# Simple and Reliable Inference for Matching Estimators: A Single-Matching Variance Approach

Xiang Meng[a,b,*], Aaron Smith[c], Luke Miratrix[d]

[a] *Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA*
[b] *Harvard T.H. Chan School of Public Health, Boston, MA, USA*
[c] *Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada*
[d] *Harvard Graduate School of Education, Harvard University, Cambridge, MA, USA*

**Abstract**

Matching estimators are widely used for causal inference, but variance estimation is difficult because reusing the same controls across matched sets induces dependence that standard resampling often ignores. We develop reuse-aware inference for a broad class of matching estimators with design-based weights, including nearest-neighbor, propensity-score, radius/caliper, and synthetic-control-style matching. Our main result reduces asymptotic normality to a simple weight-regularity condition: no control may be reused so heavily that it dominates the sampling noise. We then propose a single-matching variance estimator that uses only the treated-to-control matching already formed for the point estimate. Unlike existing analytic variance estimators, it requires no additional within-treatment-group matching, which can be computationally costly; it also accommodates non-uniform weights. We also give a valid bootstrap procedure based on the same fixed-matching principle. Simulations show that, once matching bias is handled, reuse-aware methods achieve near-nominal coverage, while reuse-blind bootstraps can severely undercover. An application to Brazilian education data illustrates the method.

*Corresponding author.
*Email addresses:* xmeng@g.harvard.edu (Xiang Meng), asmi28@uottawa.ca (Aaron Smith), luke_miratrix@gse.harvard.edu (Luke Miratrix)

*Keywords:* Matching estimators, Variance estimation, Causal inference, Bootstrap, Central limit theorem, Control reuse

## 1. Introduction

In observational studies, treatment is not randomly assigned, so treated and untreated units may differ systematically before treatment. Matching seeks to make the comparison more credible by pairing each treated unit with similar untreated units based on observed covariates (Rosenbaum and Rubin, 1983; Rubin, 1973). Valid population inference—the ability to generalize findings beyond the specific sample to the broader population—is crucial for policy decisions and scientific understanding across diverse fields including economics (Dehejia and Wahba, 1999; Heckman et al., 1997), epidemiology (Stuart, 2010), and policy evaluation (Smith and Todd, 2005).

Matching with replacement allows the same control unit to be used for more than one treated unit. The resulting comparisons are therefore not independent: they share the same control outcome. Ignoring this dependence can understate uncertainty and produce confidence intervals that are too narrow, especially when suitable controls are scarce and the same controls are used repeatedly (Abadie and Imbens, 2008).

The foundational theory of Abadie and Imbens (2006) showed that fixed-neighbor matching can have non-negligible bias and developed an analytic variance estimator that accounts for control reuse. Subsequent work established the failure of the ordinary bootstrap (Abadie and Imbens, 2008), developed bias-corrected estimators (Abadie and Imbens, 2011), provided alternative representations of the limiting distribution (Abadie and Imbens, 2012), and proposed a valid wild bootstrap for regression-adjusted matching estimators (Otsu and Rai, 2017). Related work has examined interval estimation for matching with replacement, sandwich estimators for survival outcomes, and the finite-sample performance of alternative inference procedures (Hill and Reiter, 2006; Austin and Cafri, 2020; Bodory et al., 2020).

Closest in principle, Abadie and Spiess (2022) study regression after

matching. Their procedure and ours share the principle that inference must account for the dependence created by the matching structure. Their variance estimates are constructed from residuals from a post-matching regression. Our approach instead estimates the outcome-noise scale from the dispersion of control outcomes within the matched sets already formed for the treatment-effect estimate; it therefore requires no fitted outcome regression.

This paper makes two methodological contributions and provides supporting finite-sample evidence.

*A single-matching variance estimator.* We propose a variance estimator that uses only the original treated-to-control matching. We call it *single-matching* because it requires no additional treated-to-treated or control-to-control matching solely for variance estimation. The estimator accounts for the dependence created when the same control contributes to several matched differences. It applies to general, possibly non-uniform matching weights and is consistent without requiring equal conditional variances. Relative to Abadie and Imbens (2006), it avoids additional same-group nearest-neighbor searches and is therefore substantially less costly to compute. We also develop a multiplier bootstrap based on the same fixed-matching principle and prove its validity.

*A unifying central limit theorem.* We establish asymptotic normality for matching estimators with general design-based weights, meaning weights determined by treatment assignments and covariates rather than outcomes. The result covers nearest-neighbor, propensity-score, radius/caliper, and synthetic-control-style matching. Its only matching-specific requirement is a weight-regularity condition: no single control may receive enough total weight to dominate the sampling variation. We verify this condition for standard matching schemes using existing bounds on control reuse (Abadie and Imbens, 2006, 2016; Kormos et al., 2023).

*Finite-sample evidence.* We compare six inference procedures in two data-generating settings designed to separate matching bias from variance error. The first uses nonlinear outcome surfaces, where matching bias must be removed before variance estimators can be assessed. The second uses a flat

outcome surface with zero matching bias and varies the availability of comparable controls, thereby isolating the effect of control reuse. Once matching bias is handled, our analytic estimator and covariance-corrected bootstrap achieve coverage near the nominal level, whereas the bootstrap that treats matched differences as independent can fall to 49% coverage under heavy reuse. An application to Brazilian education data illustrates implementation with radius matching and non-uniform synthetic-control weights.

The remainder of this paper is organized as follows. Section 2 introduces the matching estimator and summarizes how heavily each control is used. Section 3 develops the bias–variance decomposition and the central limit theorem, and records when the matching bias is negligible without regression adjustment. Section 4 introduces our single-matching variance estimator and proves consistency. Section 5 develops the covariance-corrected multiplier bootstrap. Section 6 compares six inference procedures across two data-generating designs. Section 7 applies our method to Brazilian education data. Section 8 concludes. Proofs and technical details are presented in the appendix. All matching, variance estimation, and bootstrap procedures used in Sections 6 and 7 are implemented in the `R` package `CSM`.[1]

## 2. Problem Setup and The Reuse-Weight Decomposition

We consider a setting with $n$ observations, each representing a unit in our study population. The sample consists of $n_T$ treated units and $n_C$ control units, with $n = n_T + n_C$.

For each unit $i$, we observe a tuple $\{Z_i, Y_i, \mathbf{X}_i\}$ where:

- $Z_i \in \{0, 1\}$ denotes its binary treatment status.
- $Y_i \in \mathbb{R}$ denotes its observed real-valued outcome.
- $\mathbf{X}_i \equiv \{X_{1i}, \ldots, X_{ki}\}^T \in \mathbb{R}^k$ denotes its $k$-dimensional real-valued covariate vector.

[1] Available at https://github.com/jche/CSMatch.

We adopt the potential outcomes framework where each unit has two potential outcomes: $Y_i(1)$ and $Y_i(0)$. Here, $Y_i(1)$ represents the outcome if unit $i$ receives treatment, and $Y_i(0)$ represents the outcome if unit $i$ does not receive treatment. The fundamental problem of causal inference is that we only observe one of these potential outcomes for each unit. Specifically, the observed outcome for unit $i$ is

$$Y_i \equiv (1 - Z_i)Y_i(0) + Z_i Y_i(1),$$

under the stable unit treatment value assumption (SUTVA).

We assume the data consist of i.i.d. draws of tuples $(Y_i(0), Y_i(1), Z_i, \mathbf{X}_i)$ from a common distribution that does not depend on the sample size $n$. For each unit $i$, the generic random variables $(Y(0), Y(1), Z, \mathbf{X})$ represent the population distribution from which the observed data are drawn. Throughout the paper, indexed variables, such as $\mathbf{X}_i$, refer to specific observations, while non-indexed variables, such as $\mathbf{X}$, refer to the corresponding generic random variables.

Our estimand of interest is the average treatment effect on the treated (ATT):

$$\tau = E\big[Y_i(1) - Y_i(0) \mid Z_i = 1\big].$$

### *2.1. Matching Estimator*

We write the set of all treated units' indices as $\mathcal{T} = \{i : Z_i = 1\}$, the set of all control units' indices as $\mathcal{C} = \{i : Z_i = 0\}$, and $t \in \mathcal{T}$, $j \in \mathcal{C}$ as individual treated and control units respectively. For each treated unit $t \in \mathcal{T}$, let $\mathcal{C}_t \subseteq \mathcal{C}$ denote an arbitrary set of control units assigned as its matches; the collection $\{\mathcal{C}_t : t \in \mathcal{T}\}$ is then called a matching. Finally, we denote the size of a set $\mathcal{S}$ as $|\mathcal{S}|$. The matching estimator of the ATT takes the form

$$\hat{\tau}(w) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \big(Y_t - \sum_{j \in \mathcal{C}_t} w_{jt} Y_j\big), \tag{1}$$

where $w_{jt} \in [0,1]$ is the weight assigned to the matched control unit $j$ for treated unit $t$, with $\sum_{j \in \mathcal{C}_t} w_{jt} = 1$ for each $t \in \mathcal{T}$. This formulation encompasses many common procedures: for instance, in $M$-nearest neighbor matching (Rubin, 1973; Abadie and Imbens, 2006; Stuart, 2010), each $\mathcal{C}_t$ consists of the $M$ nearest controls to $t$ with equal weights $w_{jt} = 1/M$, while in synthetic-control-style matching Che et al. (2024), $w_{jt}$ is chosen by solving an optimization problem to approximate $\mathbf{X}_t$ by a convex combination of $\{\mathbf{X}_j : j \in \mathcal{C}_t\}$.

Because a single control unit may be matched to several treated units, control outcomes are *reused* across matched sets. This reuse is the source of the cross-unit dependence that makes inference for matching nonstandard, so it is convenient to record, in a single per-control quantity, how heavily each control is used.

**Definition 2.1** (Reuse weight)**.** *The* reuse weight *of control $j$ is*

$$w_j \;=\; \sum_{t \in \mathcal{T}} w_{jt},$$

*the total weight placed on unit $j$ across all treated units.*

The *weight matrix* $[w_{jt}]_{j \in \mathcal{C}, t \in \mathcal{T}}$ records the full bipartite matching structure. Summing over treated units gives $w_j$; summing over controls gives 1 (each treated unit's weights sum to 1). For $M$-nearest-neighbor matching, $w_j = K_M(j)/M$ where $K_M(j)$ is the number of treated units that list $j$ among their $M$ nearest neighbors.

### *2.2. The matching index and the decomposition of the estimation error*

Matching schemes differ in what they match on, and our theory is most transparent when written in terms of that object rather than the full covariate. Let $S_i = s(\mathbf{X}_i)$ denote the matching index used to form the matched sets: $S_i = \mathbf{X}_i$ for nearest-neighbor matching on the covariates (so $\dim S = k$), and $S_i = e(\mathbf{X}_i)$ or $S_i = \hat{e}(\mathbf{X}_i)$ for matching on the true or estimated propensity score and for radius/caliper matching (so $\dim S = 1$). We require

the matching index $S = s(\mathbf{X})$ to be a valid balancing score (Rosenbaum and Rubin, 1983):

$$Z \perp\!\!\!\perp \mathbf{X} \mid S.$$

Together with unconfoundedness, $\{Y(0), Y(1)\} \perp\!\!\!\perp Z \mid \mathbf{X}$, this implies

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

(Rosenbaum and Rubin, 1983). Hence, for $z \in \{0, 1\}$,

$$\mu_z(s) := \mathbb{E}[Y(z) \mid S = s] = \mathbb{E}[Y \mid Z = z, S = s].$$

Define the conditional treatment effect projected onto the matching index by

$$\tau_S(s) := \mu_1(s) - \mu_0(s) = \mathbb{E}[Y(1) - Y(0) \mid S = s].$$

The ATT therefore satisfies

$$\tau = \mathbb{E}[\tau_S(S) \mid Z = 1].$$

Define the score-centered residuals $\varepsilon_{z,i} = Y_i(z) - \mu_z(S_i), z \in \{0, 1\}$. By construction, $\mathbb{E}[\varepsilon_{z,i} \mid S_i] = 0$, and, by consistency, for a treated unit, $Y_t = \mu_0(S_t) + \tau_S(S_t) + \varepsilon_{1,t}$, whereas for a control unit, $Y_j = \mu_0(S_j) + \varepsilon_{0,j}$.

With this notation, the estimation error decomposes into three components,

$$\hat{\tau}(w) - \tau = \big(\hat{\tau}(w) - \bar{\tau}\big) + \big(\bar{\tau} - \tau\big) = B_n + E_n + P_n, \tag{2}$$

where $\bar{\tau} = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \tau_S(S_t)$ is the sample average conditional effect among

the treated. The three terms are

$$B_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{C}_t} w_{jt} \big(\mu_0(S_t) - \mu_0(S_j)\big), \quad \text{the bias from imperfect matching on the index;} \tag{3}$$

$$E_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \varepsilon_{1,t} - \frac{1}{n_T} \sum_{j \in \mathcal{C}} w_j \, \varepsilon_{0,j}, \quad \text{the measurement error from random outcome noise;} \tag{4}$$

$$P_n = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \big\{\tau_S(S_t) - \tau\big\}, \quad \text{the representation error between sample and population effec} \tag{5}$$

where $w_j = \sum_{t \in \mathcal{T}} w_{jt}$ is the total weight assigned to control unit $j$ across all matched treated units.

**Assumption 1** (Regular outcome regression)**.** *The control regression*

$$\mu_0(s) = \mathbb{E}[Y(0) \mid S = s]$$

*is Lipschitz continuous on the compact support $\mathcal{S}$ of $S$: there exists $L_\mu < \infty$ such that*

$$|\mu_0(s) - \mu_0(s')| \le L_\mu \|s - s'\| \qquad \textit{for all } s, s' \in \mathcal{S}.$$

This condition directly controls the imperfect-matching bias:

$$|B_n| \le L_\mu \max_{t \in \mathcal{T}} \max_{j \in \mathcal{C}_t} \|S_t - S_j\|.$$

Smooth primitive outcome models together with regularity of the conditional distribution of $\mathbf{X}$ given $S$ provide sufficient conditions for this assumption; see, for example, the admissible model class of Kormos et al. (2023).

*2.2.1. Error Variance Assumptions*

Throughout the variance analysis, conditional variances are taken given the matching index of Section 2.2: $\sigma_z^2(s) = \mathrm{Var}(Y(z) \mid Z = z, S = s)$ for $z \in \{0,1\}$, with $\sigma_{1,t}^2 = \sigma_1^2(S_t)$ and $\sigma_{0,j}^2 = \sigma_0^2(S_j)$, and the relevant noises are the index-centered $\varepsilon_{1,t}, \varepsilon_{0,j}$ defined there. For propensity-score matching this conditions on the score, not on the full covariate $\mathbf{X}$, unless $Y(z)$ depends on $\mathbf{X}$ only through the score.

**Definition 2.2** (Regular variance function)**.** *Let $\mathcal{S}$ denote the (compact) support of the matching index $S$. A function $\sigma^2 : \mathcal{S} \to \mathbb{R}_+$ is a* regular variance function *if*

- ***Uniform continuity.*** *$\sigma^2(\cdot)$ is uniformly continuous on $\mathcal{S}$;*
- ***Boundedness.*** *there exist $0 < \sigma_{\min}^2 < \sigma_{\max}^2 < \infty$ with $\sigma_{\min}^2 \le \sigma^2(s) \le \sigma_{\max}^2$ for all $s \in \mathcal{S}$.*

We can now state the variance assumptions.

**Assumption 2** (Regular error variances)**.** *Both $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$ are regular variance functions on $\mathcal{S}$.*

**Assumption 3** (Error moments)**.** *There exist $\delta > 0$ and $C < \infty$ such that, for $z \in \{0,1\}$,*

$$\sup_{s \in \mathcal{S}} \mathbb{E}\big[\, |\varepsilon_{z,i}|^{2+\delta} \,\big|\, S_i = s,\, Z_i = z \big] \;\le\; C.$$

The continuity condition ensures that matched units have similar variances: whenever the matched discrepancy in the index vanishes, $\max_{j \in \mathcal{C}_t} \|S_j - S_t\| \to 0$—guaranteed for nearest-neighbor matching by standard matching-discrepancy bounds (Abadie and Imbens, 2006, Lemmas 1–2) and for radius/caliper matching by the vanishing caliper $\delta_n \to 0$—we have $\max_{j \in \mathcal{C}_t} |\sigma_0^2(S_j) - \sigma_0^2(S_t)| \to 0$, which is exactly what the within-cluster variance proxy relies on. Definition 2.2 relaxes the Lipschitz condition of Abadie and Imbens (2006, Assumption 4). Boundedness prevents degeneracy and controls the influence of outliers, and the moment bound of Assumption 3 is standard and enables the Lindeberg–Lyapunov argument of Theorem 3.2.

The asymptotic variance $V = n_T \cdot (V_E + V_P)$ decomposes into two components. Note that both $V_E$ and $V_P$ are defined at the $O(1/n_T)$ scale, so $V = O(1)$ provides the correct normalization for the $\sqrt{n_T}$-scaled CLT. Throughout, all conditional variances are taken given the matching index, as fixed in Section 2.2.1.

*Measurement Error Component $V_E$.* Conditional on the design, the variance from random outcome noise is

$$V_E = \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} \sigma_{1,t}^2 + \sum_{j\in\mathcal{C}} (w_j)^2 \sigma_{0,j}^2\right), \tag{6}$$

with $\sigma_{1,t}^2 = \sigma_1^2(S_t)$ and $\sigma_{0,j}^2 = \sigma_0^2(S_j)$. The first term reflects treated-unit outcome variance; the second captures control variance weighted by reuse. Heavily reused controls (large $w_j$) disproportionately affect total variance, creating a bias–variance tradeoff (Kallus, 2020; Che et al., 2024).

*Population Heterogeneity Component $V_P$.* Even with perfect matching, the sample ATT may deviate from the population ATT due to treatment-effect heterogeneity:

$$V_P = \frac{1}{n_T}\mathbb{E}\big[(\tau_S(S_i) - \tau)^2 \mid Z_i = 1\big]. \tag{7}$$

This component vanishes under homogeneous treatment effects and depends only on dispersion among treated units.

## 3. A Unified Lyapunov Central Limit Theorem for Matching

We now establish the asymptotic normality of $\hat{\tau}(w)$ that underlies valid confidence intervals. The theorem below introduces no new matching-specific probability device. Instead, it shows that a *single* operative condition—weight regularity of the control-reuse weights—governs the matching noise term simultaneously for the whole class of matching estimators, including nearest-neighbor matching on covariates or on the propensity score, and radius (caliper) matching. Asymptotic normality of the matching estimator

thereby reduces to the familiar Lyapunov requirement that no single reused control dominates the noise.

To proceed, we require a set of conditions on the covariates, treatment assignment, and matching procedure.

**Assumption 4** (Compact support)**.** *The covariate vector* $\mathbf{X}$ *is a* $k$*-dimensional random vector with a density with respect to Lebesgue measure on* $\mathbb{R}^k$ *with compact support* $\mathbb{X}$*. The density of* $X$ *is bounded and bounded away from zero on its support.*

The compact support assumption helps ensure that the covariate space is well-behaved, which facilitates consistent estimation and rules out pathological cases where the distribution of covariates becomes too sparse or unbounded.

**Assumption 5** (Unconfoundedness and overlap (Rubin, 1974))**.** *For almost every* $x \in \mathbb{X}$ *there exists* $\eta > 0$ *such that*

1. $(Y(1), Y(0)) \perp\!\!\!\perp Z \mid \mathbf{X}$,
2. $\eta < \Pr(Z = 1 \mid \mathbf{X} = x) < 1 - \eta$.

This assumption states that, conditional on the observed covariates, treatment assignment is independent of the potential outcomes, and that both treated and control units are sufficiently represented across the covariate space. By the law of large numbers, it follows that $n_T/n \to \Pr(Z = 1)$ and $n_C/n \to 1 - \Pr(Z = 1)$ almost surely, and hence $n_T/n_C \to \theta := \Pr(Z = 1)/(1 - \Pr(Z = 1)) \in \left(\frac{\eta}{1-\eta}, \frac{1-\eta}{\eta}\right)$.

These two assumptions, together with the design-side weight conditions of Section 3.1, are all the central limit theorem requires. The theorem is stated under the operative weight-regularity condition (Assumption 7); a stronger but directly checkable moment condition implying it is given in Lemma 3.1 and verified scheme by scheme in Section 3.2 (Proposition 3.3).

### *3.1. Weight regularity and the Central Limit Theorem*

Recall the decomposition $\hat{\tau} - \tau = B_n + E_n + P_n$ from Section 2.2. The heterogeneity term $P_n = \bar{\tau} - \tau$ of Equation (5) is an average of $n_T$ i.i.d.

treated contributions and obeys an ordinary central limit theorem with variance $n_T V_P$; the bias $B_n$ of Equation (3) is discussed in Remark 3.1 below. The only term whose limiting behavior is specific to matching is the noise term $E_n$ of Equation (4), with conditional variance $V_E$ (Equation (6)). Conditional on the design $(\mathbf{S}, \mathbf{Z})$ the summands of $E_n$ are independent and mean zero, so $E_n/\sqrt{V_E}$ is asymptotically normal by a Lyapunov argument as soon as no single control dominates the reuse. This is the content of the following two conditions, the first structural and the second the operative Lyapunov requirement.

**Assumption 6** (Design-measurable matching weights)**.** *For every treated unit $t$ and control $j$, the per-match weight $w_{jt}$* (i) *is a measurable function of the design $(\mathbf{S}, \mathbf{Z})$ alone and does not depend on the outcomes $\{Y_i\}$;* (ii) *is nonnegative with $\sum_{j \in \mathcal{C}} w_{jt} = 1$, so that $w_{jt} \in [0, 1]$.* [2]

Items (i)–(ii) are purely structural: the weights lie in $[0, 1]$, are computable from the design—covariates or scores, treatment indicators. Each treated unit distributes at most unit mass across controls. The substantive requirement is Assumption 7 below; Lemma 3.1 then supplies a stronger but directly checkable moment condition, which Proposition 3.3 discharges for nearest-neighbor and caliper matching.

**Assumption 7** (Weight regularity)**.** *The aggregate reuse weights satisfy*

$$\frac{\max_{j \in \mathcal{C}} w_j}{\sqrt{n_T + \sum_{j \in \mathcal{C}} w_j^2}} \xrightarrow{p} 0 \qquad \text{as } n_T \to \infty.$$

Weight regularity is precisely the Lindeberg/Lyapunov condition for $E_n$: it requires the most-reused control to be asymptotically negligible relative to the standard deviation of the noise.

Isolating this single condition is our main contribution to the matching central limit theory, and it is deliberately a modest, almost *tautological*

[2] Bounded mean reuse is automatic under Assumption 6: $\sum_j w_j \leq n_T$, and by exchangeability of the controls, $\sup_j \mathbb{E}[w_j] \leq \mathbb{E}[n_T/(n_C \vee 1)] = O(1)$ by inverse moments of binomial variates (Cribari-Neto et al., 2000).

one: nothing about the matching estimator requires more than the Lindeberg condition that any weighted average requires. Earlier treatments reach asymptotic normality through scheme-specific devices—reuse-count moment bounds for nearest-neighbor matching (Abadie and Imbens, 2006), a martingale representation (Abadie and Imbens, 2012), or match-count ratios for caliper matching (Kormos et al., 2023)—each of which *implies* weight regularity. By making the operative requirement transparent and showing it is the *only* matching-specific input, we unify these results rather than replace them, and we make clear that the matching estimator obeys the same central-limit logic as any other reused-weight average.

Assumption 7 is stated in the self-normalized form the proof uses, but it is rarely checked in that form. The next lemma converts it into a moment bound on the reuse weights—the form in which scheme-specific results are naturally cited. The allowance for a slowly diverging $c_n$ is deliberate: nearest-neighbor schemes satisfy the bound with $c_n = O(1)$, while caliper matching satisfies it only up to logarithmic factors (Proposition 3.3).

**Lemma 3.1** (A checkable moment condition)**.** *Suppose there exist $r > 2$ and a deterministic sequence $c_n$ with $n^{1-r/2}c_n \to 0$ such that $\max_{j\in\mathcal{C}} \mathbb{E}[w_j^r] \le c_n$. Then, under Assumption 5, Assumption 7 holds.*

*Proof.* Since $n_T/n \to \Pr(Z = 1) > 0$ a.s., the event $A_n = \{n_T \ge \kappa n\}$ has probability tending to one for small $\kappa > 0$, and the denominator in Assumption 7 is at least $\sqrt{n_T}$. For $\varepsilon > 0$, a union bound and Markov's inequality give $\Pr\big(\{\max_j w_j > \varepsilon\sqrt{n_T}\} \cap A_n\big) \le n\,\varepsilon^{-r}(\kappa n)^{-r/2}c_n \lesssim n^{1-r/2}c_n \to 0$. □

Verification of the moment condition for the standard matching schemes is carried out in Section 3.2; here we proceed directly to the main asymptotic-normality result.

We now state our main asymptotic-normality result.

**Theorem 3.2** (Central Limit Theorem)**.** *Under Assumptions 2, 4, 5, 6, and 7, as $n_T \to \infty$:*

$$\frac{\sqrt{n_T}\,\big(\hat{\tau} - B_n - \tau\big)}{V^{1/2}} \quad \xrightarrow{d} \quad N(0,1),$$

*where*

$$V = n_T \cdot (V_E + V_P).^{3}$$

Proof: See Appendix Appendix A.

This theorem both unifies and extends existing results. It subsumes the nearest-neighbor theory of Abadie and Imbens (2006), its estimated-propensity-score extension (Abadie and Imbens, 2016), and the caliper-matching result of Kormos et al. (2023), and it covers estimators outside that list: radius matching and adaptive within-caliper weights, such as the synthetic-control-style weights of caliper synthetic matching (Che et al., 2024), to which no existing matching CLT applies directly.

**Remark 3.1** (Matching bias)**.** *Our results take the imperfect-match bias $B_n$ as given. When matching is on the scalar (estimated) propensity score, the bias involves only one-dimensional score discrepancies and is asymptotically negligible: $\sqrt{n}\, B_n = o_p(1)$ for nearest-neighbor matching (Abadie and Imbens, 2006, Cor. 1), also with a first-step estimated score (Abadie and Imbens, 2016), and for caliper matching, where $|B_n| \lesssim \delta_n \asymp \log n/n$ under Lipschitz outcome regressions (Kormos et al., 2023, Thms. 1–4). When matching is on $k \geq 2$ covariates, $B_n = O_p(n^{-1/k})$ and can dominate the $\sqrt{n}$ approximation (Abadie and Imbens, 2006, Thm. 1), so our intervals should then be applied to a bias-corrected estimator Abadie and Imbens (2011); Lin and Han (2022). Since our contribution concerns inference rather than bias correction, our simulations subtract the bias term directly, isolating the performance of the proposed variance estimator. Variance estimation for regression-adjusted, bias-corrected matching estimators (Abadie and Imbens, 2011; Lin and Han, 2022) is deferred to companion work.*

**Remark 3.2** (Quantitative CLT / Berry–Esseen Rate)**.** *Theorem 3.2 establishes convergence in distribution but does not provide a finite-sample rate. For ATE estimation under a related stabilization framework, Shi et al.*

[3] Since both $V_E$ and $V_P$ are $O(1/n_T)$, we have $V = n_T \cdot (V_E + V_P) = O(1)$ and thus $V^{1/2} = O(1)$, yielding a properly standardized statistic.

*(2024) derive Gaussian and bootstrap approximations with Berry–Esseen-type bounds using Malliavin–Stein methods and geometric stabilization theory. Their results suggest that the Kolmogorov distance between the standardized statistic and $N(0,1)$ decays at rate $O(n^{-1/2})$ under analogous conditions. Adapting this approach to the ATT setting—where the treated-only summation and asymmetric control reuse introduce additional technicalities—is left for future work.*

### *3.2. Verification of weight regularity*

Weight regularity (Assumption 7) is the only matching-specific input to Theorem 3.2, and it is verified scheme by scheme through the moment condition of Lemma 3.1. Matching on the propensity score requires one further regularity condition on the score.

**Assumption 8** (Propensity-score regularity)**.** *When matching is on the propensity score $\pi(\mathbf{X})$ or an estimate thereof: conditional on $Z = z$, $\pi(\mathbf{X})$ admits a density $f_z$, and $f_0, f_1$ are continuous and bounded away from zero on their common compact support $[\underline{p}, \bar{p}] \subset (0,1)$;*

**Proposition 3.3** (Weight regularity of standard matching schemes)**.** *Suppose Assumptions 4 (with convex support), and 6 hold and units are sampled i.i.d.*

(a) (Nearest-neighbor on covariates) *If each treated unit is matched to its $M$ nearest controls in $\mathbf{X}$, then $\max_{j\in\mathcal{C}} \mathbb{E}[w_j^r] = O(1)$ for every $r > 2$.*

(b) (Nearest-neighbor on the (estimated) propensity score) *If matching is $M$-nearest-neighbor on $\pi(\mathbf{X}, \hat{\theta})$ and Assumption 8 holds, then $\max_{j\in\mathcal{C}} \mathbb{E}[w_j^r] = O(1)$ for every $r > 2$.*

(c) (Caliper/radius) *If matching is caliper matching on $\pi(\mathbf{X}, \hat{\theta})$ with caliper $\delta_n$ as in Kormos et al. (2023, eqs. (3) and (13)) and Assumption 8 holds, then for every fixed integer $r \geq 2$ and every $\varrho > 0$, $\max_{j\in\mathcal{C}} \mathbb{E}[w_j^r] = o(n^{\varrho})$.*

*In each case the moment condition of Lemma 3.1 is satisfied ($c_n = O(1)$ in (a)–(b); $c_n = o(n^{\varrho})$, in fact $c_n \lesssim (\log n)^r$, in (c)), hence Assumption 7 holds.*

The proof, in Appendix Appendix B, is by direct appeal to the reuse-count moment bounds of Abadie and Imbens (2006), Abadie and Imbens (2016), and Kormos et al. (2023): with $w_{jt} \in [0,1]$, the total reuse $w_j$ is dominated by the number of treated units matched to $j$, so the cited bounds transfer to any adaptive within-set weights. Notably, caliper matching satisfies the moment condition only in the polylogarithmic form—the number of matches per unit grows like $\log n$ (Kormos et al., 2023, Prop. 2)—which is exactly why Lemma 3.1 is stated with a diverging $c_n$ rather than $c_n = O(1)$.

## 4. The Single-Matching Variance Estimator

The central limit theorem of Section 3 gives the matching estimator $\hat{\tau}(w)$ an asymptotic variance $V = n_T(V_E + V_P)$—the sum of a measurement-error component $V_E$ and a heterogeneity component $V_P$. To turn the theorem into confidence intervals we need a consistent estimator of $V$. This section develops the single-matching estimator $\hat{V}$, which reads all of its variance proxies off the treated-to-control matching already formed for the point estimate, requiring no additional same-group matching and no outcome model.

We build it in two steps. Section 4.1 introduces the core engine of single matching: the within-cluster control variance $s_t^2$, which reads the outcome-noise scale directly off each matched set. As a first application we assemble these proxies into the measurement-error estimator $\hat{V}_E$, whose consistency on its own requires an equal-variance condition $\sigma_1^2 = \sigma_0^2$. Section 4.2 then assembles the total estimator $\hat{V}$, in which the equal-variance requirement disappears because the empirical variance of the matched differences absorbs the idiosyncratic treatment-effect variance directly; we prove its consistency there.

*4.1. Within-cluster variance proxies and the additive-noise model*

Recall from Equation (6) that the measurement-error variance is

$$V_E = \mathbb{E}[E_n^2 \mid \mathbf{S}, \mathbf{Z}] = \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} \sigma_{1,t}^2 + \sum_{j\in\mathcal{C}} (w_j)^2 \sigma_{0,j}^2\right),$$

with $w_j = \sum_{t\in\mathcal{T}} w_{jt}$. The only inputs are the per-unit conditional variances $\sigma_{1,t}^2$ and $\sigma_{0,j}^2$, which we estimate from the dispersion of control outcomes within each matched set—a proxy that requires no outcome model.

For each matched cluster $\{t\} \cup \mathcal{C}_t$ we use the within-cluster control variance

$$s_t^2 = \frac{1}{|\mathcal{C}_t| - 1} \sum_{j\in\mathcal{C}_t} \left(Y_j - \bar{Y}_t\right)^2, \quad \text{where} \quad \bar{Y}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j\in\mathcal{C}_t} Y_j. \tag{8}$$

A reused control $j$ may belong to several clusters; we assign it to one arbitrarily and write $s_j^2$ for the variance of its assigned cluster.[4][5] The proxy is model-free: it uses only control outcomes and centers each deviation at the matched-set average $\bar{Y}_t$ rather than at a fitted regression value, in the spirit of the within-matched-set variance estimators of Abadie and Spiess (2022) and Austin and Cafri (2020).

Aggregating these cluster variances over the matching, weighting each control by its squared total reuse $w_j^2$, gives the measurement-error estimator

$$\hat{V}_E = \frac{1}{n_T^2}\left(\sum_{t\in\mathcal{T}} s_t^2 + \sum_{j\in\mathcal{C}} w_j^2\, s_j^2\right). \tag{9}$$

Because $s_t^2$ and $s_j^2$ are both built from *control* dispersion, they estimate $\sigma_0^2$; they therefore reproduce the treated term $\sigma_{1,t}^2$ of $V_E$ only under the equal-variance condition $\sigma_1^2 = \sigma_0^2$. Under that condition $\hat{V}_E$ is consistent for $V_E$,

[4] Overlap does not affect the asymptotic theory, since the contribution of each $j$ is accounted for via its total weight $w_j = \sum_{t\in\mathcal{T}} w_{jt}$.

[5] One could instead average the variance estimates across all clusters containing $j$; both choices are asymptotically consistent, and the arbitrary assignment is simpler and performs identically in the limit.

as recorded next; the general additive-noise case, in which the treated and control variances differ by $\sigma_\tau^2$, is handled by the total estimator of Section 4.2, which does not require equal variances.

Beyond the weight conditions of Section 3.1, the consistency argument requires the matched-set discrepancy in the index to vanish,

$$\max_{t \in \mathcal{T}} \max_{j \in \mathcal{C}_t} \left\| S_j - S_t \right\| \xrightarrow{p} 0, \tag{10}$$

which was already established for the standard schemes in the commentary following Definition 2.2.

**Theorem 4.1** (Consistency of $\hat{V}_E$ under equal variances)**.** *Suppose Assumptions 2, 3, and 6 hold, the weights satisfy the moment condition of Lemma 3.1, Condition (10) holds, and $\sigma_1^2(s) = \sigma_0^2(s)$ for all $s \in \mathcal{S}$. Then the estimator $\hat{V}_E$ of Equation (9) satisfies $n_T |\hat{V}_E - V_E| \xrightarrow{p} 0$ as $n_T \to \infty$.*

Without equal variances, $\hat{V}_E$ instead targets $V_E^* = V_E - (1/n_T^2) \sum_t (\sigma_{1,t}^2 - \sigma_{0,t}^2)$ (Theorem Appendix C.1); this is exactly the discrepancy that the total estimator $\hat{V}$ corrects. The proof, including the treatment of matched sets containing a single control—for which $s_t^2$ is undefined—is in Appendix Appendix C. The estimator $\hat{V}_E$ also admits an illuminating pooled form: with the control effective sample size $\mathrm{ESS}(\mathcal{C}) = n_T^2 / \sum_{j \in \mathcal{C}} w_j^2$ (Potthoff et al., 2024), it equals a Welch-type two-sample $t$-test variance plus a heteroskedasticity adjustment that vanishes under uniform weights or homoskedasticity; we defer this form to Appendix Appendix F. For inference we never need $\hat{V}_E$ on its own—we use the total estimator $\hat{V}$ derived next.

### *4.2. The single-matching estimator $\hat{V}$ and its consistency*

For each treated unit $t$, let $\hat{Y}_t(0) = \sum_{j \in \mathcal{C}_t} w_{jt} Y_j$ denote the imputed counterfactual (the weighted average of matched control outcomes), and let $\hat{\Delta}_t = Y_t - \hat{Y}_t(0)$ be the matched difference. The squared deviation $(\hat{\Delta}_t - \tau)^2$ carries both treatment-effect heterogeneity and residual outcome noise. Averaging over the treated and taking expectations (neglecting $o(1/n_T)$ bias

terms from matching imperfections), a short calculation gives

$$\frac{1}{n_T}\sum_{t\in\mathcal{T}} E\Big[(Y_t(1)-\hat{Y}_t(0)-\tau)^2\Big] \approx n_T V_P + \frac{1}{n_T}\Big[\sum_{t\in\mathcal{T}}\sigma_{1,t}^2 + \sum_{j\in\mathcal{C}}\Big(\sum_{t'\in\mathcal{T}} w_{jt'}^2\Big)\sigma_{0,j}^2\Big]$$
$$= n_T V_P + n_T V_E - \frac{1}{n_T}\sum_{j\in\mathcal{C}}\sigma_{0,j}^2\Big[\Big(\sum_{t'\in\mathcal{T}} w_{jt'}\Big)^2 - \sum_{t'\in\mathcal{T}} w_{jt'}^2\Big], \tag{11}$$

where the second equality uses $w_j^2 = (\sum_{t'} w_{jt'})^2$ and adds and subtracts $\sum_j w_j^2\sigma_{0,j}^2$.

Equation (11) is the key to estimating the total variance without an equal-variance assumption. Its left-hand side is the population analog of the empirical matched-difference variance $\frac{1}{n_T}\sum_t(\hat{\Delta}_t-\hat{\tau})^2$, which is computed from the actual treated outcomes $Y_t$ and therefore carries the true treated variance $\sigma_{1,t}^2$ automatically. This means we never have to model $\sigma_{1,t}^2$, nor require equal variances as in proving consistency of $\hat{V}_E$ (Theorem 4.1). Rearranging (11) to solve for $n_T(V_E+V_P)=V$ and plugging the control proxy $s_j^2$ in for $\sigma_{0,j}^2$ in the remaining shared-control term results in the single-matching estimator

$$\hat{V} = \frac{1}{n_T}\sum_{t\in\mathcal{T}}\Big(Y_t-\hat{Y}_t(0)-\hat{\tau}\Big)^2 + \frac{1}{n_T}\sum_{j\in\mathcal{C}} s_j^2\left[\left(\sum_{t'\in\mathcal{T}} w_{jt'}\right)^2 - \left(\sum_{t'\in\mathcal{T}} w_{jt'}^2\right)\right]. \tag{12}$$

The first term captures empirical variation in matched differences across treated units. The second term of $\hat{V}$ in Equation (12) is a control-reuse covariance. It weights each control variance $s_j^2$ by

$$\Big(\textstyle\sum_t w_{jt}\Big)^2 - \sum_{t\in\mathcal{T}} w_{jt}^2 \;=\; 2\sum_{t_1<t_2} w_{jt_1}w_{jt_2} \;\geq\; 0.$$

This is precisely the covariance generated by sharing control $j$ across distinct matched differences: control $j$ contributes $-w_{jt}\varepsilon_j$ to each $\hat{\Delta}_t = Y_t - \hat{Y}_t(0)$ with $j\in\mathcal{C}_t$, so the single noise draw $\varepsilon_j$ drives several matched differences

and

$$\sum_{t\neq t'} \mathrm{Cov}\big(w_{jt}\varepsilon_j,\ w_{jt'}\varepsilon_j\big) \;=\; \sigma_{0,j}^2 \sum_{t\neq t'} w_{jt}w_{jt'} \;=\; \sigma_{0,j}^2\Big[\big(\textstyle\sum_t w_{jt}\big)^2 - \sum_t w_{jt}^2\Big],$$

which the second term of $\hat{V}$ estimates by plugging $s_j^2$ in for $\sigma_{0,j}^2$. It vanishes precisely when no control is reused, in which case $\hat{V}$ collapses to the pooled matched-difference variance.

Independent treated-side multipliers cannot reproduce this covariance; ignoring it produces systematic under-coverage whenever controls are substantially reused (Section 6).[6]

A conceptual point is worth emphasizing: the second term is not $w_j^2$, as one might expect from the variance formulas of regression-adjusted estimators. Because the empirical matched-difference term already contains each control's own-noise contribution $\sum_t w_{jt}^2$, the correction must supply only the covariance across distinct pairs of matched differences, $w_j^2 - \sum_t w_{jt}^2$. This yields an exact covariance decomposition rather than a conservative upper bound.

**Theorem 4.2** (Consistency of the Total Variance Estimator)**.** *Suppose Assumptions 1, 2, 3, and 6 hold, the reuse weights satisfy the moment condition of Lemma 3.1, and the matched-set discrepancy in the matching index vanishes (Condition (10)). Then the proposed estimator $\hat{V}$ is consistent:*

$$\left|\hat{V} - V\right| \xrightarrow{p} 0 \quad \text{as } n_T \to \infty.$$

*Proof:* See Appendix Appendix K.

The technical heart of Theorem 4.2 is not that any single within-cluster variance $s_j^2$ is close to $\sigma_{0,j}^2$—it need not be—but that the reuse-weighted aggregate $\frac{1}{n_T}\sum_{j\in\mathcal{C}} s_j^2\big[(\sum_t w_{jt})^2 - \sum_t w_{jt}^2\big]$ concentrates around its population target by averaging. The proof ("Term A" in Appendix Appendix K) decom-

[6] Equivalently, at the level of pairs, $\mathrm{Cov}(\hat{\Delta}_t, \hat{\Delta}_{t'} \mid \mathbf{S}, \mathbf{Z}) = \sum_{j\in\mathcal{C}} w_{jt}w_{jt'}\sigma_{0,j}^2$ for $t \neq t'$, and summing over $t_1 < t_2$ recovers the second term of $\hat{V}$.

poses each $s_j^2$ into a noise part, a within-cluster mean-shift controlled by the matched-set discrepancy in the index, and a cross term, and bounds each under the bounded-reuse-moment and vanishing-discrepancy conditions. The reuse-moment requirement is exactly the one that drives the central limit theorem (Lemma 3.1) and is verified scheme by scheme in Proposition 3.3; Theorem 4.2 therefore imposes no matching-specific condition beyond those already required for Theorem 3.2.

### *4.3. $\hat{V}$ as a generalization of existing variance estimators*

We now show that $\hat{V}$ acts as a generator for the variance estimators used in the matching literature: existing estimators arise as special cases, obtained by fixing a weighting scheme and a choice of control-variance proxy.

The estimator $\hat{V}$ nests most directly is Abadie and Imbens (2006). Specialize $\hat{V}$ to fixed-$M$ nearest-neighbor matching with uniform weights, $w_{jt} = 1/M$ and $\hat{Y}_t(0) = \frac{1}{M}\sum_{j\in\mathcal{C}_t} Y_j$. Because $w_{jt} = 1/M$ on exactly the $K_M(j)$ treated units that match control $j$, the off-diagonal bracket reduces to a reuse count,

$$\Big(\textstyle\sum_t w_{jt}\Big)^2 - \sum_t w_{jt}^2 \;=\; \frac{K_M(j)^2}{M^2} - \frac{K_M(j)}{M^2} \;=\; \frac{K_M(j)\big\{K_M(j)-1\big\}}{M^2}.$$

The second ingredient is the control-variance proxy. Abadie and Imbens (2006) do not read it off the matched clusters; they estimate it from a separate nearest-neighbor search within the control group,

$$\hat{\sigma}_0^2(\mathbf{X}_j) = \frac{J}{J+1}\Big(Y_j - \frac{1}{J}\sum_{\ell\in\mathcal{J}_J^{(0)}(j)} Y_\ell\Big)^2,$$

computed from the $J$ nearest neighbors of $j$ within the control group (with an analogous treated-group search for the treated term). Substituting this proxy for the within-cluster $s_j^2$ in Equation (12) yields

$$\hat{V}^{\mathrm{AI}} = \frac{1}{n_T}\sum_{t\in\mathcal{T}}\Big(Y_t - \tfrac{1}{M}\sum_{j\in\mathcal{C}_t} Y_j - \hat{\tau}\Big)^2 + \frac{1}{n_T}\sum_{j\in\mathcal{C}}\hat{\sigma}_0^2(\mathbf{X}_j)\,\frac{K_M(j)\big\{K_M(j)-1\big\}}{M^2},$$

which is exactly the Abadie and Imbens (2006) estimator. The only difference from $\hat{V}$ is therefore this proxy: $\hat{V}$ reads $s_j^2$ off the matching already formed. Thus $\hat{V}$ recovers the Abadie and Imbens (2006) reuse count as a special case, but the weight-product form exposes its origin as a shared-control covariance and extends it to arbitrary, possibly non-uniform, weights.

This is also why $\hat{V}$ is substantially cheaper to compute: it reuses the treated-to-control matching already formed for the point estimate and performs no additional same-group searches (Tables 3 and 5).

The same form applies beyond covariate matching. For propensity-score matching with a known score (Abadie and Imbens, 2016), $\hat{V}$ is used verbatim, with $s_j^2$ now the within-cluster dispersion of control outcomes among units matched on the score; when the score is estimated, $\hat{V}$ targets the matching/reuse component and a first-stage correction term must be added, which we leave to future work. For caliper matching, Kormos et al. (2023) use a regression-adjusted variance estimator, whereas $\hat{V}$ applies directly to the matched output with no outcome model.

## 5. A Multiplier Bootstrap for Single Matching

Otsu and Rai (2017) showed that valid bootstrap inference for matching is possible for regression-adjusted matching estimators by holding the matching weights fixed and resampling the centered unit-level contributions after regression adjustment. Our goal is different but closely related. We ask what the corresponding bootstrap should be for the unadjusted, single-matching estimator studied above, under the same principle of keeping the matching weights fixed after matching.

Throughout we use the matched difference $\hat{\Delta}_t = Y_t - \hat{Y}_t(0)$ and $\hat{\tau} = n_T^{-1} \sum_{t \in \mathcal{T}} \hat{\Delta}_t$ of Section 4.2. Let $\xi_t, \zeta_j$ be i.i.d. mean-zero, unit-variance multipliers, such as the two-point Rademacher or Mammen weights.[7] The

[7] Rademacher: $\xi = \pm 1$, each with probability $1/2$. Mammen: $\xi = -(\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $\xi = (\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$; both have mean zero and unit variance.

multipliers are drawn independently of the data $\mathcal{D}_n = \{(Y_i, Z_i, \mathbf{X}_i)\}$ and independently of each other. We write $\mathbb{E}^*, \mathrm{Var}^*$, and $\Pr^*$ for expectation, variance, and probability conditional on $\mathcal{D}_n$.

### *5.1. The naive matched-difference bootstrap fails*

The most immediate matched-difference bootstrap treats the $n_T$ matched differences as independent observations:

$$T^*_{\text{naive}} = \frac{1}{\sqrt{n_T}} \sum_{t \in \mathcal{T}} \xi_t \big(\hat{\Delta}_t - \hat{\tau}\big). \tag{13}$$

Because the multipliers are independent across treated units, the bootstrap variance is simply the empirical variance of the matched differences:

$$\mathrm{Var}^*\big(T^*_{\text{naive}}\big) = \frac{1}{n_T} \sum_{t \in \mathcal{T}} \big(\hat{\Delta}_t - \hat{\tau}\big)^2.$$

This bootstrap can be reasonable in matching without replacement, where each control is used at most once and the matched differences are approximately independent. With replacement, however, the same control outcome may enter several matched differences, so the independence approximation breaks down.

Comparing with the consistent estimator $\hat{V}$ of Equation (12), the failure is exactly the missing shared-control covariance term derived in Section 4.2. The consequence is systematic under-coverage in matching with replacement, as documented in Section 6.

This term is zero when controls are not reused, but positive whenever the same

### *5.2. The covariance correction*

To restore validity we add a control-side term that injects precisely the missing covariance. Define

$$T^* = \frac{1}{\sqrt{n_T}} \sum_{t \in \mathcal{T}} \xi_t \big(\hat{\Delta}_t - \hat{\tau}\big) \; + \; \frac{1}{\sqrt{n_T}} \sum_{j \in \mathcal{C}} \zeta_j \, U_j, \tag{14}$$

where $\{\xi_t\}$ and $\{\zeta_j\}$ are i.i.d. mean-zero, unit-variance multipliers drawn independently of the data, and the control factors $U_j$ are computed from the data. The bootstrap randomness on the control side therefore enters through the multipliers: the $U_j$ are chosen so that each product $\zeta_j U_j$ is, conditionally on the data, mean zero with the reuse covariance as its variance,

$$\mathbb{E}^*[\zeta_j U_j] = 0, \qquad \mathrm{Var}^*(\zeta_j U_j) = s_j^2\, c_j, \qquad c_j := w_j^2 - \sum_{t \in \mathcal{T}} w_{jt}^2 \;=\; 2 \sum_{t_1 < t_2} w_{jt_1} w_{jt_2} \;\geq\; 0. \tag{15}$$

With this choice $\mathrm{Var}^*(T^*) = \hat{V}$ exactly (Equation (12)): the treated sum reproduces the first term of $\hat{V}$, and the control sum reproduces the reuse covariance in the second.

*Choices of $U_j$.*. Any construction satisfying (15) is admissible; two are natural.

- *Variance plug-in* (`boot_cov`[8]): $U_j = s_j \sqrt{c_j}$, using the within-cluster standard deviation $s_j$. Given the data $U_j$ is a fixed scale factor, and $\zeta_j U_j$ has exactly the moments in (15); this is the cleanest object for the theory and the smoothest in finite samples.

- *Cluster-centered residual* (`boot_cov_ej`): $U_j = \hat{e}_j \sqrt{c_j}$ with $\hat{e}_j = Y_j - \bar{Y}_j^{\mathrm{sub}}$, where $\bar{Y}_j^{\mathrm{sub}}$ is the $w$-weighted mean of the control outcomes in the subclass(es) containing $j$. Here $\mathrm{Var}^*(\zeta_j U_j) = \hat{e}_j^2\, c_j$, a conditionally unbiased proxy for $s_j^2 c_j$, so (15) and the identity $\mathrm{Var}^*(T^*) = \hat{V}$ hold in expectation over the matched clusters rather than exactly. This variant is more residual-bootstrap-like and fully data-driven, at the cost of slightly noisier standard errors.

**Theorem 5.1** (Validity of the covariance-corrected bootstrap)**.** *Assume the multipliers $\{\xi_t\}_{t \in \mathcal{T}}, \{\zeta_j\}_{j \in \mathcal{C}}$ are independent, mean zero and unit variance, with a uniform moment bound $\sup_t \mathbb{E}^*|\xi_t|^{2+\delta} + \sup_j \mathbb{E}^*|\zeta_j|^{2+\delta} \leq C_\delta < \infty$ for*

[8] Method notation as used in our code and in simulation.

*some $\delta > 0$; that $U_j$ satisfies (15); the bootstrap negligibility condition*

$$\frac{\max_{t\in\mathcal{T}} |\hat{\Delta}_t - \hat{\tau}|}{\sqrt{n_T \hat{V}}} \xrightarrow{p} 0, \qquad \frac{\max_{j\in\mathcal{C}} |U_j|}{\sqrt{n_T \hat{V}}} \xrightarrow{p} 0;$$

*and that the conditions of Theorems 3.2 and 4.2 hold, so that $\hat{V} \xrightarrow{p} V$ and $\sqrt{n_T}(\hat{\tau} - B_n - \tau) \Rightarrow N(0, V)$. Then, conditional on the data,*

$$T^* \mid \mathcal{D}_n \;\Rightarrow_p\; N(0,V), \qquad \sup_{x\in\mathbb{R}} \Big| \Pr{}^*(T^* \le x) - \Pr\big\{\sqrt{n_T}(\hat{\tau} - B_n - \tau) \le x\big\} \Big| \xrightarrow{p} 0.$$

*Proof:* See Appendix Appendix L. The argument is a conditional Lyapunov central limit theorem for the triangular array $\{n_T^{-1/2}\xi_t(\hat{\Delta}_t - \hat{\tau})\} \cup \{n_T^{-1/2}\zeta_j U_j\}$, whose conditional variance is exactly $\hat{V}$; consistency of $\hat{V}$ and the original CLT then transfer the bootstrap law to the sampling law.

*5.3. Connection to Otsu–Rai.*

Our bootstrap is closely related in form to the weighted bootstrap of Otsu and Rai (2017). In their regression-adjusted setting, write

$$R_i = Y_i - \hat{m}_0(X_i)$$

for the residual after outcome regression adjustment, and write

$$\hat{R}_t(0) = \sum_{j\in\mathcal{C}_t} w_{jt} R_j$$

for the matched residualized control outcome. Their bootstrap can be written schematically as

$$T^*_{\mathrm{OR}} = \frac{1}{\sqrt{n_T}} \sum_{t\in\mathcal{T}} \xi_t \big\{ R_t - \hat{R}_t(0) - \hat{\tau}^{RA} \big\} + \frac{1}{\sqrt{n_T}} \sum_{j\in\mathcal{C}} \zeta_j V_j,$$

where the control-side term is based on the total reuse weight and residualized control outcome, schematically $V_j = -w_j R_j$. Thus OR17 and our bootstrap share the same broad principle: hold the matching weights fixed

and add a control-side term rather than resampling matched differences as if they were independent.

The difference is the target estimator. OR17 is designed for a regression-adjusted matching estimator, whereas this paper studies the unadjusted single-matching estimator. For that estimator, the missing term is not the full reuse-weighted residual variance $w_j^2$, but only the off-diagonal shared-control covariance

$$c_j = w_j^2 - \sum_{t \in \mathcal{T}} w_{jt}^2.$$

Our correction therefore uses $s_j^2 c_j$, while an OR17-style control term based on $w_j^2$ includes both diagonal and off-diagonal control-noise variation. The diagonal part is already present in the empirical variance of the matched differences, so adding it again over-corrects or otherwise targets a different first-order representation.

For this reason, OR17 is not directly comparable to our method in the simulations unless the regression-adjusted and unadjusted estimators coincide. We therefore use OR17-style procedures only as diagnostics. One variant uses a local matched-set average as the regression adjustment; this produces a control-side term analogous to using $w_j^2$ rather than $c_j$, and illustrates the importance of the diagonal correction. A second variant uses no regression adjustment. This variant is appropriate only in the flat-surface design, where the true regression adjustment is zero; in nonlinear designs it does not remove matching bias and therefore is not a valid competitor to a fully regression-adjusted OR17 implementation.

## 6. Simulation

We now study the finite-sample behavior of the procedures developed above and benchmark them against existing methods. The simulations have two goals: first, to confirm the asymptotic normality underlying Theorem 3.2; and second, to assess the coverage of confidence intervals built from the analytic estimator $\hat{V}$ (Theorem 4.2) and from the covariance-corrected bootstrap (Theorem 5.1). We use two designs that deliberately separate the

two sources of error in matching: a nonlinear-surface design after Otsu and Rai (2017), in which matching bias can be large but overlap remains adequate, and a flat-surface design with no matching bias, in which we vary overlap and hence the amount of control reuse. Design 1 asks whether the variance estimators deliver valid inference once matching bias is removed; Design 2 asks whether inference remains valid as replacement-induced control reuse becomes more severe. In both designs, we use 5-nearest-neighbor matching with uniform weights $w_{jt} = 1/5$ and nominal 95% confidence intervals.

### *6.1. The six inference procedures*

We compare six procedures, each producing an estimate $\widehat{\mathrm{Var}}(\hat{\tau})$ and the interval $\hat{\tau} \pm 1.96\,\widehat{\mathrm{Var}}(\hat{\tau})^{1/2}$; Table 1 lists the procedures and their building blocks. They fall into two families. The *analytic* family writes the variance in closed form: `smv` is our proposed *single-matching* variance estimator $\hat{V}$ of Equation (12), which reads all its variance proxies off the treated-to-control matching already formed for the point estimate; `smv_noW` is an ablation that drops the reuse diagonal $\sum_t w_{jt}^2$ (and so over-counts the covariance); and `ai06` is Abadie and Imbens (2006), with each $\hat{\sigma}_i^2$ estimated from $M = 5$ same-group nearest neighbors; in this ATT setting that means matching control units to control units to estimate the $\sigma_{0,j}^2$ that enters the shared-control covariance term of Equation (12).

The bootstrap family comprises the reuse-blind `boot_naive` of Equation (13) and our two covariance-corrected variants of Section 5.2: `boot_cov` ($U_j = s_j\sqrt{c_j}$) and `boot_cov_ej` ($U_j = \hat{e}_j\sqrt{c_j}$). The names make the distinction explicit: `boot_naive` omits the shared-control covariance, while the `boot_cov` variants restore it.

The bootstrap and analytic families coincide when there is no reuse: the control correction in `boot_cov` vanishes and the $-\sum_t w_{jt}^2$ diagonal cancels the $w_j^2$ term in `smv`.

Table 1: The six inference procedures and their variance building blocks. Each method sums a treated-side and a control-side contribution. `smv` is our proposed single-matching estimator $\hat{V}$ of Equation (12), and `smv_noW` its ablation; both read all proxies off the treated-to-control matching, so they need *no* extra matching. Abadie and Imbens (2006) (`ai06`) instead requires two extra same-group matchings (treated-to-treated for $\hat{\sigma}^2_{1,t}$ and control-to-control for $\hat{\sigma}^2_{0,j}$). Here $\hat{\Delta}_t = Y_t - \hat{Y}_t(0)$, $w_j = \sum_t w_{jt}$, $s_j^2$ is the within-cluster variance, $\hat{e}_j = Y_j - \bar{Y}_j^{\text{sub}}$, and $\hat{\sigma}^2_{z,i}$ uses $M = 5$ same-group neighbors.

| Method | Treated term | Control term | Extra matching |
|---|---|---|---|
| `smv` (ours) | $\frac{1}{n_T}\sum_t(\hat{\Delta}_t - \hat{\tau})^2$ | $\frac{1}{n_T^2}\sum_j s_j^2(w_j^2 - \sum_t w_{jt}^2)$ | none |
| `smv_noW` | $\frac{1}{n_T}\sum_t(\hat{\Delta}_t - \hat{\tau})^2$ | $\frac{1}{n_T^2}\sum_j s_j^2\, w_j^2$ | none |
| `ai06` | $\frac{1}{n_T}\sum_t(\hat{\Delta}_t - \hat{\tau})^2$ | $\frac{1}{n_T^2}\sum_j \hat{\sigma}^2_{0,j} w_j^2$ | treated–treated & control–control $M$-NN |
| `boot_naive` | $(\hat{\Delta}_t - \hat{\tau})^2$ | — | none (reuse-blind) |
| `boot_cov` | $(\hat{\Delta}_t - \hat{\tau})^2$ | $(w_j^2 - \sum_t w_{jt}^2)\, s_j^2$ | none |
| `boot_cov_ej` | $(\hat{\Delta}_t - \hat{\tau})^2$ | $(w_j^2 - \sum_t w_{jt}^2)\, \hat{e}_j^2$ | none |

### *6.2. Design 1: the Otsu–Rai nonlinear surface*

We begin with the simulation design of Otsu and Rai (2017), which features nonlinear response surfaces known to be challenging for bootstrap inference. Formally, the data generating process is:

$$
\begin{aligned}
&\{Y_i, Z_i, \mathbf{X}_i\}_{i=1}^n, \\
&Y_i(0) = m(\|\mathbf{X}_i\|) + \varepsilon_i, \quad Y_i(1) = \tau + Y_i(0), \\
&Z_i = \mathbb{I}\{P(\mathbf{X}_i) \geq v_i\}, \quad v_i \sim U[0,1], \\
&P(\mathbf{X}_i) = \gamma_1 + \gamma_2\|\mathbf{X}_i\|, \quad \mathbf{X}_i = (X_{1i}, \ldots, X_{Ki})', \\
&X_{ji} = \xi_i|\zeta_{ji}|/\|\boldsymbol{\zeta}_i\|, \quad j = 1, \ldots, K, \\
&\xi_i \sim U[0,1], \quad \boldsymbol{\zeta}_i \sim N(\mathbf{0}, I_K),
\end{aligned}
$$

where we set $(\gamma_1, \gamma_2) = (0.15, 0.7)$ for the propensity score.[9] We fix the treatment effect at $\tau = 0$ and vary the covariate dimension $K \in \{2, 4, 8\}$. The error term is drawn as $\varepsilon_i \sim N(0, 0.2^2)$. Outcome functions $m(\cdot)$ are taken from the six nonlinear curves reported in Otsu and Rai (2017) and reproduced in Table 2.

We implement 5-nearest neighbor matching with uniform weighting ($w_{jt} = 1/5$) across 1000 replications, varying the total sample size $n$ from 250 to 5000.

Table 2: Nonlinear outcome functions $m(z)$ used in simulations

| Curves | $m(z)$ |
|---|---|
| 1 | $0.15 + 0.7z$ |
| 2 | $0.1 + z/2 + \exp\left(-200(z-0.7)^2\right)/2$ |
| 3 | $0.8 - 2(z-0.9)^2 - 5(z-0.7)^3 - 10(z-0.6)^{10}$ |
| 4 | $0.2 + \sqrt{1-z} - 0.6(0.9-z)^2$ |
| 5 | $0.2 + \sqrt{1-z} - 0.6(0.9-z)^2 - 0.1z\cos(30z)$ |
| 6 | $0.4 + 0.25\sin(8z-5) + 0.4\exp\left(-16(4z-2.5)^2\right)$ |

In this design, most of the control surfaces $f_0(\mathbf{X}) = m(\|\mathbf{X}\|)$ are strongly nonlinear, so for $K \geq 3$ the matching bias $B_n$ is the dominant error (Abadie and Imbens, 2006). We therefore report coverage after a bias correction by subtracting the oracle bias calculated from true $f_1, f_0$.

Figure 1 visualizes the coverage across all $K$, $n_T$, and curve combinations for the three key methods. Four points summarize results. First, our analytic (non-bootstrap) estimator $\hat{V}$ (`smv`) attains correct coverage once the matching bias is removed. Second, our covariance-corrected bootstrap (`boot_cov`) is equally valid and is not computationally expensive—only about 1.5× the cost of the naive matched-difference bootstrap. Third, both are competitive in accuracy with Abadie and Imbens (2006) while being markedly cheaper to

[9] This construction keeps the treated/control ratio at $\approx$ 1:1 regardless of dimension. Since $\|\mathbf{X}_i\| = \xi_i \|\boldsymbol{\zeta}_i\|/\|\boldsymbol{\zeta}_i\| = \xi_i \sim \text{Uniform}(0,1)$ for every $K$ (the Euclidean norm of the vector of $|\zeta_{ji}|$ equals $\|\boldsymbol{\zeta}_i\|$), the propensity score $P(\mathbf{X}_i) = \gamma_1 + \gamma_2\|\mathbf{X}_i\|$ has the same marginal distribution in all dimensions, with $\mathbb{E}[n_T/n] = \gamma_1 + \gamma_2/2 = 0.15 + 0.35 = 0.50$, yielding $n_T : n_C \approx$ 1:1.

compute: Table 3 reports the median per-replication compute time broken down by $K$ and $n_T$; compute time is nearly identical across $K$ for fixed $n_T$ (confirming it is driven by the matching cost, not the curve or dimension), while `ai06` is one to two orders of magnitude slower (e.g. 0.29 s versus 31.8 s at $K = 2$, $n = 1000$) because it must perform two additional within-group $M$-NN searches for variance estimation. Fourth, `smv_noW` is strictly worse than `smv` throughout, confirming the theoretical necessity of the diagonal correction $\sum_t w_{jt}^2$. . The occasional mild over-coverage of `smv` at high $K$

Table 3: Design 1 (nonlinear surface): median compute time per replication (seconds) by method, $K$, and $n_T$, averaged over curves 1–6 and matching arms. Time is determined primarily by $n_T$ and $K$ (matching cost), not the outcome curve.

| | $K = 2$ | | | $K = 4$ | | | $K = 8$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $n = 250$ | $n = 500$ | $n = 1000$ | $n = 250$ | $n = 500$ | $n = 1000$ | $n = 250$ | $n = 500$ | $n = 1000$ |
| `smv` | 0.146 | 0.193 | 0.287 | 0.146 | 0.195 | 0.284 | 0.148 | 0.197 | 0.286 |
| `smv_noW` | 0.145 | 0.192 | 0.282 | 0.145 | 0.193 | 0.277 | 0.147 | 0.196 | 0.278 |
| `boot_cov` | 0.157 | 0.206 | 0.303 | 0.157 | 0.208 | 0.299 | 0.159 | 0.209 | 0.299 |
| `boot_cov_ej` | 0.138 | 0.174 | 0.237 | 0.139 | 0.175 | 0.233 | 0.141 | 0.178 | 0.232 |
| `ai06` | 7.749 | 15.699 | 31.814 | 7.818 | 15.954 | 31.838 | 7.986 | 16.300 | 32.250 |

is benign and shrinks as $n_T$ grows. CI-length details are in Appendix Appendix M.

OR17-style results are reported in Appendix Appendix M.1 as diagnostics: the no-adjustment version (or17_Y) unsurprisingly over-covers after oracle bias removal, while the local-adjustment version (or17) is not a faithful implementation of the regression-adjusted OR17 estimator and should not be interpreted as a direct competitor.

### *6.3. Design 2: control reuse with a flat outcome surface*

The second design isolates control reuse from matching bias. Treated units and good controls share $X_i \sim N(\mathbf{0}, I_K)$; bad controls have $X_j \sim N(a\,\mathbf{1}_K, I_K)$ with $a = 8$ ($K = 1$) or $a = 6$ ($K > 1$), placing them outside the caliper and effectively invisible to the matcher. The control pool has $n_C = 4n_T$ units total, of which $n_G = \lfloor q\, n_T \rfloor$ are good controls and $n_B = n_C - n_G$ are bad; varying $q \in \{0.125, 0.25, 0.5, 1, 2, 4\}$ tunes the degree of reuse without altering $n_C$. The control outcome surface is flat, $Y_j(0) = \varepsilon_j$

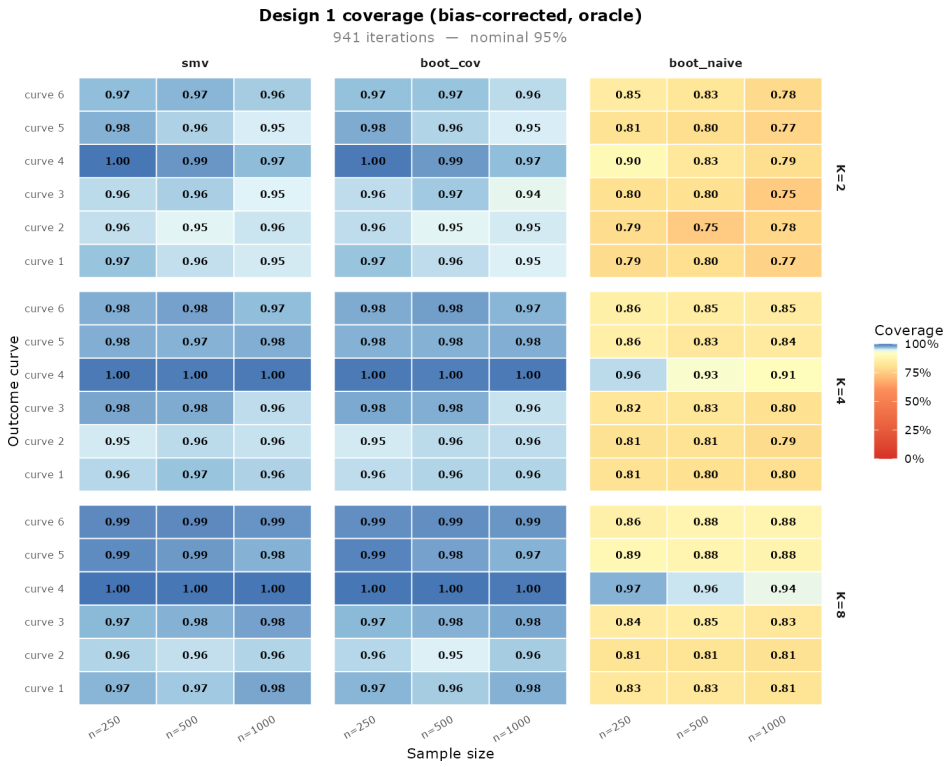


Figure 1: Design 1 coverage (oracle bias-corrected, nominal 95%) for three methods across covariate dimensions $K \in \{2, 4, 8\}$, sample sizes $n_T \in \{250, 500, 1000\}$, and six nonlinear outcome curves. Cell color: blue $\approx$ nominal, red = undercoverage, dark blue = overcoverage. `smv` and `boot_cov` hold coverage throughout; `boot_naive` fails under heavy reuse.

with $\mathbb{E}[\varepsilon_j] = 0$, so $f_0 \equiv 0$ and the matching bias $B_n$ is structurally zero no matter how imperfect the covariate match; the treatment effect is the constant null $\tau = 0$.

We vary $K \in \{1, 2, 4\}$ and $n_T$ up to 1,000, again using 5-NN matching. Since $B_n \equiv 0$, any coverage error reflects variance estimation alone, making this a clean test of control reuse. It also provides the cleanest comparison with OR17-style bootstraps: because $m_0 \equiv 0$, the correct regression adjustment is zero, so the no-adjustment OR17 variant is correctly specified, whereas the local-average adjustment remains deliberately misspecified.

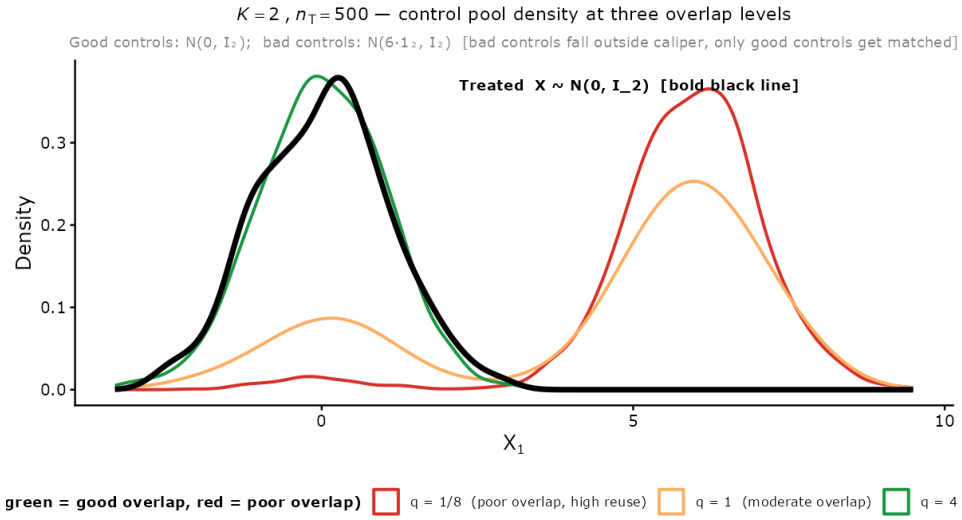


Figure 2: Design 2 DGP: $X_1$ density of the control pool at three overlap levels ($K = 2$, $n_T = 500$). **Bold black** = treated $N(0, I_2)$. Green ($q = 4$): pool is mostly good controls with high overlap. Amber ($q = 1$): moderate overlap. Red ($q = 1/8$): pool dominated by bad controls at $N(6 \cdot \mathbf{1}_2, I_2)$; good controls are scarce and each is reused many times.

Figure 2 illustrates how the control pool composition shifts with $q$: at high $q$ the pool is dominated by good controls that overlap with the treated distribution, while at low $q$ bad controls make up the bulk of the pool, forcing the matcher to reuse the few eligible controls repeatedly. Figure 3 tracks the weight-regularity diagnostics as $q$ decreases: $\mathrm{ESS}/n_T$ falls and $\max_j w_j / \sqrt{n_T + \sum_j w_j^2}$ rises, confirming that the regularity condition from Assumption 7 becomes more binding under heavy reuse.

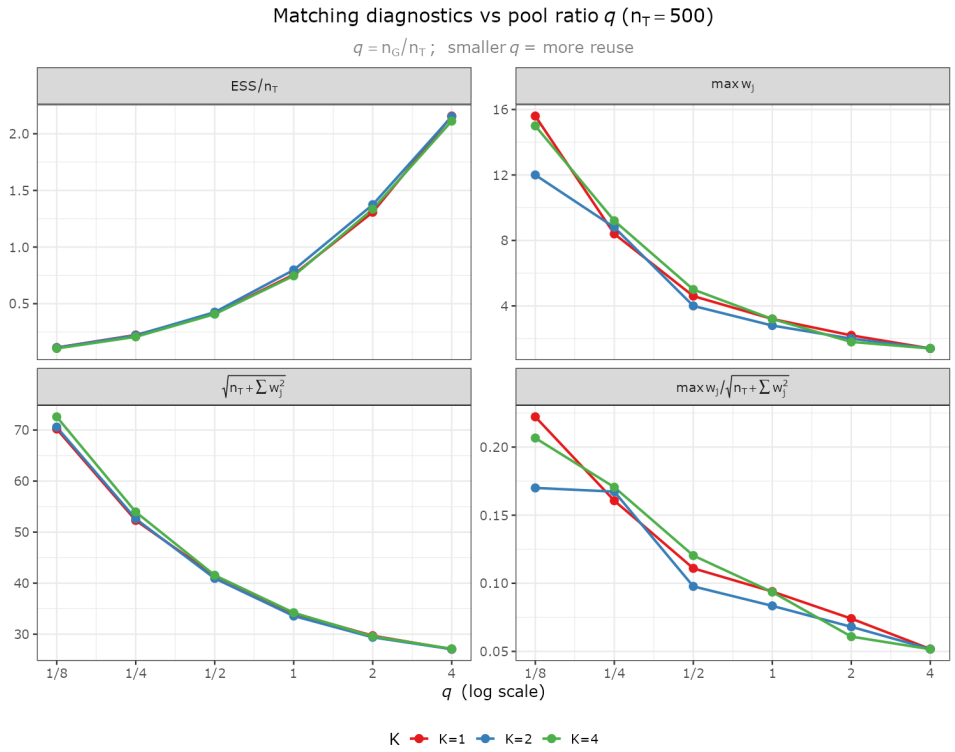


Figure 3: Matching diagnostics vs. pool ratio $q$ ($n_T = 500$, $K \in \{1, 2, 4\}$). ESS/$n_T$ measures effective distinct controls; $\max_j w_j / \sqrt{n_T + \sum_j w_j^2}$ is the weight-regularity ratio from Assumption 7: it must tend to zero for the CLT to apply. Both decreases as $q \downarrow$.

Table 4 reports coverage by $q$. The naive `boot_naive` fails dramatically as overlap tightens, collapsing from 0.92 at $q = 4$ to 0.49 at $q = 0.125$—a direct, quantitative demonstration of the deleted shared-control covariance (Section 5.1). In contrast, our `smv` and `boot_cov`, along with `ai06`, hold 0.94–0.95 across every overlap level. The ablation `smv_noW` runs slightly conservative (up to 0.96) because dropping the $-\sum_t w_{jt}^2$ diagonal over-counts the correction. Figure 4 shows this pattern is stable across $K$.

The OR17-style methods reinforce the same point. The no-adjustment variant performs well in Design 2, as expected, because $m_0 \equiv 0$ is the correct regression adjustment. The local-average variant under-covers as reuse increases, showing that a residualized control term based on the full reuse

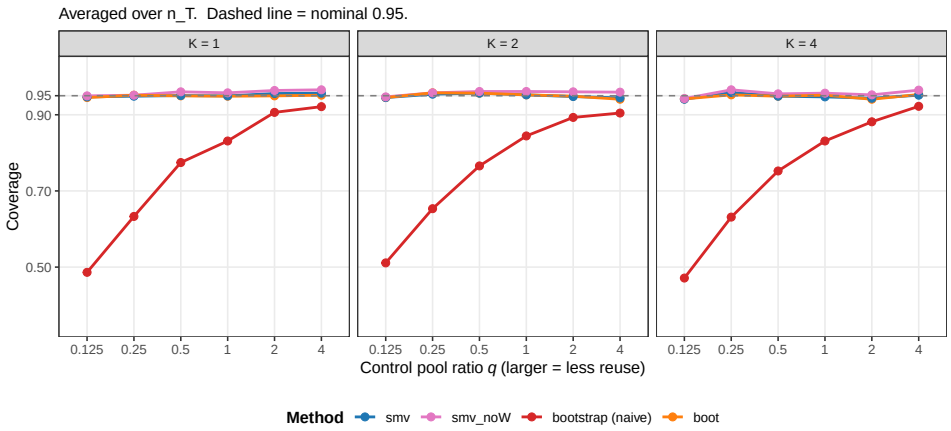


Figure 4: Design 2: Coverage vs. pool ratio $q$ by $K \in \{1, 2, 4\}$, four estimators: `smv`, `smv_noW`, `boot_cov`, `boot_naive`. Averaged over $n_T \in \{250, 500, 1000\}$. Dashed line: nominal 0.95.

weight $w_j^2$ does not target the same variance decomposition as the unadjusted estimator. The issue is therefore not OR17 itself, but the mismatch between the regression-adjusted linearization and the unadjusted single-matching estimand. See Appendix Appendix M.2 for result and discussion.

On computation cost, we observe the advantage of our method over Abadie and Imbens (2006)'s method, same as the previous simulation. Table 5 reports the median per-replication time for this design: `smv` and its ablation cost about a fifth of a second, the wild bootstraps (`boot_cov`, `boot_cov_ej`) run at roughly 1.5× the naive bootstrap, while `ai06` is more than two orders of magnitude slower (56.5 s versus $< 0.25$ s).

## 7. Application: Education Program Evaluation in Brazil

To illustrate the practical importance of our variance estimation framework, we analyze data from Brazil's "Jovem de Futuro" (Young of the Future) education program, following the experimental design of Barros et al. (2012) and Ferman (2021). This application demonstrates how our robust inference methods affect substantive conclusions in a setting with extensive control unit reuse—precisely the scenario where existing methods can fail.

Table 4: Control-reuse design: coverage of nominal 95% intervals by overlap parameter $q$ (pooled over $K \in \{1, 2, 4\}$ and $n_T$); smaller $q$ means heavier reuse. No bias correction is applied. The matched-difference `boot_naive`, which would be much more plausible without replacement, collapses as replacement-induced reuse increases; the reuse-aware methods stay near nominal throughout.

| Method | $q$=0.125 | $q$=0.25 | $q$=0.5 | $q$=1 | $q$=2 | $q$=4 |
|---|---|---|---|---|---|---|
| `smv` (ours) | .944 | .954 | .952 | .950 | .949 | .951 |
| `smv_noW` | .946 | .958 | .959 | .958 | .959 | .963 |
| `ai06` | .944 | .953 | .953 | .949 | .949 | .951 |
| `boot_cov` | .944 | .954 | .952 | .951 | .946 | .948 |
| `boot_naive` | .490 | .639 | .764 | .836 | .894 | .916 |

Table 5: Design 2 (control reuse): median compute time per replication (seconds), averaged over $K \in \{1, 2, 4\}$, $n_T$, and $q$. `ai06` is more than two orders of magnitude slower because of its two extra within-group matchings.

| Method | Median time (s) |
|---|---|
| `smv` (ours) | 0.217 |
| `smv_noW` | 0.193 |
| `boot_cov` | 0.229 |
| `boot_cov_ej` | 0.166 |
| `boot_naive` | 0.145 |
| `ai06` | 56.531 |

### *7.1. Data and Matching Design*

The Jovem de Futuro program offered management strategies and conditional grants to schools in Rio de Janeiro and São Paulo from 2010-2012. Following Ferman (2021)'s approach, we employ a within-study comparison design where experimental control schools (those randomized to receive no intervention) serve as our "treatment" group. We match these $n_T = 54$ experimental control schools to $n_C = 4{,}447$ non-participating schools to estimate what should be a null effect if matching successfully removes selection bias. See also Che et al. (2024).

Pre-treatment covariates consist of standardized test scores from 2007-2009 and a state indicator. Table 6 shows substantial pre-treatment dif-

ferences between experimental and non-participating schools, with experimental schools having consistently lower baseline scores across all years, motivating the use of matching methods.

Table 6: Summary Statistics for Brazilian School Data

| Variable | Non-participating Schools (Control) | Experimental Schools (Treated) | Standardized Difference |
|---|---|---|---|
| Score 2007 | 0.047 | −0.028 | −0.075 |
| Score 2008 | 0.008 | −0.010 | −0.018 |
| Score 2009 | 0.023 | −0.025 | −0.048 |
| São Paulo (%) | 78.1 | 72.2 | −5.9 |
| Sample size | 4,447 | 54 | |

Note: Test scores are standardized with mean 0 and standard deviation 1 in the full sample. São Paulo percentage indicates the proportion of schools from São Paulo state.

We implement radius matching following Che et al. (2024), using the $L^\infty$ distance metric with a distance caliper of $c = 0.35$. We impose covariate-specific calipers of 0.2 standard deviations for each pre-treatment test score (2007-2009) and require near-exact matching on state with a caliper of 0.001. This configuration ensures high-quality matches while maintaining adequate sample size—49 of 54 treated units (91%) find at least one match within the specified radius, with the remaining 5 units matched adaptively to their nearest neighbor. For matched units, we apply synthetic control weights to minimize covariate imbalance within each matched set. The matching procedure achieves excellent covariate balance, with post-matching standardized differences below 0.1 for all covariates. On average, each treated unit receives 3.22 controls with nonzero synthetic control weights among the 49 feasible matched units (3.0 when averaged over all 54 treated units including the 5 adaptive matches).

### *7.2. Control Unit Reuse and Effective Sample Size*

A key feature of this application is the minimal reuse of control schools in our matched sample. Table 7 presents diagnostics that characterize the dependency structure created by matching. The mean control reuse of 1.02

indicates that control schools are rarely matched to multiple treated units—on average, each control school is matched to just one treated unit, with only a small fraction matched to two treated units.

Table 7: Control Unit Reuse and Effective Sample Size in the Matched Sample

| Statistic | Value |
|---|---|
| Mean control reuse | 1.02 |
| Median control reuse | 1 |
| Maximum control reuse | 2 |
| Proportion of controls matched to multiple treated units | 1.9% |
| Effective sample size (ESS) | 95 |
| Number of unique controls | 155 |
| ESS/Number of unique controls | 61.3% |

Note: Mean control reuse measures the average number of times each control unit is matched to treated units. A value of 1 indicates no reuse; higher values indicate greater dependency in the matched sample. The effective sample size (ESS) accounts for unequal weighting across controls, capturing the reduction in effective independent information.

This minimal control reuse arises from the combination of a small treated sample ($n_T = 54$) relative to the large control reservoir ($n_C = 4{,}447$), along with our radius matching design that allows variable numbers of matches per treated unit. The maximum reuse of only 2 indicates that even the best control schools are matched to at most two treated units.

The effective sample size (ESS, defined in Section 4) of 95 is notably lower than the 155 unique controls used, with an ESS ratio of 61.3%. These two quantities measure different things: mean reuse $\bar{w} = \frac{1}{n_C}\sum_j w_j$ is an unweighted count of how many treated units each control is matched to, while $\text{ESS} = n_T^2 / \sum_j w_j^2$ accounts for the within-matched-set synthetic-control weights assigned by the CSM algorithm, which are generally unequal. Even when $\bar{w} \approx 1$ (each control is used by at most one treated unit on average), the CSM weights within each matched set can differ substantially across controls—those receiving larger synthetic-control weights contribute $w_j^2$ more to $\sum_j w_j^2$, reducing the ESS. Our variance estimator properly accounts for this through the $w_j^2$ terms in $\hat{V}$, ensuring valid inference regardless of the

weighting scheme employed.

### *7.3. Treatment Effect Estimates and Inference*

Table 8 presents estimates of the average treatment effect on the treated using our analytic estimator (`smv`) and the naive matched-difference bootstrap (`boot_naive`). The point estimate of 0.035 is close to zero, as expected in this within-study comparison. This null effect is by design: we are comparing experimental control schools (randomized to receive no treatment) to observationally similar non-participating schools. If our matching procedure successfully removes selection bias, we should find no systematic difference between these groups, validating the matching method's ability to create appropriate counterfactuals.

Table 8: Brazil application: ATT estimates for 2010 test scores using radius matching with synthetic control weights.

| Method | Point Estimate | SE | 95% CI | $n_T\hat{V}_E$ | $n_T\hat{V}_P$ |
|---|---|---|---|---|---|
| `smv` (ours) | 0.035 | 0.030 | $(-0.024, 0.094)$ | 0.055 | $-0.400$ |
| `boot_naive` | 0.035 | 0.029 | $(-0.022, 0.092)$ | — | — |

Note: `boot_naive` based on 1,000 wild-bootstrap replications.
$n_T\hat{V}_P$ is obtained by subtracting $n_T\hat{V}_E$ from $\hat{V}$; a small negative value can arise in finite samples and indicates minimal treatment-effect heterogeneity here.

Both inference methods produce similar standard errors (0.030 vs 0.029), with 95% confidence intervals that include zero. This similarity is expected given the minimal control reuse. With limited dependency structure, the wild bootstrap performs adequately, and both methods lead to the same substantive conclusion: we cannot reject the null hypothesis of no selection bias after matching.

An important feature of our variance estimator is the decomposition into measurement error variance ($\hat{V}_E$) and population heterogeneity variance ($\hat{V}_P$). The scaled variance components show $n_T\hat{V}_E = 0.055$, indicating modest sampling variability due to residual outcome noise. The estimate $n_T\hat{V}_P = -0.400$ is negative, which occurs because we obtain $n_T\hat{V}_P$ through subtraction: we subtract $n_T\hat{V}_E$ from $\hat{V}$ in Equation (12). While theoretically this decomposition is accurate in probability limit, in finite samples

it is possible for $\hat{V} < n_T \hat{V}_E$ due to sampling variability of both estimators. The negative value suggests minimal treatment effect heterogeneity in this sample. Developing improved estimators for $n_T \hat{V}_P$ that ensure non-negativity while maintaining consistency remains an interesting direction for future work.

## 8. Conclusion

We have presented a refined framework for inference with matching estimators. Our analysis establishes a central limit theorem for a broad class of matching procedures under heteroskedastic errors, reducing asymptotic normality to a single weight-regularity condition—verified through an exponential-tail/reuse-count argument for nearest-neighbor matching and, for radius/caliper matching, directly through caliper count concentration; introduces a computationally simple, consistent single-matching variance estimator whose control-reuse term is a transparent covariance correction generalizing Abadie and Imbens (2006); and derives a covariance-corrected multiplier bootstrap that repairs the otherwise-invalid matched-difference bootstrap.

Simulations across six procedures and two designs—one that stresses matching bias, one that isolates control reuse—demonstrate the practical payoff. While the intuitive matched-difference bootstrap can undercover severely, falling below 50% under heavy reuse, our analytic estimator and covariance-corrected bootstrap consistently deliver coverage close to the nominal rate (after matching bias is handled), even in moderately sized samples. The improvements are not subtle: coverage gaps can reach 20 percentage points, underscoring how minor-seeming theoretical adjustments can yield dramatic empirical benefits. We hypothesize this superior performance stems from our approach's robustness to control unit reuse—when the same high-quality controls are matched to multiple treated units, bootstrap resampling schemes struggle to properly account for the induced dependence, while our analytical variance estimator correctly captures this structure through its explicit treatment of within-cluster correlation.

Methodologically, our work parallels the role of heteroskedasticity-robust

variance estimation in regression: it provides a theoretically justified and broadly applicable correction that improves inference reliability. Practically, our results equip applied researchers with a *single-matching* variance estimator that is easy to compute—requiring only the treated-to-control matching already performed for the point estimate—and produces trustworthy confidence intervals in settings where existing approaches falter.

We view these contributions as refinements to a well-studied problem rather than a wholesale rethinking of matching inference. Yet the payoff of these refinements is large: by carefully addressing overlooked variance components and grounding inference in rigorous asymptotics, we achieve both theoretical clarity and dramatic gains in empirical performance. Future work may extend these tools to weighting methods and other causal estimators, further unifying the inferential foundations of design-based approaches in causal inference.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.


### Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.


### Data and code availability

Replication code for this article is available at https://github.com/jche/CSMatch. The Brazilian education data used in the empirical application are available from the same repository.

**Declaration of generative AI and AI-assisted technologies in the writing process**

During the preparation of this work the authors used a generative AI large language model in order to edit and improve the clarity of the written text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

**References**


Abadie, A., Imbens, G.W., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74, 235–267.

Abadie, A., Imbens, G.W., 2008. On the failure of the bootstrap for matching estimators. Econometrica 76, 1537–1557.

Abadie, A., Imbens, G.W., 2011. Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics 29, 1–11.

Abadie, A., Imbens, G.W., 2012. A martingale representation for matching estimators. Journal of the American Statistical Association 107, 833–843.

Abadie, A., Imbens, G.W., 2016. Matching on the estimated propensity score. Econometrica 84, 781–807.

Abadie, A., Spiess, J., 2022. Robust post-matching inference. Journal of the American Statistical Association 117, 1811–1827.

Austin, P.C., Cafri, G., 2020. Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. Statistics in Medicine 39, 1623–1640.

Barros, R., Carvalho, M.d., Franco, S., Rosalém, A., 2012. Impacto do projeto jovem de futuro. Est. Aval. Educ , 214–226.

Bodory, H., Camponovo, L., Huber, M., Lechner, M., 2020. The finite sample performance of inference methods for propensity score matching and weighting estimators. Journal of Business & Economic Statistics 38, 183–200.

Che, J., Meng, X., Miratrix, L., 2024. Caliper synthetic matching: Generalized radius matching with local synthetic controls. arXiv preprint arXiv:2411.05246 .

Cribari-Neto, F., Garcia, N.L., Vasconcellos, K.L., 2000. A note on inverse moments of binomial variates. Brazilian Review of Econometrics 20, 269–277.

Dehejia, R.H., Wahba, S., 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association 94, 1053–1062.

Ferman, B., 2021. Matching estimators with few treated and many control observations. Journal of Econometrics 225, 295–307.

Heckman, J.J., Ichimura, H., Todd, P.E., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The Review of Economic Studies 64, 605–654.

Hill, J., Reiter, J.P., 2006. Interval estimation for treatment effects using propensity score matching. Statistics in Medicine 25, 2230–2256.

Kallus, N., 2020. Generalized optimal matching methods for causal inference. J. Mach. Learn. Res. 21, 62–1.

Kormos, M., van der Pas, S.L., van der Vaart, A.W., 2023. Asymptotics of caliper matching estimators for average treatment effects. arXiv preprint arXiv:2304.08373 .

Lin, Z., Han, F., 2022. On regression-adjusted imputation estimators of the average treatment effect. arXiv preprint arXiv:2212.05424 .

Otsu, T., Rai, Y., 2017. Bootstrap inference of matching estimators for average treatment effects. Journal of the American Statistical Association 112, 1720–1732.

Potthoff, R.F., Woodbury, M.A., Manton, K.G., 2024. "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models .

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rubin, D.B., 1973. Matching to remove bias in observational studies. Biometrics , 159–183.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66, 688.

Shi, Z., Bhattacharjee, C., Balasubramanian, K., Polonik, W., 2024. Gaussian and bootstrap approximation for matching-based average treatment effect estimators. arXiv preprint arXiv:2412.17181 .

Smith, J.A., Todd, P.E., 2005. Does matching overcome lalonde's critique of nonexperimental estimators? Journal of Econometrics 125, 305–353.

Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. Statistical Science 25, 1–21.