0 - \mu_0(\mathbf{X})\}\left\{\mu_0(\mathbf{X}) - \tau_0\right\}\rho(\mathbf{X}, Y_0)\{1 - \rho(\mathbf{X}, Y_0)\}\right],$$
$$\mathbb{E}(S_3S_4) = \frac{1}{\kappa^2(1 - \pi)}\mathbb{E}\left[(1 - T)\{Y_0 - \mu_0(\mathbf{X})\}\{\mu_0(\mathbf{X}) - \tau_0\}\rho(\mathbf{X}, Y_0)\{1 - \rho(\mathbf{X}, Y_0)\}\right].$$

Therefore, we have

$$\begin{aligned}\mathbb{E}(\psi^2) - \mathbb{E}(\widetilde{\psi}^2) &= \mathbb{E}(S_3^2) + \mathbb{E}(S_4^2) + 2\mathbb{E}(S_1S_4) + 2\mathbb{E}(S_2S_3) + 2\mathbb{E}(S_2S_4) + 2\mathbb{E}(S_3S_4) \\ &= -\frac{1}{\kappa^2}\mathbb{E}\left[\{\mu_0(\mathbf{X}) - \tau_0\}^2 k(\mathbf{X})\left\{1 - k(\mathbf{X})\right\}\right]\end{aligned}$$

$$-\frac{1}{\kappa^2(1-\pi)^2}\mathbb{E}\left[(1-T)\{Y-\mu_0(\mathbf{X})\}^2\rho(\mathbf{X},Y)\left\{1-\rho(\mathbf{X},Y)\right\}\right]\leq 0.$$

The equality holds if and only if $k(\mathbf{X})=\rho(\mathbf{X},Y)=1$ almost surely, equivalently, $\kappa=1$.

Finally, we consider ψ. Denote $\Delta(\mathbf{x})=\mu_1(\mathbf{x})-\mu_0(\mathbf{x})$. Note that

$$\psi-\widetilde{\psi}=-\frac{1}{\kappa}\{\Delta(\mathbf{X})-\tau\}\{R-k(\mathbf{X})\}+\frac{1-T}{\kappa(1-\pi)}\{Y_0-\mu_0(\mathbf{X})\}\left\{R-\rho(\mathbf{X},Y_0)\right\}.$$

We slightly abuse the notation and denote

$$\begin{aligned}
S_1&=\frac{RT}{\kappa\pi}\{Y_1-\mu_1(\mathbf{X})\},\\
S_2&=-\frac{R(1-T)}{\kappa(1-\pi)}\{Y_0-\mu_0(\mathbf{X})\},\\
S_3&=\frac{R}{\kappa}\left\{\Delta(\mathbf{X})-\tau\right\},\\
S_4&=-\frac{1}{\kappa}\{\Delta(\mathbf{X})-\tau\}\{R-k(\mathbf{X})\},\\
S_5&=\frac{1-T}{\kappa(1-\pi)}\{Y_0-\mu_0(\mathbf{X})\}\left\{R-\rho(\mathbf{X},Y_0)\right\}.
\end{aligned}$$

Then we have $\widetilde{\psi}=S_1+S_2+S_3$ and $\psi=S_1+S_2+S_3+S_4+S_5$. It is easy to check $\mathbb{E}(S_1)=\mathbb{E}(S_2)=\mathbb{E}(S_3)=\mathbb{E}(S_4)=\mathbb{E}(S_5)=\mathbb{E}(S_1S_2)=\mathbb{E}(S_1S_3)=\mathbb{E}(S_1S_4)=\mathbb{E}(S_1S_5)=\mathbb{E}(S_2S_3)=\mathbb{E}(S_2S_4)=0$. Thus, we have $\mathbb{E}(\widetilde{\psi}^2)=\mathbb{E}(S_1^2)+\mathbb{E}(S_2^2)+\mathbb{E}(S_3^2)$, and

$\mathbb{E}(\psi^2)=\mathbb{E}(S_1^2)+\mathbb{E}(S_2^2)+\mathbb{E}(S_3^2)+\mathbb{E}(S_4^2)+\mathbb{E}(S_5^2)+2\mathbb{E}(S_2S_5)+2\mathbb{E}(S_3S_4)+2\mathbb{E}(S_3S_5)+2\mathbb{E}(S_4S_5)$.

We can calculate that

$$\begin{aligned}
\mathbb{E}(S_4^2)&=\frac{1}{\kappa^2}\mathbb{E}\left[\{\Delta(\mathbf{X})-\tau\}^2k(\mathbf{X})\{1-k(\mathbf{X})\}\right],\\
\mathbb{E}(S_5^2)&=\frac{1}{\kappa^2(1-\pi)^2}\mathbb{E}\left[(1-T)\{Y_0-\mu_0(\mathbf{X})\}^2\rho(\mathbf{X},Y_0)\{1-\rho(\mathbf{X},Y_0)\}\right],\\
\mathbb{E}(S_2S_5)&=-\frac{1}{\kappa^2(1-\pi)^2}\mathbb{E}\left[(1-T)\{Y_0-\mu_0(\mathbf{X})\}^2\rho(\mathbf{X},Y_0)\{1-\rho(\mathbf{X},Y_0)\}\right],\\
\mathbb{E}(S_3S_4)&=-\frac{1}{\kappa^2}\mathbb{E}\left[\{\Delta(\mathbf{X})-\tau\}\{\Delta(\mathbf{X})-\tau\}k(\mathbf{X})\{1-k(\mathbf{X})\}\right],\\
\mathbb{E}(S_3S_5)&=\frac{1}{\kappa^2(1-\pi)}\mathbb{E}\left[(1-T)\{Y_0-\mu_0(\mathbf{X})\}\left\{\Delta(\mathbf{X})-\tau\right\}\rho(\mathbf{X},Y_0)\{1-\rho(\mathbf{X},Y_0)\}\right],\\
\mathbb{E}(S_4S_5)&=-\frac{1}{\kappa^2(1-\pi)}\mathbb{E}\left[(1-T)\{Y_0-\mu_0(\mathbf{X})\}\{\Delta(\mathbf{X})-\tau\}\rho(\mathbf{X},Y_0)\{1-\rho(\mathbf{X},Y_0)\}\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}(\psi^2)-\mathbb{E}(\widetilde{\psi}^2)&=\mathbb{E}(S_4^2)+\mathbb{E}(S_5^2)+2\mathbb{E}(S_2S_5)+2\mathbb{E}(S_3S_4)+2\mathbb{E}(S_3S_5)+2\mathbb{E}(S_4S_5)\\
&=-\frac{1}{\kappa^2}\mathbb{E}\left[\{\Delta(\mathbf{X})-\tau\}^2k(\mathbf{X})\left\{1-k(\mathbf{X})\right\}\right]\\
&\quad-\frac{1}{\kappa^2(1-\pi)^2}\mathbb{E}\left[(1-T)\{Y-\mu_0(\mathbf{X})\}^2\rho(\mathbf{X},Y)\left\{1-\rho(\mathbf{X},Y)\right\}\right]\leq 0.
\end{aligned}$$

The equality holds if and only if $k(\mathbf{X})=\rho(\mathbf{X},Y)=1$ almost surely, equivalently, $\kappa=1$. This completes the proof.

□

F Proof of Theorem 1

Proof. By the continuous mapping theorem, we have $\widehat{\tau}_1\xrightarrow{p}\tau_1^*$, where

$$\tau_1^*=\frac{\mathbb{E}\left[\frac{RT}{\pi}\{Y_1-\mu_1(\mathbf{X};\boldsymbol{\gamma}_1^*)\}+k(\mathbf{X};\boldsymbol{\alpha}^*)\mu_1(\mathbf{X};\boldsymbol{\gamma}_1^*)\right]}{\mathbb{E}\{k(\mathbf{X};\boldsymbol{\alpha}^*)\}}.$$

Suppose $k(\mathbf{x};\boldsymbol{\alpha}^*) = k(\mathbf{x})$, then we have $\mathbb{E}\{k(\mathbf{X};\boldsymbol{\alpha}^*)\} = \kappa$, and

$$\begin{aligned}\tau_1^* &= \mathbb{E}\left[\frac{RT}{\kappa\pi}\{Y_1 - \mu_1(\mathbf{X};\boldsymbol{\gamma}_1^*)\} + \frac{k(\mathbf{X})}{\kappa}\mu_1(\mathbf{X};\boldsymbol{\gamma}_1^*)\right] \\ &= \iint\{y_1 - \mu_1(\mathbf{x};\boldsymbol{\gamma}_1^*)\}f(\mathbf{x}, y_1 \mid R=1, T=1)\mathrm{d}y_1\mathrm{d}\mathbf{x} + \int \mu_1(\mathbf{x};\boldsymbol{\gamma}_1^*)\frac{\mathrm{pr}(R=1\mid\mathbf{x})}{\kappa}f(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \iint\{y_1 - \mu_1(\mathbf{x};\boldsymbol{\gamma}_1^*)\}f(\mathbf{x}, y_1 \mid R=1)\mathrm{d}y_1\mathrm{d}\mathbf{x} + \int \mu_1(\mathbf{x};\boldsymbol{\gamma}_1^*)f(\mathbf{x}\mid R=1)\mathrm{d}\mathbf{x} \\ &= \int y_1 f(y_1 \mid R=1)\mathrm{d}y_1 = \tau_1.\end{aligned}$$

Similarly, we have $\widehat{\tau}_0 \xrightarrow{p} \tau_0^*$, where

$$\tau_0^* = \frac{\mathbb{E}\left[\frac{1-T}{(1-\pi)}\rho(\mathbf{X}, Y_0;\boldsymbol{\beta}^*)\{Y - \mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)\} + k(\mathbf{X};\boldsymbol{\alpha}^*)\mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)\right]}{\mathbb{E}\{k(\mathbf{X};\boldsymbol{\alpha}^*)\}}.$$

Suppose $k(\mathbf{x};\boldsymbol{\alpha}^*) = k(\mathbf{x})$ and $\rho(y_0, \mathbf{x};\boldsymbol{\beta}^*) = \rho(y_0, \mathbf{x})$, we have

$$\begin{aligned}\tau_0^* &= \mathbb{E}\left[\frac{1-T}{\kappa(1-\pi)}\rho(\mathbf{X}, Y_0)\{Y_0 - \mu_0^*(\mathbf{X})\} + \frac{k(\mathbf{X})}{\kappa}\mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)\right] \\ &= \iint\{y_0 - \mu_0(\mathbf{x};\boldsymbol{\gamma}_0^*)\}\frac{\mathrm{pr}(R=1\mid\mathbf{x}, y_0, T=0)}{\kappa}f(\mathbf{x}, y_0 \mid T=0)\mathrm{d}y_0\mathrm{d}\mathbf{x} \\ &\quad + \int \mu_0(\mathbf{x};\boldsymbol{\gamma}_0^*)\frac{\mathrm{pr}(R=1\mid\mathbf{x})}{\kappa}f(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \iint\{y_0 - \mu_0(\mathbf{x};\boldsymbol{\gamma}_0^*)\}f(\mathbf{x}, y_0 \mid R=1)\mathrm{d}y_0\mathrm{d}\mathbf{x} + \int \mu_0(\mathbf{x};\boldsymbol{\gamma}_0^*)f(\mathbf{x}\mid R=1)\mathrm{d}\mathbf{x} \\ &= \int y_0 f(y_0 \mid R=1)\mathrm{d}y_0 = \tau_0.\end{aligned}$$

This completes the proof. □

G Proof of Theorem 2

Proof.

$$\widehat{\tau}_1 = \frac{1}{N}\sum_{i=1}^N\left[\frac{r_t t_i}{\widehat{\pi}}\{y_i - \mu_1(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_1)\} + k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}})\mu_1(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_1)\right] \Big/ \frac{1}{N}\sum_{i=1}^N k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}}),$$

$$\widehat{\tau}_0 = \frac{1}{N}\sum_{i=1}^N\left[\frac{1-t_i}{1-\widehat{\pi}}\rho(y_i, \mathbf{x}_i;\widehat{\boldsymbol{\beta}})\{y_i - \mu_0(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_0)\} + k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}})\mu_0(\mathbf{x}_i;\widehat{\boldsymbol{\gamma}}_0)\right] \Big/ \frac{1}{N}\sum_{i=1}^N k(\mathbf{x}_i;\widehat{\boldsymbol{\alpha}}),$$

By Theorem 1, we have $\widehat{\tau}_1 \xrightarrow{p} \tau_1$ and $\widehat{\tau}_0 \xrightarrow{p} \tau_0$. Let $\mathbb{P}_N Z = N^{-1}\sum_{i=1}^N z_i$ denote the sample average of any random variable Z. Noting that $\widehat{\tau}_1$ solves that

$$\mathbb{P}_N\{\psi_1(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_1, \widehat{\pi}, \tau_1)\} = 0.$$

The standard Taylor expansion yields that

$$\begin{aligned}0 &= \mathbb{P}_N\{\psi_1(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_1, \widehat{\pi}, \widehat{\tau}_1)\} \\ &= \mathbb{P}_N\{\psi_1(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}_1^*, \pi, \tau_1)\} + \mathbb{P}_N\left\{\frac{\partial}{\partial\tau_1^{\mathrm{T}}}\psi_1(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\gamma}}_1, \bar{\pi}, \bar{\tau}_1)\right\}(\widehat{\tau}_1 - \tau_1) \\ &\quad + \mathbb{P}_N\left\{\frac{\partial}{\partial\pi^{\mathrm{T}}}\psi_1(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\gamma}}_1, \bar{\pi}, \bar{\tau}_1)\right\}(\widehat{\pi} - \pi) + \mathbb{P}_N\left\{\frac{\partial}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\psi_1(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\gamma}}_1, \bar{\pi}, \bar{\tau}_1)\right\}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\end{aligned}$$

$$+\mathbb{P}_N\left\{\frac{\partial}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\psi_1\big(\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\gamma}}_1,\bar{\pi},\bar{\tau}_1\big)\right\}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_1^*).$$

where $(\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\gamma}},\bar{\pi},\bar{\tau}_1)$ lie between $(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}},\widehat{\pi},\widehat{\tau}_1)$ and $(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1)$. Then following uniform weak law of large number (Newey & McFadden, 1994) under the regularity conditions, we have

$$\begin{aligned}\sqrt{N}(\widehat{\tau}_1-\tau_1)=&-\left[\mathbb{E}\left\{\frac{\partial}{\partial\tau_1^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}\right]^{-1}\\&\times\left[\frac{1}{\sqrt{N}}\sum_{i=1}^N\psi_{1,i}+\mathbb{E}\left\{\frac{\partial}{\partial\pi^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}\sqrt{N}(\widehat{\pi}-\pi)\right.\\&\quad+\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}\sqrt{N}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\\&\quad\left.+\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}\sqrt{N}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_1^*)\right]+o_p(1).\end{aligned}$$

Note that

$$\mathbb{E}\left\{\frac{\partial}{\partial\tau_1^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}=-\mathbb{E}\left\{\frac{k(\mathbf{X})}{\kappa}\right\}=-1,$$

$$\mathbb{E}\left\{\frac{\partial}{\partial\pi^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}=-\mathbb{E}\left[\frac{RT}{\kappa\pi^2}\{Y-\mu_1(\mathbf{X})\}\right]=0,$$

$$\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}=\mathbb{E}\left[\frac{\{\mu_1(\mathbf{X})-\tau_1\}}{\kappa}\frac{\partial k(\mathbf{X};\boldsymbol{\alpha}^*)}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right],$$

and

$$\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\gamma}_1^{\mathrm{T}}}\psi_1\big(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}_1^*,\pi,\tau_1\big)\right\}=\mathbb{E}\left[\left\{-\frac{RT}{\kappa\pi}+\frac{k(\mathbf{X})}{\kappa}\right\}\frac{\partial\mu_1(\mathbf{X})}{\partial{\boldsymbol{\gamma}_1}^{\mathrm{T}}}\right]=0.$$

Therefore, we have $\sqrt{N}(\widehat{\tau}_1-\tau_1)=N^{-1/2}\sum_{i=1}^N\psi_{1,i}+\mathcal{R}_1+o_p(1)$, where

$$\mathcal{R}_1=\mathbb{E}\left[\frac{\{\mu_1(\mathbf{X})-\tau_1\}}{\kappa}\frac{\partial k(\mathbf{X};\boldsymbol{\alpha}^*)}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right]\cdot\sqrt{N}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*).$$

Similarly, $\widehat{\tau}_0$ solves that

$$\mathbb{P}_N\left\{\psi_0(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}_0,\widehat{\pi},\tau_0)\right\}=0,$$

and we have

$$\begin{aligned}\sqrt{N}(\widehat{\tau}_0-\tau_0)=&-\left[\mathbb{E}\left\{\frac{\partial}{\partial\tau_0^{\mathrm{T}}}\psi_0\big(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0\big)\right\}\right]^{-1}\times\left[\frac{1}{\sqrt{N}}\sum_{i=1}^N\psi_{0,i}\right.\\&+\mathbb{E}\left\{\frac{\partial}{\partial\pi^{\mathrm{T}}}\psi_0\big(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0\big)\right\}\sqrt{N}(\widehat{\pi}-\pi)+\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\psi_0\big(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0\big)\right\}\sqrt{N}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\\&\left.+\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\psi_0\big(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0\big)\right\}\sqrt{N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\psi_0\big(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0\big)\right\}\sqrt{N}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0^*)\right]\\&+o_p(1).\end{aligned}$$

Note that

$$\mathbb{E}\left\{\frac{\partial}{\partial\tau_0^{\mathrm{T}}}\psi_0\big(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0\big)\right\}=-\mathbb{E}\left\{\frac{k(\mathbf{X})}{\kappa}\right\}=-1,$$

$$\mathbb{E}\left\{\frac{\partial}{\partial\pi^{\mathrm{T}}}\psi_0\big(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0\big)\right\}=\mathbb{E}\left[\frac{1-T}{\kappa(1-\pi)^2}\rho(\mathbf{X},Y)\{Y-\mu_0(\mathbf{X})\}\right]=0,$$

$$\mathbb{E}\left\{\frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\psi_0(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0)\right\}=\mathbb{E}\left[\frac{\{\mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)-\tau_0\}}{\kappa}\frac{\partial k(\mathbf{X};\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right],$$

$$\mathbb{E}\left\{\frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\psi_0(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0)\right\}=\mathbb{E}\left[\frac{1-T}{\kappa(1-\pi)}\{Y_0-\mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)\}\frac{\partial \rho(\mathbf{X},Y_0;\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right],$$

and

$$\mathbb{E}\left\{\frac{\partial}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}\psi_0(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\boldsymbol{\gamma}_0^*,\pi,\tau_0)\right\}=\mathbb{E}\left[\left\{-\frac{1-T}{\kappa(1-\pi)}\rho(\mathbf{X},Y)+\frac{k(\mathbf{X})}{\kappa}\right\}\frac{\partial \mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)}{\partial {\boldsymbol{\gamma}_0}^{\mathrm{T}}}\right]=0,$$

where the last equality is because

$$\begin{aligned}
&\mathbb{E}\left\{\frac{1-T}{\kappa(1-\pi)}\rho(\mathbf{X},Y)\mid \mathbf{x}\right\}=\int\frac{\mathrm{pr}(R=1\mid \mathbf{x},y_0,T=0)}{\kappa(1-\pi)}f(T=0,y_0\mid \mathbf{x})dy_0\\
&=\int\frac{\mathrm{pr}(R=1\mid \mathbf{x})\,\mathrm{pr}(T=0\mid R=1,\mathbf{x})f(y_0\mid R=1,T=0,\mathbf{x})}{\kappa(1-\pi)}dy_0\\
&=\frac{\mathrm{pr}(R=1\mid \mathbf{x})}{\kappa}=\frac{k(\mathbf{x})}{\kappa}.
\end{aligned}$$

Therefore, we have $\sqrt{N}(\widehat{\tau}_0-\tau_0)=N^{-1/2}\sum_{i=1}^N\psi_{0,i}+\mathcal{R}_2+\mathcal{R}_3+o_p(1)$, where

$$\begin{aligned}
\mathcal{R}_2&=\mathbb{E}\left[\frac{\{\mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)-\tau_0\}}{\kappa}\frac{\partial k(\mathbf{X};\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right]\cdot\sqrt{N}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*),\\
\mathcal{R}_3&=\mathbb{E}\left[\frac{1-T}{\kappa(1-\pi)}\{Y_0-\mu_0(\mathbf{X};\boldsymbol{\gamma}_0^*)\}\frac{\partial \rho(\mathbf{X},Y_0;\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right]\cdot\sqrt{N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*).
\end{aligned}$$

□

H Proof of Theorem 3

Proof. Suppose $\widetilde{\tau}_t$ and $\widehat{\tau}_t$ are $\sqrt{N}$-consistent in the sense that

$$\sqrt{N}\begin{pmatrix}\widetilde{\tau}_t-\tau_t\\ \widehat{\tau}_t-\tau_t^*\end{pmatrix}\xrightarrow{d}\mathbf{Z},\ \text{ with }\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}),\ \text{and }\boldsymbol{\Sigma}=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{12}&\Sigma_{22}\end{pmatrix},$$

where τ_t^* is the probability limit of $\widehat{\tau}_t$, which may or may not equal τ_t. Let $\widehat{\boldsymbol{\Sigma}}$ be the standard sandwich estimator of $\boldsymbol{\Sigma}$ such that $\widehat{\boldsymbol{\Sigma}}\xrightarrow{p}\boldsymbol{\Sigma}$ (White, 1982; Newey & McFadden, 1994). Then we have

$$\begin{aligned}
&\widehat{\mathrm{var}}(\widetilde{\tau}_t)=\widehat{\Sigma}_{11}/N,\ \widehat{\mathrm{var}}(\widehat{\tau}_t)=\widehat{\Sigma}_{22}/N,\ \widehat{\mathrm{cov}}(\widetilde{\tau}_t,\widehat{\tau}_t)=\widehat{\Sigma}_{12}/N,\\
&\text{and }\ \widehat{\mathrm{var}}(\widehat{\tau}_t-\widetilde{\tau}_t)=(\widehat{\Sigma}_{11}+\widehat{\Sigma}_{22}-2\widehat{\Sigma}_{12})/N.
\end{aligned}$$

Then, we can estimate λ_t^* and $\delta_{t,N}$ via

$$\widehat{\lambda}_t^*=\frac{\widehat{\mathrm{var}}(\widetilde{\tau}_t)-\widehat{\mathrm{cov}}(\widehat{\tau}_t,\widetilde{\tau}_t)}{\widehat{\mathrm{var}}(\widehat{\tau}_t-\widetilde{\tau}_t)},$$

and, writing $\widehat{V}_{t,N}=\widehat{\mathrm{var}}(\widehat{\tau}_t-\widetilde{\tau}_t)=(\widehat{\Sigma}_{11}+\widehat{\Sigma}_{22}-2\widehat{\Sigma}_{12})/N$,

$$\widehat{\delta}_{t,N}=\frac{\widehat{V}_{t,N}^2}{\widehat{V}_{t,N}^2+(\widehat{\tau}_t-\widetilde{\tau}_t)^4/N^\gamma}=\left\{1+N^{-\gamma}\left(\frac{\widehat{\tau}_t-\widetilde{\tau}_t}{\sqrt{\widehat{V}_{t,N}}}\right)^4\right\}^{-1}.$$

The continuous mapping theorem implies that $\widehat{\lambda}_t^*\xrightarrow{p}\lambda_t^*$. Note that $N\widehat{V}_{t,N}\xrightarrow{p}\Sigma_{11}+\Sigma_{22}-2\Sigma_{12}$ and hence $\widehat{V}_{t,N}=O_p(N^{-1})$. If $\widehat{\tau}_t\xrightarrow{p}\tau_t$, then $\widehat{\tau}_t-\widetilde{\tau}_t=O_p(N^{-1/2})$ and

$$1-\widehat{\delta}_{t,N}=\frac{(\widehat{\tau}_t-\widetilde{\tau}_t)^4/N^\gamma}{\widehat{V}_{t,N}^2+(\widehat{\tau}_t-\widetilde{\tau}_t)^4/N^\gamma}\le\frac{(\widehat{\tau}_t-\widetilde{\tau}_t)^4}{N^\gamma\widehat{V}_{t,N}^2}=\frac{O_p(N^{-2})}{O_p(N^{\gamma-2})}=O_p(N^{-\gamma}),$$

so that $\widehat{\delta}_{t,N} = 1 + O_p(N^{-\gamma})$. If instead $\widehat{\tau}_t - \widetilde{\tau}_t \xrightarrow{p} c_t^* := \tau_t^* - \tau_t \neq 0$, then $(\widehat{\tau}_t - \widetilde{\tau}_t)^{-4} = O_p(1)$ and

$$\widehat{\delta}_{t,N} = \frac{\widehat{V}_{t,N}^2}{\widehat{V}_{t,N}^2 + (\widehat{\tau}_t - \widetilde{\tau}_t)^4 / N^\gamma} \leq \frac{N^\gamma \widehat{V}_{t,N}^2}{(\widehat{\tau}_t - \widetilde{\tau}_t)^4} = \frac{O_p(N^{\gamma-2})}{O_p(1)} = O_p(N^{\gamma-2}).$$

Let $\widehat{\lambda}_{t,N} = \widehat{\delta}_{t,N} \widehat{\lambda}_t^*$. Then, $\widehat{\lambda}_{t,N} \xrightarrow{p} \lambda_t^*$ if $\widehat{\tau}_t \xrightarrow{p} \tau_t$, and $\widehat{\lambda}_{t,N} = O_p(N^{\gamma-2})$ otherwise. Recall that $\gamma \in (0, 3/2)$ and $\lambda_t^\dagger = \lambda_t^* \, \mathbb{1}(\tau_t^* = \tau_t)$. Thus, in both cases, we have $\widehat{\lambda}_{t,N} \xrightarrow{p} \lambda_t^\dagger$.

Write $\widehat{\tau}_t^{(s)} = \boldsymbol{c}(\widehat{\lambda}_{t,N})^{\mathrm{T}} (\widetilde{\tau}_t, \widehat{\tau}_t)^{\mathrm{T}}$. If $\widehat{\tau}_t$ is consistent, then $\tau_t^* = \tau_t$ and $\boldsymbol{c}(\lambda)^{\mathrm{T}} (\tau_t, \tau_t^*)^{\mathrm{T}} = \tau_t$ for every λ, so

$$\sqrt{N}\{\widehat{\tau}_t^{(s)} - \tau_t\} = \boldsymbol{c}(\widehat{\lambda}_{t,N})^{\mathrm{T}} \sqrt{N}\{(\widetilde{\tau}_t - \tau_t), (\widehat{\tau}_t - \tau_t^*)\}^{\mathrm{T}} \xrightarrow{d} \boldsymbol{c}(\lambda_t^\dagger)^{\mathrm{T}} \mathbf{Z}$$

by Slutsky's theorem, which is the stated limit. If $\widehat{\tau}_t$ is inconsistent, write $c_t^* = \tau_t^* - \tau_t \neq 0$. Then

$$\begin{aligned}
\sqrt{N}\{\widehat{\tau}_t^{(s)} - \tau_t\} &= \sqrt{N}\{\widetilde{\tau}_t - \tau_t\} + \widehat{\lambda}_{t,N} \sqrt{N}\{\widehat{\tau}_t - \widetilde{\tau}_t\} \\
&= \sqrt{N}\{\widetilde{\tau}_t - \tau_t\} + \widehat{\lambda}_{t,N} \left[\sqrt{N}\{(\widehat{\tau}_t - \tau_t^*) - (\widetilde{\tau}_t - \tau_t)\} + \sqrt{N}\, c_t^* \right] \\
&= \sqrt{N}\{\widetilde{\tau}_t - \tau_t\} + O_p(N^{\gamma-2}) O_p(1) + O_p(N^{\gamma-3/2}) \\
&= \sqrt{N}\{\widetilde{\tau}_t - \tau_t\} + o_p(1) \\
&\xrightarrow{d} \mathcal{N}(0, \Sigma_{11}),
\end{aligned}$$

where the third equality uses $\widehat{\lambda}_{t,N} = O_p(N^{\gamma-2})$ and the fourth uses $\gamma < 3/2$. Since $\lambda_t^\dagger = 0$ here, $\Sigma_{11} = v_t$ and the stated limit again holds. Finally $\widehat{v}_t \xrightarrow{p} v_t$ by the continuous mapping theorem, since $\widehat{\boldsymbol{\Sigma}}_t \xrightarrow{p} \boldsymbol{\Sigma}_t$ and $\widehat{\lambda}_{t,N} \xrightarrow{p} \lambda_t^\dagger$. This completes the proof. □