Supplementary Appendix: Difference-in-differences with "bad controls"

Carolina Caetano[*] Brantly Callaway[†] Stroud Payne[‡] Hugo Sant'Anna[§]

August 4, 2026

This supplementary appendix contains proofs for some of the results in the main text as well as additional results and supporting details for the arguments in the main text. Appendix SA proves asymptotic normality of the imputation estimator provided in Section 6.1 in the main text. Appendix SB proves double robustness of the Neyman-orthogonal estimator proposed in Section 6.2 and then proves asymptotic normality when nuisance functions are estimated by machine learning. Appendix SC develops an alternative identification strategy under parallel trends for the bad control. Finally, Appendix SD provides Monte Carlo simulations.

SA Proofs for Imputation Estimator

The following assumption collects standard regularity conditions for Proposition 4. Define the vectors of regressors $R_i := (X_{it^*}, X_{i,t^*-1}, Z_i')'$ and $S_i := (X_{i,t^*-1}, W_i', Z_i')'$.

Assumption S1 (Regularity Conditions for Imputation Estimator)**.**

(i) $\mathbb{E}\|X_{t^*}\|^4$, $\mathbb{E}\|X_{t^*-1}\|^4$, $\mathbb{E}\|W\|^4$, $\mathbb{E}\|Z\|^4$, *and* $\mathbb{E}[\Delta Y_{t^*}^4]$ *are finite.*

(ii) The matrices $\mathbb{E}[RR' \mid D = 0]$ *and* $\mathbb{E}[SS' \mid D = 0]$ *are positive definite.*

Throughout this section, we use notation from Section 6.1 and write $\pi = \mathrm{P}(D = 1)$ and $n_0 = \sum_i (1-D_i)$. Under Assumption 8, we have $m_0(X_{t^*}, X_{t^*-1}, Z) = R'\beta$, $\mathbb{E}[X_{t^*} \mid S, D = 0] = S'\gamma$, and $\nu_0(X_{t^*-1}, W, Z) = S'\gamma\beta_1 + X_{t^*-1}\beta_2 + Z'\beta_3 = \tilde{R}'\beta$. The OLS estimators use only untreated observations:

$$\widehat{\beta} = \Big(\sum_{i=1}^{n} (1 - D_i) R_i R_i'\Big)^{-1} \sum_{i=1}^{n} (1 - D_i) R_i \Delta Y_{it^*},$$

$$\widehat{\gamma} = \Big(\sum_{i=1}^{n} (1 - D_i) S_i S_i'\Big)^{-1} \sum_{i=1}^{n} (1 - D_i) S_i X_{it^*}.$$

[*]John Munro Godfrey, Sr. Department of Economics, University of Georgia. carol.caetano@uga.edu

[†]John Munro Godfrey, Sr. Department of Economics, University of Georgia. brantly.callaway@uga.edu

[‡]Department of Economics, Vanderbilt University. robert.s.payne@vanderbilt.edu

[§]Department of Economics, University of Alabama at Birmingham. hsantanna@uab.edu

Define the error $u_i = \Delta Y_{it^*} - R_i'\beta$ and $v_i = X_{it^*} - S_i'\gamma$, which satisfy $\mathbb{E}[u \mid R, D = 0] = 0$ and $\mathbb{E}[v \mid S, D = 0] = 0$ under Assumption 8. Let $\widetilde{R}_i = \big((S_i'\gamma)', X_{i,t^*-1}', Z_i'\big)'$ correspond to R_i above but with X_{it^*} replaced by its conditional mean $S_i'\gamma$, and let

$$\boldsymbol{\Sigma}_R = \mathbb{E}\left[\frac{1-D}{1-\pi} RR'\right], \qquad \boldsymbol{\Sigma}_S = \mathbb{E}\left[\frac{1-D}{1-\pi} SS'\right].$$

The following lemma provides the asymptotic linear representations for $\widehat{\beta}$ and $\widehat{\gamma}$, which are standard consequences of OLS theory.

Lemma S1. *Under Assumption S1,*

$$\sqrt{n}(\widehat{\beta} - \beta) = \boldsymbol{\Sigma}_R^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1-D_i}{1-\pi} R_i\, u_i + o_p(1), \tag{S1}$$

$$\sqrt{n}(\widehat{\gamma} - \gamma) = \boldsymbol{\Sigma}_S^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1-D_i}{1-\pi} S_i\, v_i + o_p(1). \tag{S2}$$

Proof. For $\sqrt{n}(\widehat{\beta} - \beta)$, notice that

$$\begin{aligned}
\sqrt{n}\big(\widehat{\beta} - \beta\big) &= \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1-D_i}{1-\hat{\pi}} R_i R_i'\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1-D_i}{1-\hat{\pi}} R_i u_i \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1-D_i}{1-\pi} R_i R_i'\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1-D_i}{1-\pi} R_i u_i \\
&= \boldsymbol{\Sigma}_R^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1-D_i}{1-\pi} R_i\, u_i + o_p(1)
\end{aligned}$$

where the first equality holds by the definition of $\hat{\beta}$, the second equality holds by canceling the $(1-\hat{\pi})$ terms and then multiplying and dividing by $(1-\pi)$, and the last equality holds by the law of large numbers applied to the first term. An analogous argument holds for $\sqrt{n}(\widehat{\gamma} - \gamma)$. □

Proof of Proposition 4. First, we decompose

$$\sqrt{n}\Big(\widehat{\mathrm{ATT}}_{ra} - \mathrm{ATT}\Big) = \underbrace{\sqrt{n}\big(\widehat{m}_1 - \mathbb{E}[\Delta Y_{t^*} \mid D = 1]\big)}_{A} - \underbrace{\sqrt{n}\big(\widehat{\tau}_{ra} - \mathbb{E}[\nu_0 \mid D = 1]\big)}_{B}. \tag{S3}$$

For term A in Equation (S3),

$$\begin{aligned}
\sqrt{n}\big(\widehat{m}_1 - \mathbb{E}[\Delta Y_{t^*} \mid D = 1]\big) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_i}{\hat{\pi}} \Big(\Delta Y_{it^*} - \mathbb{E}[\Delta Y_{t^*} \mid D = 1]\Big) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_i}{\pi} \Big(\Delta Y_{it^*} - \mathbb{E}[\Delta Y_{t^*} \mid D = 1]\Big) + o_p(1)
\end{aligned} \tag{S4}$$

where the first equality holds by the definition of $\hat{m}_1$, the second equality holds by multiplying and dividing by π and because $\pi/\hat{\pi} \xrightarrow{p} 1$.

For term B in Equation (S3), from Equation (7) in the main text, we have that

$$\hat{\tau}_{ra} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{\pi}}\hat{\nu}_{0,i} \quad \text{where } \hat{\nu}_{0,i} := \hat{\nu}_0(X_{t^*-1}, W, Z) = S'\hat{\gamma}\hat{\beta}_1 + X_{t^*-1}\hat{\beta}_2 + Z'\hat{\beta}_3,$$

and we also use the shorthand notation $\nu_{0,i} := \nu_0(X_{i,t^*-1}, W_i, Z_i)$. Next, notice that

$$\sqrt{n}\Big(\widehat{\tau}_{ra} - \mathbb{E}[\nu_0 \mid D=1]\Big) = \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{\hat{\pi}}\big(\widehat{\nu}_{0,i} - \nu_{0,i}\big)}_{\text{B}_1} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{\hat{\pi}}\Big(\nu_{0,i} - \mathbb{E}[\nu_0 \mid D=1]\Big)}_{\text{B}_2},$$

For B_1, notice that

$$\begin{aligned}\widehat{\nu}_{0,i} - \nu_{0,i} &= \widehat{\nu}_{0,i} - S_i'\gamma\hat{\beta}_1 + S_i'\gamma\hat{\beta}_1 - \nu_{0,i} \\ &= S_i'\big(\hat{\gamma} - \gamma\big)\hat{\beta}_1 + \tilde{R}_i'\big(\hat{\beta} - \beta\big).\end{aligned} \tag{S5}$$

which implies that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{\hat{\pi}}\big(\widehat{\nu}_{0,i} - \nu_{0,i}\big) &= \left(\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\pi}\beta_1 S_i'\right)\sqrt{n}(\hat{\gamma} - \gamma) + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\pi}\tilde{R}_i'\right)\sqrt{n}(\hat{\beta} - \beta) + o_p(1) \\ &= \mathbb{E}\left[\frac{D}{\pi}\beta_1 S_i'\right]\sqrt{n}(\hat{\gamma} - \gamma) + \mathbb{E}\left[\frac{D}{\pi}\tilde{R}'\right]\sqrt{n}(\hat{\beta} - \beta) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1-D_i}{1-\pi}\left(\mathbb{E}\left[\frac{D}{\pi}\beta_1 S'\right]\boldsymbol{\Sigma}_S^{-1}S_i v_i + \mathbb{E}\left[\frac{D}{\pi}\widetilde{R}\right]'\boldsymbol{\Sigma}_R^{-1}R_i u_i\right) + o_p(1).\end{aligned} \tag{S6}$$

where the first equality holds by plugging in Equation (S5) and by consistency of $\hat{\pi}$ and $\hat{\beta}_1$, the second equality holds by applying the law of large numbers and continuous mapping theorem to the two sample average terms in the previous line, and the last equality holds by plugging in the result in Lemma S1.

For B_2, it holds immediately that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{\hat{\pi}}\Big(\nu_{0,i} - \mathbb{E}[\nu_0 \mid D=1]\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{\pi}\Big(\nu_{0,i} - \mathbb{E}[\nu_0 \mid D=1]\Big) + o_p(1) \tag{S7}$$

because $\hat{\pi}$ is consistent for π and by the continuous mapping theorem.

Next,

$$\psi_i^{ra} := \frac{D_i}{\pi}\Big(\Delta Y_{it^*} - \mathbb{E}[\Delta Y_{t^*} \mid D=1] - \nu_{0,i} + \mathbb{E}[\nu_0 \mid D=1]\Big)$$
$$- \frac{1-D_i}{1-\pi}\Big(\mathbb{E}\left[\frac{D}{\pi}\widetilde{R}\right]' \boldsymbol{\Sigma}_R^{-1} R_i u_i + \mathbb{E}\left[\frac{D}{\pi}\beta_1' S'\right] \boldsymbol{\Sigma}_S^{-1} S_i v_i\Big). \tag{S8}$$

Then, by combining Equations (S4), (S6) and (S7), it follows that

$$\sqrt{n}(\widehat{\text{ATT}}_{ra} - \text{ATT}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_i^{ra} + o_p(1) \xrightarrow{d} \mathcal{N}\big(0, \text{Var}(\psi^{ra})\big)$$

□

SB Proofs for Neyman Orthogonal Estimator

SB.1 Neyman Orthogonal Estimand

Proof of Proposition 5. We consider the expectation of each line in Equation (10) in the main text in turn.

$$\mathbb{E}\left[\frac{D}{\pi}\Delta Y_{t^*} - \frac{D}{\pi}\nu_0(X_{t^*-1}, W, Z)\right] = \mathbb{E}[\Delta Y_{t^*} \mid D=1] - \mathbb{E}[\nu_0(X_{t^*-1}, W, Z) \mid D=1] = \text{ATT},$$

which holds by the law of iterated expectations, Theorem 2, and the definition of ν_0. Next,

$$\begin{aligned}
&\mathbb{E}\left[-\frac{1-D}{\pi}\left(m_0(X_{t^*}, X_{t^*-1}, Z) - \nu_0(X_{t^*-1}, W, Z)\right)\frac{p(X_{t^*-1}, W, Z)}{1-p(X_{t^*-1}, W, Z)}\right] \\
&\qquad = -\frac{1-\pi}{\pi}\mathbb{E}\left[m_0(X_{t^*}, X_{t^*-1}, Z)\frac{p(X_{t^*-1}, W, Z)}{1-p(X_{t^*-1}, W, Z)} \;\middle|\; D=0\right] \\
&\qquad\quad + \frac{1-\pi}{\pi}\mathbb{E}\left[\nu_0(X_{t^*-1}, W, Z)\frac{p(X_{t^*-1}, W, Z)}{1-p(X_{t^*-1}, W, Z)} \;\middle|\; D=0\right] \\
&\qquad = 0,
\end{aligned}$$

where the first equality holds by the law of iterated expectations, and the second equality holds by applying the law of iterated expectations again to the first term and by the definition of ν_0. Next,

$$\begin{aligned}
&\mathbb{E}\left[-\frac{1-D}{\pi}\big(\Delta Y_{t^*} - m_0(X_{t^*}, X_{t^*-1}, Z)\big)\omega_0(X_{t^*}, X_{t^*-1}, Z)\right] \\
&\qquad = -\frac{1-\pi}{\pi}\mathbb{E}\Big[\big(\Delta Y_{t^*} - m_0(X_{t^*}, X_{t^*-1}, Z)\big)\omega_0(X_{t^*}, X_{t^*-1}, Z) \;\Big|\; D=0\Big] \\
&\qquad = 0
\end{aligned}$$

where the first equality holds by the law of iterated expectations, and the second equality holds by applying the law of iterated expectations again and by the definition of m_0. Combining the previous three displays implies the result. $\square$

Lemma S2. *Under Assumptions 6 and 7, for any integrable function $h(X_{t^*}(0), X_{t^*-1}, Z)$,*

$$\mathbb{E}\big[h(X_{t^*}(0), X_{t^*-1}, Z)\,\omega_0(X_{t^*}(0), X_{t^*-1}, Z) \mid D=0\big] = \frac{\pi}{1-\pi}\,\mathbb{E}\big[h(X_{t^*}(0), X_{t^*-1}, Z) \mid D=1\big].$$

Proof. Notice that, for $h := h(X_{t^*}(0), X_{t^*-1}, Z)$, we have that

$$\begin{aligned}
\mathbb{E}\big[h\,\omega_0(X_{t^*}(0), X_{t^*-1}, Z) \mid D=0\big] &= \mathbb{E}\left[h\,\frac{p(X_{t^*-1}, W, Z)}{1-p(X_{t^*-1}, W, Z)} \;\middle|\; D=0\right] \\
&= \mathbb{E}\left[\frac{p(X_{t^*-1}, W, Z)}{1-p(X_{t^*-1}, W, Z)}\,\mathbb{E}\big[h \mid X_{t^*-1}, W, Z, D=0\big] \;\middle|\; D=0\right] \\
&= \mathbb{E}\left[\frac{p(X_{t^*-1}, W, Z)}{1-p(X_{t^*-1}, W, Z)}\mathbb{E}\big[h \mid X_{t^*-1}, W, Z, D=1\big] \;\middle|\; D=0\right] \\
&= \frac{\pi}{1-\pi}\mathbb{E}\Big[\mathbb{E}\big[h \mid X_{t^*-1}, W, Z, D=1\big] \;\Big|\; D=1\Big] \\
&= \frac{\pi}{1-\pi}\mathbb{E}\big[h \mid D=1\big]
\end{aligned}$$

where the first equality holds by the definition of ω_0 in Equation (9) in the main text and the law of iterated expectations, the second equality holds by the law of iterated expectations, the third equality holds by Assumption 6, the fourth equality holds by a change of measure argument, and the last equality holds by the law of iterated expectations. $\square$

Proof of Proposition 6. Since $\widehat{\mathrm{ATT}}_{dr}(\hat\theta) \xrightarrow{p} \mathrm{ATT}_{dr}(\theta)$ by assumption, it suffices to show $\mathrm{ATT}_{dr}(\theta) =$ ATT under each condition, where

$$\mathrm{ATT}_{dr}(\theta) = m_1 - \tau(\theta) - T_3(\theta) - T_4(\theta), \tag{S9}$$

with $m_1 = \mathbb{E}[\Delta Y_{t^*} \mid D=1]$, $\tau(\theta) = \mathbb{E}[\nu_0(\theta_\nu) \mid D=1]$,

$$\begin{aligned}
T_3(\theta) &= \frac{1-\pi}{\pi}\mathbb{E}\left[\big(m_0(\theta_m) - \nu_0(\theta_\nu)\big)\frac{p(\theta_p)}{1-p(\theta_p)} \;\middle|\; D=0\right], \\
T_4(\theta) &= \frac{1-\pi}{\pi}\mathbb{E}\left[\big(\Delta Y_{t^*} - m_0(\theta_m)\big)\omega_0(\theta_\omega) \;\middle|\; D=0\right].
\end{aligned}$$

Since $\mathrm{ATT} = m_1 - \tau$ by Theorem 2, we aim to show that $\mathrm{ATT}_{dr}(\theta) = m_1 - \tau$.

Case (i): $(m_0(\theta_m), \nu_0(\theta_\nu)) = (m_0, \nu_0)$.

In this case, $\tau(\theta) = \tau$. For $T_3(\theta)$, notice that

$$\frac{1-\pi}{\pi}\mathbb{E}\left[(m_0 - \nu_0)\frac{p(\theta_p)}{1-p(\theta_p)} \;\middle|\; D=0\right] = \frac{1-\pi}{\pi}\mathbb{E}\left[\frac{p(\theta_p)}{1-p(\theta_p)}\underbrace{\mathbb{E}\Big[(m_0-\nu_0) \;\Big|\; X_{t^*-1}, W, Z, D=0\Big]} \;\middle|\; D=0\right] = 0$$

where the first equality holds by the law of iterated expectations and because $p(\theta_p)$ is a deterministic function given (X_{t^*-1}, W, Z), and the underlined term is equal to 0 by the definition of ν_0 in Equation (5). Next, for $T_4(\theta)$, notice that

$$\frac{1-\pi}{\pi}\mathbb{E}\Big[(\Delta Y_{t^*} - m_0)\omega_0(\theta_\omega) \;\Big|\; D=0\Big] = \frac{1-\pi}{\pi}\mathbb{E}\left[\omega_0(\theta_\omega)\underbrace{\mathbb{E}\Big[(\Delta Y_{t^*}-m_0) \;\Big|\; X_{t^*}, X_{t^*-1}, Z, D=0\Big]} \;\middle|\; D=0\right] = 0,$$

where the first equation holds by the law of iterated expectations and because $\omega_0(\theta_\omega)$ is a deterministic function given (X_{t^*}, X_{t^*-1}, Z), and the underlined term is equal to 0 by the definition of m_0 in Equation (1).

Finally, since $\tau(\theta) = \tau$ and $T_3(\theta) = T_4(\theta) = 0$, it imples that $\text{ATT}_{dr}(\theta) = \text{ATT}$.

Case (ii): $(p(\theta_p), \omega_0(\theta_\omega)) = (p, \omega_0)$.

For $T_3(\theta)$, write

$$T_3(\theta) = \frac{1-\pi}{\pi}\mathbb{E}\left[m_0(\theta_m)\frac{p}{1-p} \;\middle|\; D=0\right] - \frac{1-\pi}{\pi}\mathbb{E}\left[\nu_0(\theta_\nu)\frac{p}{1-p} \;\middle|\; D=0\right] =: T_3^A(\theta) - T_3^B(\theta),$$

which uses that $p(\theta_p) = p$, and we consider each of these terms individually.

$$T_3^A(\theta) = \frac{1-\pi}{\pi}\mathbb{E}\Big[m_0(\theta_m)\omega_0 \;\Big|\; D=0\Big] = \mathbb{E}[m_0(\theta_m) \mid D=1] \tag{S10}$$

where the first equality holds by the law of iterated expectations, the definition of ω_0 in Equation (9), and because $m_0(\theta_m)$ is a deterministic function given (X_{t^*}, X_{t^*-1}, Z), and the second equality holds by Lemma S2. Next,

$$T_3^B(\theta) = \mathbb{E}[\nu_0(\theta_\nu) \mid D=1] = \tau(\theta). \tag{S11}$$

where the first equality holds by a change of measure argument, and the second equality holds by the definition of $\tau(\theta)$. Next, decompose $T_4(\theta)$ as

$$T_4(\theta) = \frac{1-\pi}{\pi}\mathbb{E}\Big[\Delta Y_{t^*}\omega_0 \;\Big|\; D=0\Big] - \frac{1-\pi}{\pi}\mathbb{E}\Big[m_0(\theta_m)\omega_0 \;\Big|\; D=0\Big] =: T_4^A(\theta) - T_4^B(\theta),$$

which uses that $\omega_0(\theta_\omega) = \omega_0$. Notice that

$$T_4^A(\theta) = \frac{1-\pi}{\pi}\mathbb{E}\Big[\mathbb{E}[\Delta Y_{t^*} \mid X_{t^*}, X_{t^*-1}, Z, D=0]\omega_0 \;\Big|\; D=0\Big]$$

$$= \frac{1-\pi}{\pi}\mathbb{E}\big[m_0\omega_0 \mid D=0\big] = \mathbb{E}[m_0 \mid D=1] = \tau, \tag{S12}$$

where the first equality holds by the law of iterated expectations and because ω_0 is fully determined given (X_{t^*}, X_{t^*-1}, Z), the second equality holds by the definition of m_0, the third equality holds by Lemma S2, and the last equality holds by the definition of τ. Next, it immediately follows from Lemma S2 that

$$T_4^B(\theta) = \mathbb{E}[m_0(\theta_m) \mid D=1] \tag{S13}$$

Combining Equations (S9) to (S13) implies the result. □

SB.2 Asymptotic Normality

The following assumption collects the regularity conditions needed for Proposition 7.

Assumption S2 (Regularity Conditions for DR Estimator)**.**

(i) $\mathbb{E}\big[\varphi(O;\eta)^2\big] < \infty$ *and* $Var\big(\varphi(O;\eta)\big) > 0$.

(ii) $\|\hat{m}_0^{-k} - m_0\|_2 = o_p(1)$, $\|\hat{\nu}_0^{-k} - \nu_0\|_2 = o_p(1)$, $\|\hat{p}^{-k} - p\|_2 = o_p(1)$, *and* $\|\hat{\omega}_0^{-k} - \omega_0\|_2 = o_p(1)$ *for each fold* k.

(iii) $p(X_{t^*-1}, W, Z)$ *is uniformly bounded away from 1 and* $\hat{p}^{-k}(X_{t^*-1}, W, Z)$ *is uniformly bounded away from 1 for each fold* k.

(iv) $\hat{\omega}_0^{-k}$ *is uniformly bounded for each fold* k.

(v) $\mathbb{E}[\Delta Y_{t^*}^2 \mid X_{t^*}, X_{t^*-1}, Z, D=0]$ *is uniformly bounded.*

Proof of Proposition 7. To start with, notice that $\widehat{\mathrm{ATT}}_{dr} = \frac{1}{n}\sum_{i=1}^n \frac{\pi}{\hat{\pi}}\varphi_1(O_i, \hat{\eta}^{-k})$. Then,

$$\begin{aligned}
\sqrt{n}\big(\widehat{\mathrm{ATT}}_{dr} - \mathrm{ATT}\big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(\varphi_1(O_i, \hat{\eta}^{-k}) - \mathrm{ATT}\Big) + \frac{1}{n}\sum_{i=1}^n \varphi_1(O_i, \hat{\eta}^{-k})\sqrt{n}\left(\frac{\pi}{\hat{\pi}} - 1\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(\varphi_1(O_i, \hat{\eta}^{-k}) - \mathrm{ATT}\Big) + \mathrm{ATT}\sqrt{n}\left(\frac{\pi}{\hat{\pi}} - 1\right) + o_p(1) \\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\varphi_1(O_i, \hat{\eta}^{-k}) - \mathrm{ATT}\big)}_{A} - \underbrace{\frac{\mathrm{ATT}}{\pi}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n (D_i - \pi)}_{B} + o_p(1).
\end{aligned}$$

where the first equality holds by adding and subtracting $\frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_1(O_i, \hat{\eta}^{-k})$, the second equality holds because $\frac{1}{n}\sum_{i=1}^n \varphi_1(O_i, \hat{\eta}^{-k})$ is consistent for ATT and by the continuous mapping theorem, and the last equality holds by rearranging terms and the continuous mapping theorem. For term

A, notice that

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\varphi_1(O_i,\hat\eta^{-k})-\text{ATT}\Big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\varphi_1(O_i,\eta)-\text{ATT}\Big)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\varphi_1(O_i,\hat\eta^{-k})-\varphi_1(O_i,\eta)\Big)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\varphi_1(O_i,\eta)-\text{ATT}\Big)+\sqrt{n}\sum_{k=1}^{K}\frac{|\mathcal{I}_k|}{n}\mathbb{E}\left[\Big(\varphi_1(O_i,\hat\eta^{-k})-\varphi_1(O_i,\eta)\Big)\ \Big|\ \hat\eta^{-k}\right]\\
&+\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left\{\Big(\varphi_1(O_i,\hat\eta^{-k})-\varphi_1(O_i,\eta)\Big)-\mathbb{E}\left[\Big(\varphi_1(O_i,\hat\eta^{-k})-\varphi_1(O_i,\eta)\Big)\ \Big|\ \hat\eta^{-k}\right]\right\}\\
&=: T_1+T_2+T_3
\end{aligned}
$$

where the first equality holds by adding and subtracting $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_1(O_i,\eta)$, and the second equality holds by adding and subtracting $\sqrt{n}\sum_{k=1}^{K}\frac{|\mathcal{I}_k|}{n}\mathbb{E}\left[\Big(\varphi_1(O_i,\hat\eta^{-k})-\varphi_1(O_i,\eta)\Big)\ \Big|\ \hat\eta^{-k}\right]$.

We show in Lemma S3 that $T_2=o_p(1)$, and we show in Lemma S4 that $T_3=o_p(1)$. Thus, combining the results above, we have that

$$
\sqrt{n}\big(\widehat{\text{ATT}}_{dr}-\text{ATT}\big)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(O_i)+o_p(1)\xrightarrow{d}\mathcal{N}(0,V_{dr})
$$

with $\varphi(O_i,\eta)$ as defined in the proposition. □

Lemma S3. *Under Assumptions 1, 2, 6, 7 and 9, and S2,*

$$
\sqrt{n}\sum_{k=1}^{K}\frac{|\mathcal{I}_k|}{n}\mathbb{E}\left[\Big(\varphi_1(O_i,\hat\eta^{-k})-\varphi_1(O_i,\eta)\Big)\ \Big|\ \hat\eta^{-k}\right]=o_p(1).
$$

Proof. Define $T_2(k):=\varphi_1(O,\hat\eta^{-k})-\varphi_1(O,\eta)$, so that

$$
\sqrt{n}\sum_{k=1}^{K}\frac{|\mathcal{I}_k|}{n}\mathbb{E}\left[\Big(\varphi_1(O_i,\hat\eta^{-k})-\varphi_1(O_i,\eta)\Big)\ \Big|\ \hat\eta^{-k}\right]=\sqrt{n}\sum_{k=1}^{K}\frac{|\mathcal{I}_k|}{n}\mathbb{E}\Big[T_2(k)\ \Big|\ \hat\eta^{-k}\Big],
$$

For a particular fold k, it holds that

$$
\begin{aligned}
\mathbb{E}\Big[T_2(k)\ \Big|\ \hat\eta^{-k}\Big] &= \mathbb{E}\left[-\frac{D}{\pi}\hat\nu_0^{-k}-\frac{(1-D)}{\pi}\big(\hat m_0^{-k}-\hat\nu_0^{-k}\big)\frac{\hat p^{-k}}{1-\hat p^{-k}}-\frac{(1-D)}{\pi}(\Delta Y_{t^*}-\hat m_0^{-k})\hat\omega_0^{-k}\ \Bigg|\ \hat\eta^{-k}\right]\\
&+\mathbb{E}\left[\frac{D}{\pi}\nu_0+\frac{(1-D)}{\pi}\big(m_0-\nu_0\big)\frac{p}{1-p}+\frac{(1-D)}{\pi}(\Delta Y_{t^*}-m_0)\omega_0\ \Bigg|\ \hat\eta^{-k}\right]\\
&= -\underbrace{\mathbb{E}\left[\frac{D}{\pi}\Big(\hat\nu_0^{-k}-\nu_0\Big)\ \Bigg|\ \eta^{-k}\right]}_{A}\\
&-\underbrace{\mathbb{E}\left[\frac{1-D}{\pi}\Big(\hat m_0^{-k}-m_0\Big)\frac{\hat p^{-k}}{1-\hat p^{-k}}\ \Bigg|\ \eta^{-k}\right]}_{B}-\underbrace{\mathbb{E}\left[\frac{1-D}{\pi}m_0\left(\frac{\hat p^{-k}}{1-\hat p^{-k}}-\frac{p}{1-p}\right)\ \Bigg|\ \eta^{-k}\right]}_{C}
\end{aligned}
$$

$$+\underbrace{\mathbb{E}\left[\frac{1-D}{\pi}\Big(\hat{\nu}_0^{-k}-\nu_0\Big)\frac{\hat{p}^{-k}}{1-\hat{p}^{-k}}\;\Bigg|\;\eta^{-k}\right]}_{D}+\underbrace{\mathbb{E}\left[\frac{1-D}{\pi}\nu_0\left(\frac{\hat{p}^{-k}}{1-\hat{p}^{-k}}-\frac{p}{1-p}\right)\;\Bigg|\;\eta^{-k}\right]}_{E}$$

$$+\underbrace{\mathbb{E}\left[\frac{(1-D)}{\pi}\Big(\hat{m}_0^{-k}-m_0\Big)\hat{\omega}_0^{-k}\;\Bigg|\;\eta^{-k}\right]}_{F}-\underbrace{\mathbb{E}\left[\frac{(1-D)}{\pi}(\Delta Y_{t^*}-m_0)\big(\hat{\omega}_0^{-k}-\omega_0\big)\;\Bigg|\;\eta^{-k}\right]}_{G}$$

which the first equality holds by the definition of $T_2(k)$, and the second equality holds by adding and subtracting terms and rearranging. Terms C and E cancel by the law of iterated expectations. Term $G=0$ by the law of iterated expectations.

$$\begin{aligned}
-A+D &= -\mathbb{E}\Big[\Big(\hat{\nu}_0^{-k}-\nu_0\Big)\;\Big|\;\hat{\eta}^{-k},D=1\Big]+\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{\nu}_0^{-k}-\nu_0\Big)\frac{\hat{p}^{-k}}{1-\hat{p}^{-k}}\;\Bigg|\;\eta^{-k},D=0\right]\\
&= -\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{\nu}_0^{-k}-\nu_0\Big)\frac{p}{1-p}\;\Bigg|\;\eta^{-k},D=0\right]+\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{\nu}_0^{-k}-\nu_0\Big)\frac{\hat{p}^{-k}}{1-\hat{p}^{-k}}\;\Bigg|\;\eta^{-k},D=0\right]\\
&= \frac{1-\pi}{\pi}\mathbb{E}\left[\frac{(\hat{\nu}_0^{-k}-\nu_0)(\hat{p}^{-k}-p)}{(1-p)(1-\hat{p}^{-k})}\;\Bigg|\;\eta^{-k},D=0\right]
\end{aligned}$$

where the first equality holds by the definitions of A and D, the second equality holds by a change of measure argument, and the third equality holds by combining terms and canceling. Next,

$$\begin{aligned}
-B+F &= -\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{m}_0^{-k}-m_0\Big)\frac{\hat{p}^{-k}}{1-\hat{p}^{-k}}\;\Bigg|\;\eta^{-k},D=0\right]+\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{m}_0^{-k}-m_0\Big)\hat{\omega}_0^{-k}\;\Big|\;\eta^{-k},D=0\right]\\
&= -\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{m}_0^{-k}-m_0\Big)\left(\frac{\hat{p}^{-k}}{1-\hat{p}^{-k}}-\frac{p}{1-p}\right)\;\Bigg|\;\eta^{-k},D=0\right]\\
&\quad+\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{m}_0^{-k}-m_0\Big)\Big(\hat{\omega}_0^{-k}-\omega_0\Big)\;\Big|\;\eta^{-k},D=0\right]\\
&= -\frac{1-\pi}{\pi}\mathbb{E}\left[\frac{(\hat{m}_0^{-k}-m_0)(\hat{p}^{-k}-p)}{(1-p)(1-\hat{p}^{-k})}\;\Bigg|\;\eta^{-k},D=0\right]\\
&\quad+\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{m}_0^{-k}-m_0\Big)\Big(\hat{\omega}_0^{-k}-\omega_0\Big)\;\Big|\;\eta^{-k},D=0\right]
\end{aligned}$$

where the first equality holds by the law of iterated expectations, the second equality holds by adding and subtracting $\frac{1-\pi}{\pi}\mathbb{E}\left[\Big(\hat{m}_0^{-k}-m_0\Big)\frac{p}{1-p}\;\Big|\;\eta^{-k},D=0\right]$ and by applying the law of iterated expectations and the definition of ω_0 in the second term, and the last equality holds by rearranging the first term.

The previous results imply that

$$\begin{aligned}
\mathbb{E}\Big[T_2(k)\;\Big|\;\hat{\eta}^{-k}\Big] &= \frac{1-\pi}{\pi}\mathbb{E}\left[\frac{(\hat{\nu}_0^{-k}-\nu_0)(\hat{p}^{-k}-p)}{(1-p)(1-\hat{p}^{-k})}\;\Bigg|\;\eta^{-k},D=0\right]\\
&\quad-\frac{1-\pi}{\pi}\mathbb{E}\left[\frac{(\hat{m}_0^{-k}-m_0)(\hat{p}^{-k}-p)}{(1-p)(1-\hat{p}^{-k})}\;\Bigg|\;\eta^{-k},D=0\right]
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1-\pi}{\pi}\mathbb{E}\left[\left(\hat{m}_0^{-k}-m_0\right)\left(\hat{\omega}_0^{-k}-\omega_0\right)\;\middle|\;\eta^{-k}, D=0\right] \\
&\leq C\Big(||\hat{\nu}_0^{-k}-\nu_0||_2\times||\hat{p}^{-k}-p||_2+||\hat{m}_0^{-k}-m_0||_2\times||\hat{p}^{-k}-p||_2+||\hat{m}_0^{-k}-m_0||_2\times||\hat{\omega}_0^{-k}-\omega_0||_2\Big) \\
&=o_p(n^{-1/2})
\end{aligned}
$$

where the first equality holds by combining the previous results, the inequality on the fourth line holds by Assumptions S2 and 7 and by the Cauchy-Schwarz inequality, and the last line holds by Assumption 9.

Since this result holds for any fold k and $|\mathcal{I}_k|$ is the same order as n, it implies the claim in the lemma. ☐

Lemma S4. *Under Assumptions 1, 2, 6, 7 and 9, and S2,*

$$
\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left\{\left(\varphi_1(O_i,\hat{\eta}^{-k})-\varphi_1(O_i,\eta)\right)-\mathbb{E}\left[\left(\varphi_1(O_i,\hat{\eta}^{-k})-\varphi_1(O_i,\eta)\right)\;\middle|\;\hat{\eta}^{-k}\right]\right\}=o_p(1)
$$

Proof. Define $\dot{\varphi}_{1,i}(k):=\varphi_1(O_i,\hat{\eta}^{-k})-\varphi_1(O_i,\eta)-\mathbb{E}\left[\left(\varphi_1(O_i,\hat{\eta}^{-k})-\varphi_1(O_i,\eta)\right)\;\middle|\;\hat{\eta}^{-k}\right]$, so that

$$
\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left\{\left(\varphi_1(O_i,\hat{\eta}^{-k})-\varphi_1(O_i,\eta)\right)-\mathbb{E}\left[\left(\varphi_1(O_i,\hat{\eta}^{-k})-\varphi_1(O_i,\eta)\right)\;\middle|\;\hat{\eta}^{-k}\right]\right\}=\sum_{k=1}^{K}T_3(k)
$$

where

$$
T_3(k):=\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_k}\dot{\varphi}_{1,i}(k).
$$

For any fold k, notice that $\mathbb{E}[\dot{\varphi}_1(k)\mid\hat{\eta}^{-k}]=0$, which holds immediately, and implies that

$$
\mathbb{E}\left[T_3(k)\mid\hat{\eta}^{-k}\right]=0.
$$

Further, notice that

$$
\begin{aligned}
\operatorname{Var}\big(\dot{\varphi}_1(k)\mid\hat{\eta}^{-k}\big) &\leq \mathbb{E}\Big[\big(\varphi_1(O;\hat{\eta}^{-k})-\varphi_1(O;\eta)\big)^2\;\Big|\;\hat{\eta}^{-k}\Big] \\
&\leq C\big(\|\hat{m}_0-m_0\|_2^2+\|\hat{\nu}_0-\nu_0\|_2^2+\|\hat{p}-p\|_2^2+\|\hat{\omega}_0-\omega_0\|_2^2\big) \\
&=o_p(1) \qquad\qquad (S14)
\end{aligned}
$$

where the first line holds by the definition of variance, the second line holds by adding and subtracting terms and then applying the Cauchy-Schwarz inequality under Assumption S2(iii)-(v), and the last line holds because the nuisance functions are L^2 consistent by Assumption S2(ii).

This implies that

$$\operatorname{Var}(T_3(k) \mid \hat{\eta}^{-k}) = \operatorname{Var}\left(\frac{1}{\sqrt{n}} \sum_{i\in\mathcal{I}_k} \dot{\varphi}_{1,i}(k) \,\middle|\, \hat{\eta}^{-k}\right) = \frac{|\mathcal{I}_k|}{n} \operatorname{Var}(\dot{\varphi}_1(k) \mid \hat{\eta}^{-k}) = o_p(1) \tag{S15}$$

which holds by Assumption 1, because $|\mathcal{I}_k|$ is the same order as n, and by Equation (S14).

Now, for any $\epsilon > 0$ and $\delta > 0$, first note that on the event $\operatorname{Var}\left(T_3(k) \mid \hat{\eta}^{-k}\right) \leq \delta$, by Chebyshev's inequality we have that

$$\mathrm{P}\left(|T_3(k)| > \epsilon \mid \hat{\eta}^{-k}\right) \leq \frac{\operatorname{Var}\left(T_3(k) \mid \hat{\eta}^{-k}\right)}{\epsilon^2} \leq \frac{\delta}{\epsilon^2}. \tag{S16}$$

Then,

$$\begin{aligned}
\mathrm{P}\left(|T_3(k)| > \epsilon\right) &= \mathbb{E}\left[\mathrm{P}\left(|T_3(k)| > \epsilon \mid \hat{\eta}^{-k}\right)\right] \\
&\leq \mathbb{E}\left[\mathrm{P}\left(|T_3(k)| > \epsilon \mid \hat{\eta}^{-k}\right) \mathbf{1}\{\operatorname{Var}(T_3 \mid \hat{\eta}^{-k}) \leq \delta\} + \mathbf{1}\{\operatorname{Var}(T_3 \mid \hat{\eta}^{-k}) > \delta\}\right] \\
&\leq \frac{\delta}{\epsilon^2}\mathrm{P}\Big(\operatorname{Var}(T_3 \mid \hat{\eta}^{-k}) \leq \delta\Big) + \mathrm{P}\Big(\operatorname{Var}(T_3 \mid \hat{\eta}^{-k}) > \delta\Big) \\
&\leq \frac{\delta}{\epsilon^2} + o(1)
\end{aligned}$$

where the first equality holds by the law of iterated expectations, the second line holds because the indicator terms sum to one and probabilities are bounded below one, the third line holds by Equation (S16), and the fourth inequality holds by Equation (S15). Since $\delta > 0$ is arbitrary, taking $\delta \to 0$ implies that $\mathrm{P}\big(|T_3(k)| > \epsilon\big) \to 0$. Applying this argument for all $k = 1, \ldots, K$ implies $T_3 = o_p(1)$. □

SB.3 Asymptotic Normality with Staggered Adoption

Appendices SA and SB showed that the imputation and Neyman Orthogonal estimators discussed in the main text are consistent and asymptotically normal in the setting with two time periods. This section discusses how to extend those results to the staggered adoption setting considered in Section 5. We focus on the estimator implied by the identification result in Proposition 2 as we think this is the most likely result to be used in empirical work. Estimating $\mathrm{ATT}(g,t)$ based on this expression is as simple as replacing $(Y_{it^*} - Y_{it^*-1})$ with $(Y_{it} - Y_{ig-1})$, X_{it^*} with X_{it}, and X_{it^*-1} with X_{ig-1} in the preceding estimators. Consistency and asymptotic normality follow from effectively the same arguments as above and involve the same influence function and asymptotic variance up to adjusting the periods, which, following the notation used in the main text, we denote by $\psi_{g,t}(O_i)$ and $\varphi_{g,t}(O_i)$, respectively.

Inference for aggregated parameters like $\mathrm{ATT}^{es}(e)$ and ATT^{o}, which were discussed in Remark 4, is also straightforward. We only provide a sketch of these arguments as they follow effectively immediately from existing results in Callaway and Sant'Anna (2021). Notice that all of these parameters can be expressed as

$$\theta = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{\mathrm{T}} w_\theta(g,t) \mathrm{ATT}(g,t),$$

where θ generically denotes some aggregated parameter that depends on the weighting function $w_\theta(g,t)$. For example, for $\mathrm{ATT}^{es}(e)$, the weights are given by $\mathbf{1}\{t = g + e\}\mathrm{P}(G = g \mid G + e \leq \mathrm{T})$. For ATT^o, the weights are given by $\mathrm{P}(G = g \mid G \in \bar{\mathcal{G}})/(\mathrm{T} - g + 1)$. One additional challenge in this case is that the weights generally need to be estimated. Thus, an estimator for an aggregated parameter is given by

$$\hat{\theta} = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{\mathrm{T}} \hat{w}_\theta(g,t) \widehat{\mathrm{ATT}}(g,t).$$

We assume that, for all $g \in \bar{\mathcal{G}}$ and $t = g, \ldots, \mathrm{T}$,

$$\sqrt{n}\big(\hat{w}_\theta(g,t) - w_\theta(g,t)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi^w_{g,t}(O_i) + o_p(1),$$

where $\mathbb{E}[\xi^w_{g,t}] = 0$ and $\mathrm{Var}(\xi^w_{g,t})$ is positive definite and finite. This is a mild high-level assumption about estimating the weights that is satisfied for all the weights considered in Callaway and Sant'Anna (2021) under mild regularity conditions. Under this condition, it follows that

$$\sqrt{n}\big(\hat{\theta} - \theta\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_\theta(O_i) + o_p(1) \xrightarrow{d} \mathcal{N}\big(0, \mathbb{E}\big[\xi_\theta(O)^2\big]\big)$$

where

$$\xi_\theta(O) := \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{\mathrm{T}} \Big(w_\theta(g,t)\psi_{g,t}(O) + \xi^w_{g,t}(O)\mathrm{ATT}(g,t) \Big)$$

The expression above is for the influence function from the imputation estimator, but the same argument applies for the Neyman Orthogonal estimator with $\varphi_{g,t}$ replacing $\psi_{g,t}$.

SC Identification under Parallel Trends for the Bad Control

In this section, we consider identification of the ATT under a parallel trends assumption for the bad control. We focus on the case with two periods, as the extension to staggered treatment adoption follows along the same lines as the arguments in Section 5 in the main text. Consider the following assumption.

Assumption S3 (Bad Control Parallel Trends)**.**

$$\mathbb{E}[\Delta X_{t^*}(0) \mid W, Z, D = 1] = \mathbb{E}[\Delta X_{t^*}(0) \mid W, Z, D = 0]$$

Under Assumption S3, it is straightforward to see that the average effect of the treatment on the treated of the bad control is identified. In particular, it follows from standard difference-in-differences arguments that

$$\text{ATT}_X := \mathbb{E}[X_{t^*}(1) - X_{t^*}(0) \mid D = 1] = \mathbb{E}[\Delta X_{t^*} \mid D = 1] - \mathbb{E}\Big[\mathbb{E}[\Delta X_{t^*} \mid W, Z, D = 0] \Bigm| D = 1\Big]. \tag{S17}$$

However, identifying ATT_X is not sufficient for identifying ATT as it depends on the entire distribution of $X_{t^*}(0)$ for the treated group (see Equation (3) in the main text). That being said, what we show next is that if we additionally impose linearity on the parallel trends assumption in Assumption 2, then we can recover the ATT under Assumption S3. Consider the following assumption:

Assumption S4 (Linear Conditional Expection of $\Delta Y_{t^*}(0)$)**.**

$$\mathbb{E}[\Delta Y_{t^*}(0) \mid X_{t^*}(0), X_{t^*-1}, Z] = X_{t^*}(0)\beta_1 + X_{t^*-1}\beta_2 + Z'\beta_3$$

$(\beta_1, \beta_2, \beta_3)$ are identified from the regression regression of ΔY_{t^*} on X_{t^*}, X_{t^*-1}, and Z among the untreated group. This is the same assumption as we used for the imputation estimator discussed in the main text. Then, we have the following identification result:

Proposition S1. *Under Assumptions 1, 2, 7, S3, and S4,*

$$\begin{aligned} ATT = \mathbb{E}[\Delta Y_{t^*} \mid D = 1] - \Big(\mathbb{E}\big[\mathbb{E}[\Delta X_{t^*} \mid W, Z, D = 0] \bigm| D = 1\big]\beta_1 \\ + \mathbb{E}[X_{t^*-1} \mid D = 1](\beta_2 + \beta_1) + \mathbb{E}[Z \mid D = 1]'\beta_3\Big), \end{aligned}$$

which is identified.

Proof. Notice that

$$\begin{aligned} \mathbb{E}[\Delta Y_{t^*}(0) \mid D = 1] &= \mathbb{E}\Big[\mathbb{E}[\Delta Y_{t^*}(0) \mid X_{t^*}(0), X_{t^*-1}, Z, D = 0] \Bigm| D = 1\Big] \\ &= \mathbb{E}\Big[X_{t^*}(0)\beta_1 + X_{t^*-1}\beta_2 + Z'\beta_3 \Bigm| D = 1\Big] \\ &= \mathbb{E}\Big[\Delta X_{t^*}(0) \Bigm| D = 1\Big]\beta_1 + \mathbb{E}[X_{t^*-1} \mid D = 1](\beta_2 + \beta_1) + \mathbb{E}[Z \mid D = 1]'\beta_3 \\ &= \mathbb{E}\Big[\mathbb{E}[\Delta X_{t^*} \mid W, Z, D = 0] \Bigm| D = 1\Big]\beta_1 + \mathbb{E}[X_{t^*-1} \mid D = 1](\beta_2 + \beta_1) + \mathbb{E}[Z \mid D = 1]'\beta_3 \end{aligned}$$

where the first equality holds by Assumption 2 and the same argument used to prove Theorems 1 and 2, the second equality holds by Assumption S4, the third equality holds by adding and subtracting $\mathbb{E}[X_{t^*-1} \mid D=1]\beta_1$ and rearranging, and the last equality holds by the same argument as in Equation (S17). Plugging this expression into the definition of ATT implies the result. □

To conclude this section, we add one more linearity assumption and derive a corollary of Proposition S1.

Assumption S5 (Linear Conditional Expection of $\Delta X_{t^*}(0)$)**.**

$$\mathbb{E}[\Delta X_{t^*}(0) \mid W, Z] = W'\delta_1 + Z'\delta_2$$

Notice that, under Assumption S3, (δ_1, δ_2) are identified from the regression of ΔX_{t^*} on W, Z using the untreated group.

Corollary S1. *Under Assumptions 1, 2, 7, S3, S4, and S5,*

$$ATT = \mathbb{E}[\Delta Y_{t^*} \mid D=1] - \Big(\mathbb{E}[W \mid D=1]'\delta_1\beta_1 + \mathbb{E}[X_{t^*-1} \mid D=1](\beta_2+\beta_1) \\ + \mathbb{E}[Z \mid D=1]'(\delta_2\beta_1+\beta_3)\Big),$$

which is identified.

Proof. The result holds immediately by plugging the expression in Assumption S5 into the result of Proposition S1. □

The expression in Corollary S1 suggests a straightforward imputation estimator of the ATT. In a first step, estimate $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ and $(\hat{\delta}_1, \hat{\delta}_2)$ from regressing ΔY_{t^*} on (X_{t^*}, X_{t^*-1}, Z) and ΔX_{t^*} on (W, Z), both using the untreated group. Then, estimate

$$\widehat{\text{ATT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{\hat{\pi}}\Big(\Delta Y_{i,t^*} - W_i'\hat{\delta}_1\hat{\beta}_1 - X_{i,t^*-1}(\hat{\beta}_2+\hat{\beta}_1) - Z_i'(\hat{\delta}_2\hat{\beta}_1+\hat{\beta}_3)\Big)$$

SD Monte Carlo Simulations

In this section, we conduct a Monte Carlo study to evaluate the finite-sample performance of the estimators discussed in the main text. We focus on the setting with two time periods and consider five different data generating processes (DGPs) below. The DGPs mainly differ in terms of the model for the bad control, so we start by describing their common structure first. Let $Z_i \sim \mathcal{N}(0,1)$ and $\eta_i \sim \mathcal{N}(0,1)$ be mutually independent, with η_i an unobserved heterogeneity term, and let $\varepsilon_i^W, \varepsilon_i^D, \varepsilon_i^{X_1}, \varepsilon_i^{X_2}, \varepsilon_{i1}^Y, \varepsilon_{i2}^Y \sim \mathcal{N}(0,1)$ be mutually independent noise terms, independent of (Z_i, η_i). From these primitives, we construct the pre-treatment covariate $W_i = 0.8\eta_i + 0.3Z_i + 0.2\varepsilon_i^W$. The treatment is assigned as

$$D_i = \mathbf{1}\{0.2Z_i + 0.4W_i + 0.3\eta_i + \varepsilon_i^D > 0\}.$$

The pre-treatment bad control is $X_{i1} = 0.5\eta_i + 0.4Z_i + 0.3\varepsilon_i^{X_1}$, and treatment shifts the bad control by $\lambda = 0.5$ units, so that the treated potential covariate is $X_{i2}(1) = X_{i2}(0) + \lambda$. The untreated potential outcome $Y_{it}(0)$ is linear in $X_{it}(0)$ across all DGPs,

$$Y_{it}(0) = \theta_t + 0.5\eta_i + 0.3Z_i + X_{it}(0) + 0.3\varepsilon_{it}^Y,$$

with $\theta_1 = 0.3$ and $\theta_2 = 0.6$. We set the treated potential outcome at $t = t^*$ to be $Y_{it^*}(1) = Y_{it^*}(0) + \big(X_{it^*}(1) - X_{it^*}(0)\big) + \delta$, where δ is the direct effect of the treatment, which we set at $\delta = 0.5$. Thus, the true ATT $= \delta + \lambda = 1.00$. Next, we describe the differences across DGPs.

`DGP 1 --- Covariate Unconfoundedness, Linear W`

$$X_{i2}(0) = 0.7X_{i1} + 0.3Z_i + 0.2W_i + 0.15 + 0.3\varepsilon_i^{X_2}.$$

`DGP 2 --- Covariate Unconfoundedness, Nonlinear W`

$$X_{i2}(0) = 0.7X_{i1} + 0.3Z_i + 0.2W_i + 0.03W_i^2 + 0.15 + 0.3\varepsilon_i^{X_2}.$$

`DGP 3 --- Simple Covariate Unconfoundedness, Nonlinear X`

$$X_{i2}(0) = 0.7X_{i1} + 0.3Z_i + 0.4X_{i1}Z_i + 0.2X_{i1}^2 + 0.15 + 0.3\varepsilon_i^{X_2}.$$

`DGP 4 --- Bad Control Parallel Trends`

$$X_{i2}(0) = X_{i1} + 0.3Z_i + 0.2W_i + 0.15 + 0.3\varepsilon_i^{X_2}.$$

`DGP 5 --- Covariate Unconfoundedness, Nonlinear W, X/W Interaction`

$$X_{i2}(0) = 0.7X_{i1} + 0.3Z_i + 0.2W_i + 0.03W_i^2 + 0.05X_{i1}W_i + 0.15 + 0.3\varepsilon_i^{X_2}.$$

We use seven of the nine estimators that were discussed in detail in the application in the main text: `TWFE: include BC`, `TWFE: exclude BC`, `PT for X`, `ML: pre-treatment`, `Imputation`, `DR`, `ML`, and exclude `Imp. include BC` and `Imp. exclude BC` for concisness. Table S1 summarizes which estimators are consistent for which DGPs.

Even for estimators that are not generally consistent for a given DGP, it is still interesting to check their performance. For example, the functional form assumptions for the imputation and

Table S1: Theoretical Consistency of Each Estimator, by DGP

Method	dgp1	dgp2	dgp3	dgp4	dgp5
TWFE Include BC	×	×	×	×	×
TWFE Exclude BC	×	×	×	×	×
PT for X	×	×	×	✓	×
ML (Pre-treatment)	×	×	✓	×	×
Imputation	✓	×	×	✓	×
DR	✓	×	×	✓	×
ML	✓	✓	✓	✓	✓

doubly robust estimator make them theoretically not consistent for the nonlinear DGPs, but they could still potentially have comparable or better performance than the machine learning estimator in finite samples. Similarly, the assumptions that rationalize `PT for X` and `ML: pre-treatment` do not hold in most DGPs, though they could still potentially perform relatively well. Finally, it is also interesting to compare the finite sample performance across different consistent estimators in DGPs (e.g., 1 and 4) where several estimators are consistent.

In the results below, for each combination of estimator, sample size, and DGP, we report bias ($\widehat{\text{ATT}} - \text{ATT}$), standard deviation across replications (SD), root mean squared error (RMSE), the ratio of mean standard error to SD (SE/SD), and empirical coverage of nominal 95% confidence intervals.[1] All results provided below are based on 1000 Monte Carlo simulations. We provide these results in Figure S1 and Tables S2 to S6.

The results are in line with the theoretical properties of different estimators that we have discussed in the paper. Mainly, in line with the discussion in the text, traditional approaches that directly include the bad control can perform extremely poorly (`TWFE: include BC`, all five DGPs) as well as approaches that exclude the bad control but make no other adjustment (`TWFE: exclude BC`, DGPs 3 and 4).[2] Next, even in DGPs where they are not consistent, `PT for X` and `ML: pre-treatment` perform decently well. We conjecture that in most realistic applications, the alternative assumptions imposed by these approaches are likely to, in some sense, be closer to holding than assumptions that lead to fully including or excluding the bad control. Finally, among our estimators that are based on covariate unconfoundedness (`Imputation`, `DR`, and `ML`), all of them perform decently well across DGPs, including the nonlinear DGPs that violate the functional form assumptions needed by `Imputation` and `DR`. For some DGPs, `ML`, which is consistent for all of the DGPs considered here, sometimes undercovers. This is driven by standard errors that are sometimes too small relative to the standard deviation of the estimator. Finally, even in the DGPs where

[1] An SE/SD ratio near 1.0 indicates that the analytical standard errors are well calibrated, and coverage near 0.95 indicates that the confidence intervals have correct size.

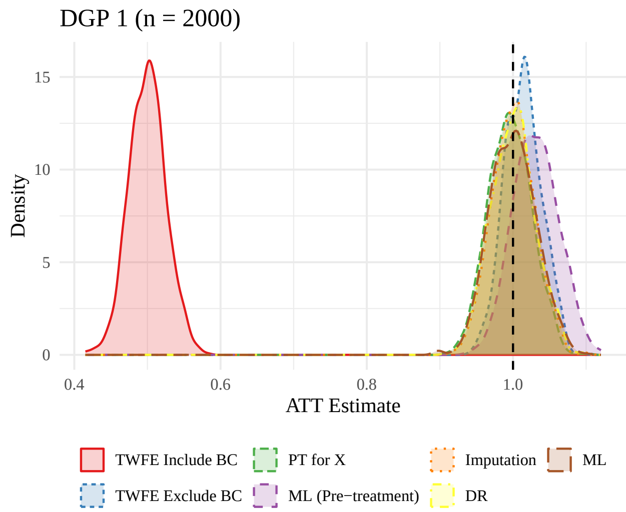
[2] That excluding the bad control performs fairly well in some DGPs is an artifact of the parameters of these DGPs, and it would be straightforward to make this approach perform poorly by simply changing the parameters in DGPs 1, 2, and 5. It is also noteworthy that this approach is biased in all five DGPs, which results in coverage rates shrinking in the sample size.

its functional form assumptions do not hold, the `DR` estimator performs on par with or sometimes slightly better than the `ML` estimator.

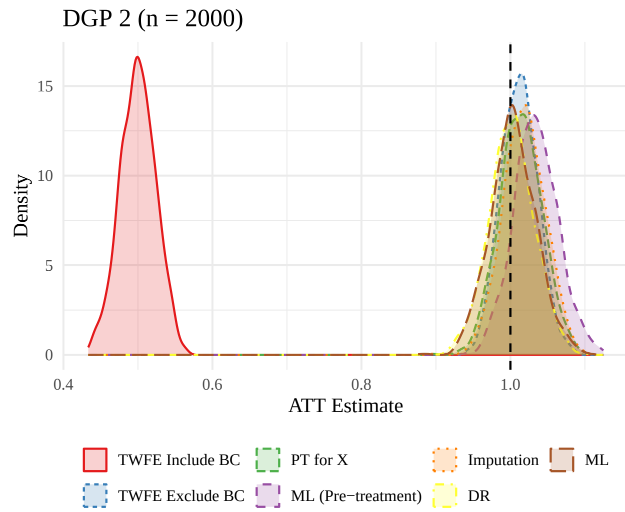
References

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-differences with multiple time periods". *Journal of Econometrics* 225.2, pp. 200–230.

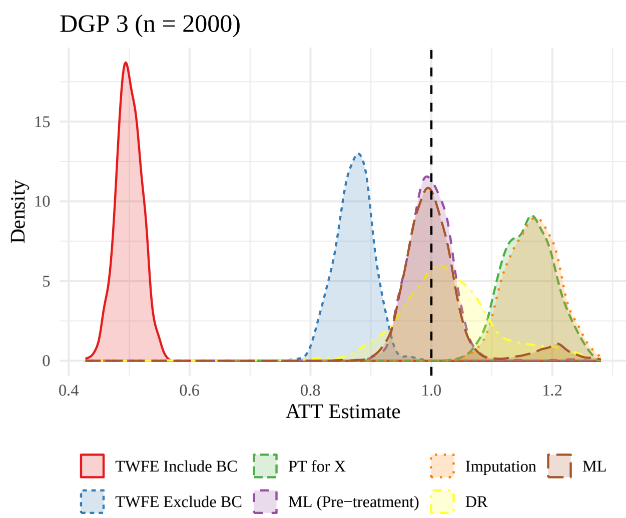
Figure S1: Monte Carlo estimate densities across DGPs



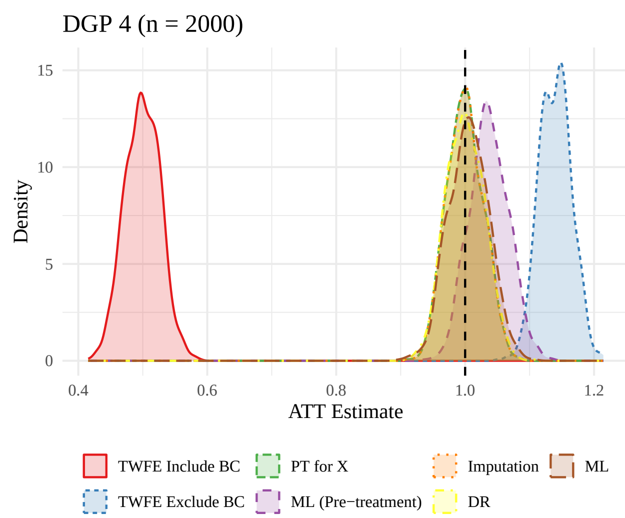
(a) DGP 1



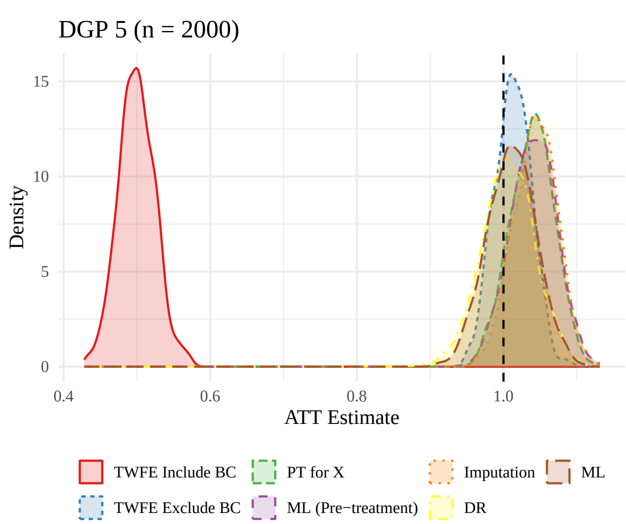
(b) DGP 2



(c) DGP 3



(d) DGP 4



(e) DGP 5

Notes: The figures provides kernel density estimates of $\widehat{\text{ATT}}$ across 1,000 replications at $n = 2,000$, for the four main DGPs. The dashed vertical line marks the true ATT $= 1.00$.

Table S2: Monte Carlo Results: DGP 1

Method	n	Bias	SD	RMSE	SE/SD	Coverage
TWFE Include BC	500	-0.499	0.051	0.501	0.96	0.000
TWFE Include BC	1000	-0.500	0.035	0.502	0.98	0.000
TWFE Include BC	2000	-0.500	0.025	0.501	0.98	0.000
TWFE Exclude BC	500	0.013	0.049	0.050	0.99	0.944
TWFE Exclude BC	1000	0.012	0.034	0.036	1.01	0.934
TWFE Exclude BC	2000	0.014	0.025	0.028	0.97	0.890
PT for X	500	-0.004	0.057	0.057	0.99	0.948
PT for X	1000	-0.006	0.040	0.040	1.01	0.939
PT for X	2000	-0.004	0.029	0.029	0.98	0.946
ML (Pre-treatment)	500	0.021	0.056	0.060	0.90	0.905
ML (Pre-treatment)	1000	0.024	0.041	0.047	0.90	0.874
ML (Pre-treatment)	2000	0.031	0.031	0.044	0.84	0.741
Imputation	500	-0.000	0.056	0.056	1.00	0.945
Imputation	1000	-0.001	0.039	0.039	1.01	0.943
Imputation	2000	0.000	0.029	0.028	0.98	0.944
DR	500	-0.000	0.059	0.059	1.01	0.951
DR	1000	-0.002	0.041	0.041	1.01	0.948
DR	2000	0.000	0.030	0.030	0.97	0.940
ML	500	0.003	0.057	0.057	0.86	0.921
ML	1000	-0.001	0.041	0.041	0.87	0.893
ML	2000	0.000	0.031	0.031	0.82	0.897

Notes: The table provides Monte Carlo results for DGP 1. The rows are organized by the seven estimators discussed in the main text. Results are based on 1000 Monte Carlo simulations.

Table S3: Monte Carlo Results: DGP 2

Method	n	Bias	SD	RMSE	SE/SD	Coverage
TWFE Include BC	500	-0.496	0.050	0.499	0.99	0.000
TWFE Include BC	1000	-0.499	0.034	0.500	1.03	0.000
TWFE Include BC	2000	-0.500	0.024	0.501	1.04	0.000
TWFE Exclude BC	500	0.016	0.049	0.051	0.99	0.940
TWFE Exclude BC	1000	0.012	0.035	0.037	0.98	0.938
TWFE Exclude BC	2000	0.012	0.025	0.027	0.99	0.917
PT for X	500	0.018	0.057	0.060	1.00	0.938
PT for X	1000	0.014	0.041	0.043	1.00	0.941
PT for X	2000	0.014	0.029	0.032	1.00	0.925
ML (Pre-treatment)	500	0.030	0.056	0.064	0.90	0.883
ML (Pre-treatment)	1000	0.031	0.042	0.052	0.87	0.820
ML (Pre-treatment)	2000	0.034	0.030	0.045	0.88	0.723
Imputation	500	0.023	0.055	0.060	1.01	0.933
Imputation	1000	0.019	0.040	0.044	1.00	0.925
Imputation	2000	0.018	0.028	0.034	0.99	0.897
DR	500	0.003	0.061	0.061	1.01	0.953
DR	1000	0.001	0.043	0.043	0.98	0.952
DR	2000	0.002	0.030	0.030	0.98	0.936
ML	500	0.012	0.057	0.058	0.87	0.904
ML	1000	0.006	0.042	0.043	0.84	0.901
ML	2000	0.004	0.030	0.031	0.84	0.894

Notes: The table provides Monte Carlo results for DGP 2. The rows are organized by the seven estimators discussed in the main text. Results are based on 1000 Monte Carlo simulations.

Table S4: Monte Carlo Results: DGP 3

Method	n	Bias	SD	RMSE	SE/SD	Coverage
TWFE Include BC	500	-0.502	0.042	0.504	0.98	0.000
TWFE Include BC	1000	-0.499	0.029	0.500	1.00	0.000
TWFE Include BC	2000	-0.500	0.021	0.501	0.97	0.000
TWFE Exclude BC	500	-0.129	0.060	0.142	1.00	0.402
TWFE Exclude BC	1000	-0.128	0.043	0.135	0.98	0.151
TWFE Exclude BC	2000	-0.128	0.030	0.131	0.98	0.011
PT for X	500	0.164	0.079	0.182	1.02	0.489
PT for X	1000	0.160	0.059	0.171	0.97	0.210
PT for X	2000	0.161	0.041	0.166	0.99	0.018
ML (Pre-treatment)	500	0.005	0.064	0.064	0.95	0.933
ML (Pre-treatment)	1000	0.001	0.048	0.048	0.90	0.928
ML (Pre-treatment)	2000	-0.002	0.036	0.036	0.83	0.929
Imputation	500	0.171	0.080	0.189	1.02	0.462
Imputation	1000	0.168	0.060	0.178	0.97	0.187
Imputation	2000	0.168	0.041	0.173	1.00	0.016
DR	500	0.013	0.133	0.133	0.92	0.897
DR	1000	0.016	0.101	0.102	0.85	0.878
DR	2000	0.028	0.082	0.086	0.74	0.829
ML	500	0.004	0.066	0.066	0.92	0.927
ML	1000	0.002	0.059	0.059	0.73	0.905
ML	2000	0.011	0.064	0.065	0.48	0.848

Notes: The table provides Monte Carlo results for DGP 3. The rows are organized by the seven estimators discussed in the main text. Results are based on 1000 Monte Carlo simulations.

Table S5: Monte Carlo Results: DGP 4

Method	n	Bias	SD	RMSE	SE/SD	Coverage
TWFE Include BC	500	-0.498	0.053	0.501	0.99	0.000
TWFE Include BC	1000	-0.501	0.038	0.503	0.99	0.000
TWFE Include BC	2000	-0.500	0.027	0.501	0.97	0.000
TWFE Exclude BC	500	0.142	0.051	0.151	0.97	0.185
TWFE Exclude BC	1000	0.138	0.035	0.143	0.99	0.018
TWFE Exclude BC	2000	0.140	0.025	0.142	1.00	0.000
PT for X	500	0.003	0.056	0.056	1.00	0.947
PT for X	1000	-0.002	0.039	0.039	1.02	0.955
PT for X	2000	-0.000	0.028	0.028	1.01	0.958
ML (Pre-treatment)	500	0.046	0.058	0.074	0.89	0.827
ML (Pre-treatment)	1000	0.037	0.041	0.055	0.91	0.794
ML (Pre-treatment)	2000	0.037	0.030	0.048	0.88	0.690
Imputation	500	0.003	0.056	0.056	1.00	0.945
Imputation	1000	-0.002	0.038	0.038	1.03	0.956
Imputation	2000	-0.000	0.028	0.028	1.01	0.957
DR	500	0.003	0.062	0.062	0.97	0.945
DR	1000	-0.002	0.040	0.040	1.03	0.958
DR	2000	-0.000	0.029	0.029	0.99	0.956
ML	500	0.027	0.059	0.064	0.87	0.872
ML	1000	0.011	0.042	0.044	0.86	0.906
ML	2000	0.005	0.031	0.031	0.84	0.901

Notes: The table provides Monte Carlo results for DGP 4. The rows are organized by the seven estimators discussed in the main text. Results are based on 1000 Monte Carlo simulations.

Table S6: Monte Carlo Results: DGP 5

Method	n	Bias	SD	RMSE	SE/SD	Coverage
TWFE Include BC	500	-0.498	0.048	0.500	1.02	0.000
TWFE Include BC	1000	-0.500	0.034	0.501	1.01	0.000
TWFE Include BC	2000	-0.500	0.025	0.501	0.98	0.000
TWFE Exclude BC	500	0.014	0.048	0.050	1.02	0.943
TWFE Exclude BC	1000	0.012	0.034	0.036	1.01	0.932
TWFE Exclude BC	2000	0.013	0.025	0.028	0.98	0.913
PT for X	500	0.041	0.058	0.071	0.99	0.887
PT for X	1000	0.039	0.041	0.056	1.01	0.846
PT for X	2000	0.039	0.029	0.049	0.98	0.717
ML (Pre-treatment)	500	0.039	0.059	0.070	0.87	0.834
ML (Pre-treatment)	1000	0.039	0.041	0.057	0.89	0.780
ML (Pre-treatment)	2000	0.041	0.031	0.051	0.86	0.620
Imputation	500	0.046	0.057	0.074	0.99	0.860
Imputation	1000	0.044	0.040	0.059	1.00	0.814
Imputation	2000	0.044	0.029	0.052	0.98	0.640
DR	500	0.006	0.065	0.066	0.97	0.932
DR	1000	0.004	0.045	0.045	1.00	0.947
DR	2000	0.006	0.036	0.036	0.88	0.902
ML	500	0.021	0.059	0.062	0.85	0.879
ML	1000	0.014	0.042	0.044	0.86	0.884
ML	2000	0.011	0.033	0.035	0.79	0.865

Notes: The table provides Monte Carlo results for DGP 5. The rows are organized by the seven estimators discussed in the main text. Results are based on 1000 Monte Carlo simulations.