

Efficient Difference-in-Differences and Event Study Estimators

Xiaohong Chen
Yale University

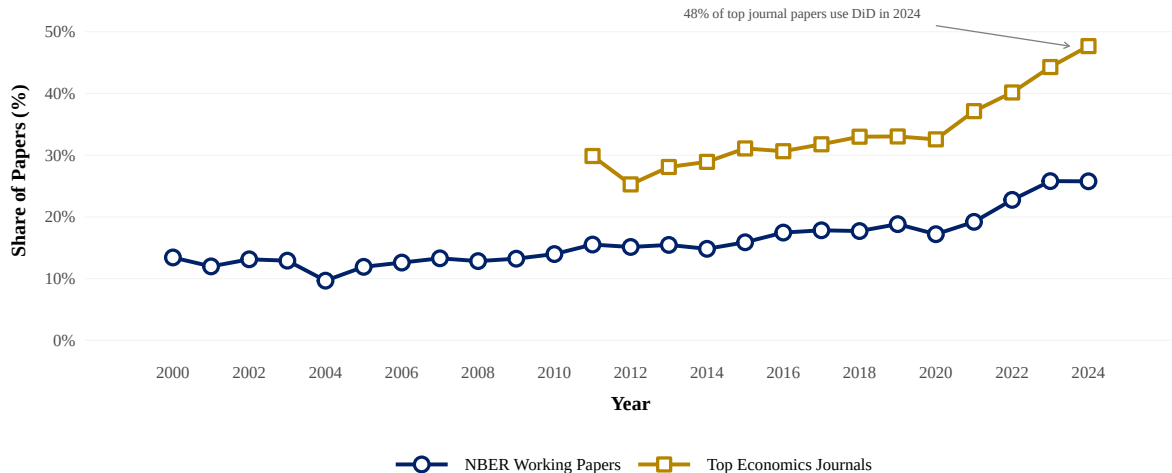
Pedro H. C. Sant'Anna
Emory University

Haitian Xie
Peking University

CESG
October 2026

The Rise of Difference-in-Differences in Economics

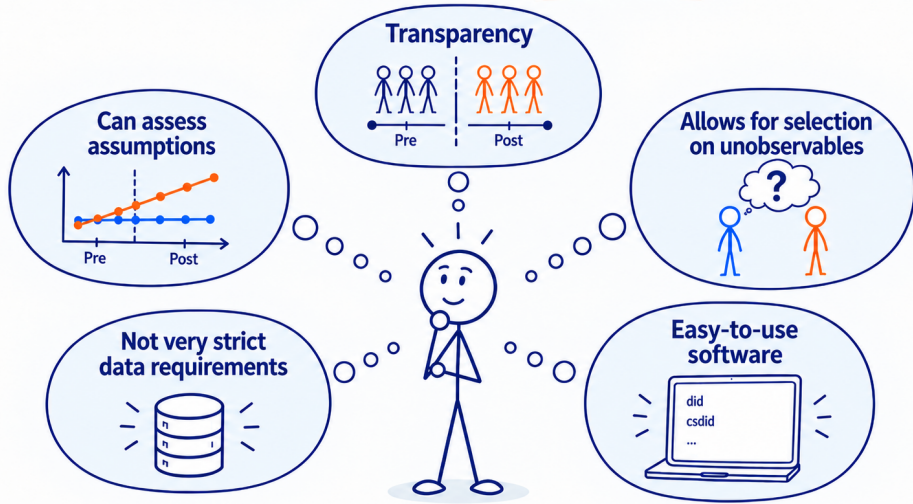
Top Journals: AER, QJE, JPE, AEJ:Applied, AEJ:Policy



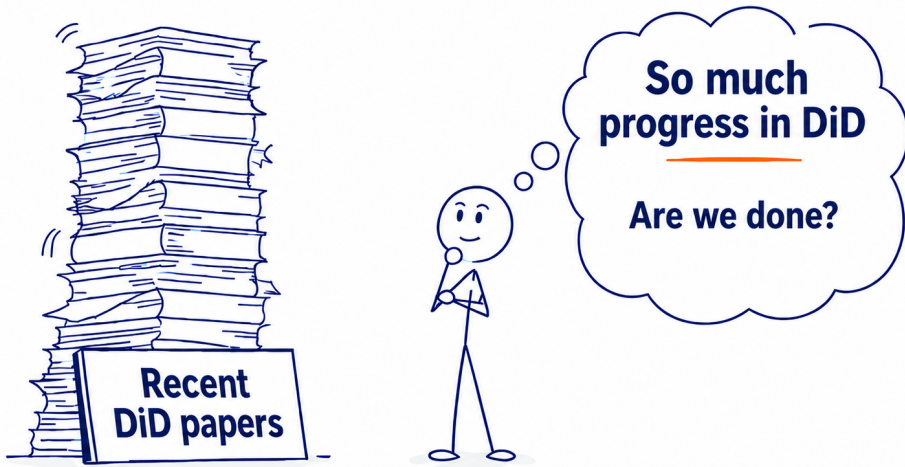
Source: Goldsmith-Pinkham (2024), "Tracking the Credibility Revolution across Fields." Data: paulgp.com/econlit-pipeline

Search terms: "Difference-in-Differences" | "Difference-in-Difference" | "DiD" | "DD" | "D-i-D" | "Difference in Differences" | "Difference in Difference" | "Diff in Diff" | "Diff-in-Diff"

Why is DiD so popular?

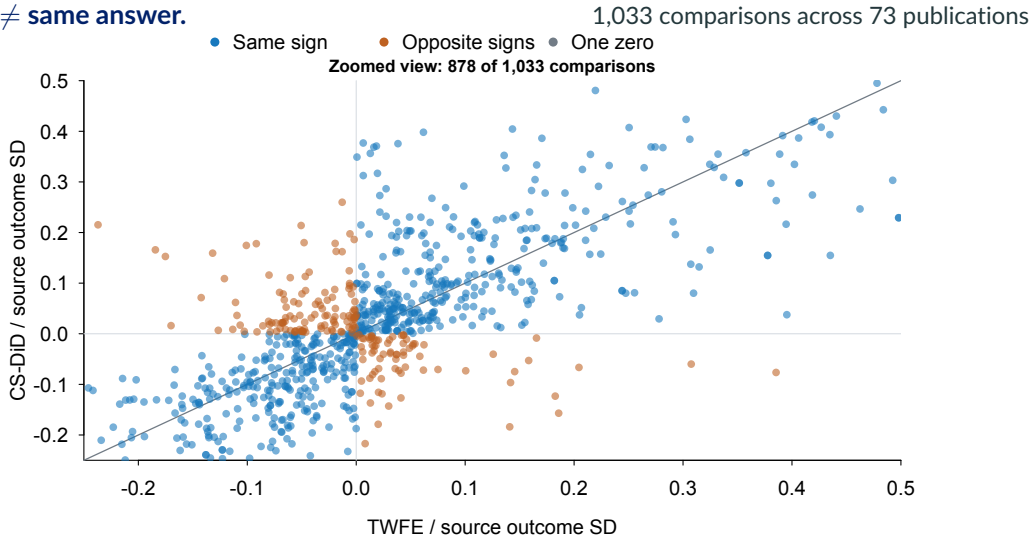


DiD has evolved so much in recent years...



Do these advances actually matter? TWFE vs CSDiD

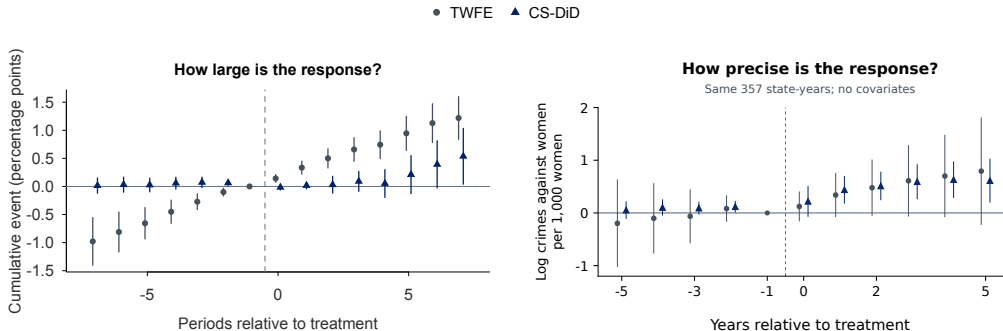
Same sign $\neq$ same answer.



Source: ongoing exercise with Victor Rangel; Rangel (2026) and Peter Hull's October 9 collection.

Would you tell the same policy story?

A scalar average can hide differences in the dynamic path.



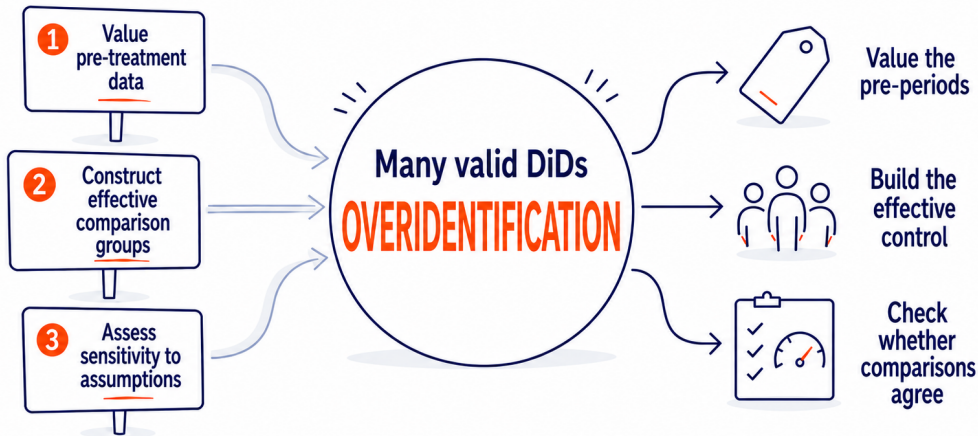
Pointwise 95% intervals; original clustering, supplied frame and reference $e = -1$; no added covariates.

Left: additional application outside the 73-paper inventory; not-yet-treated CS vs. reconstructed unstacked TWFE.

But so many basic questions **remain**



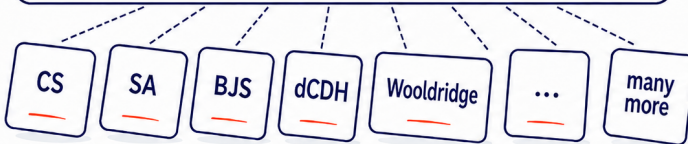
DiD is often nonparametrically overidentified



This paper leverages NP overidentification to better understand DiD

What does this mean?

**Under (conditional) PTs,
many different DiD estimators
can identify the same effect**

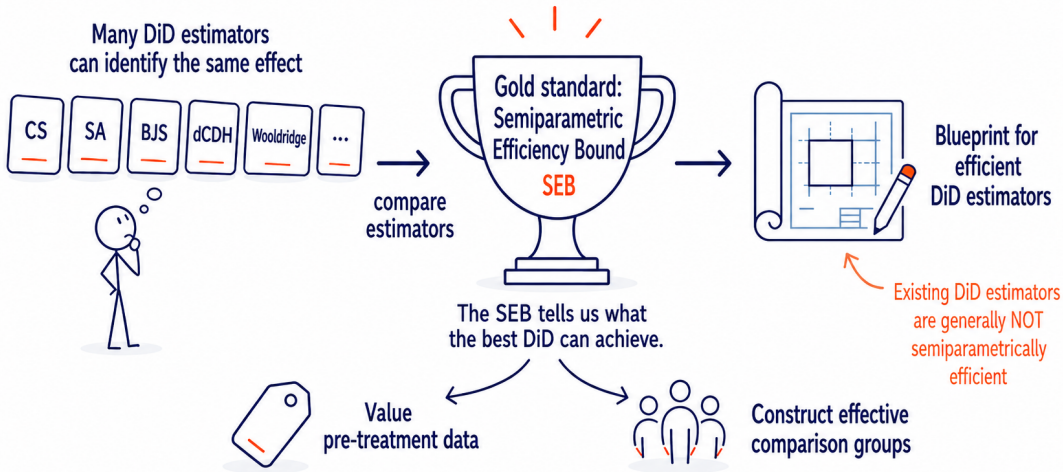


The familiar procedures are only
some of the possibilities



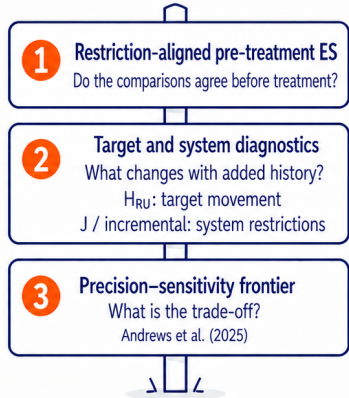
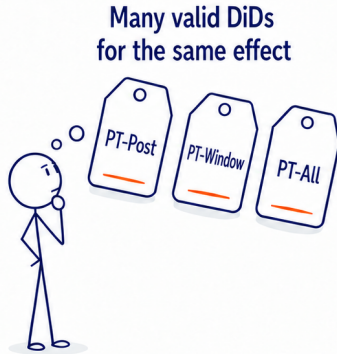
How should we choose between the many options?

We need a gold standard for comparison



—— For block and staggered adoption; with and without covariates; and for unit or clustered sampling. ——

Overidentification also gives us diagnostics



Overidentification helps us assess the coherence of our design.

These tools do not prove parallel trends; they help us understand agreement and sensitivity.

What the 9 applications reveal



1 Precision gains vary



0.88–4.43

Fitted variance ratios
Post/Window and Post/All
across 8 other applications

Different maintained histories;
not same-model efficiency ratios

2 Different tests, different questions



8 history ladders:
no scalar rejection at 5%
(Holm within application)
Some pre-treatment tests
reject zero

Precision and credibility
are not the same thing.

3 Diagnostics keep them apart



Pre-treatment ES



Directed comparisons



Precision-sensitivity
frontier

Non-rejection does not
establish parallel trends.

Across very different designs, the framework **separates** precision gains from sensitivity.

A taste of the econometrics

- “Short” panel $\{W_i\}_{i=1}^n = \{(Y_{i,1}, \dots, Y_{i,T}, X_i', G_i)'\}_{i=1}^n$: n large, T finite and fixed.
- Treatment is binary, an **absorbing state**, with possibly different starting dates.
- $Y_{i,t}(g)$: potential outcome for unit i at time t if first treated in period g .
- G_i : period unit i is first treated, with $G_i = \infty$ if never treated by T ; $G_i \in \mathcal{G} \subseteq \{2, \dots, T, \infty\}$.
 - ▶ Single date: $G_i = g$ (treated) or $G_i = \infty$ (untreated). Let $G_g = 1\{G = g\}$.

- Group-time average treatment effects on the treated:

$$ATT(g, t) := \mathbb{E}[Y_t(g) - Y_t(\infty) \mid G = g].$$

- Aggregate over cohorts into event-study summaries ($e = t - g$):

$$\begin{aligned} ES(e) &:= \mathbb{E}[ATT(G, G + e) \mid G + e \in [1, T]], \\ ES_{\text{avg}} &:= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} ES(e). \end{aligned}$$

Maintained Assumptions Today: Sampling, Overlap, and No-anticipation

Assumption (Maintained Assumption (M))

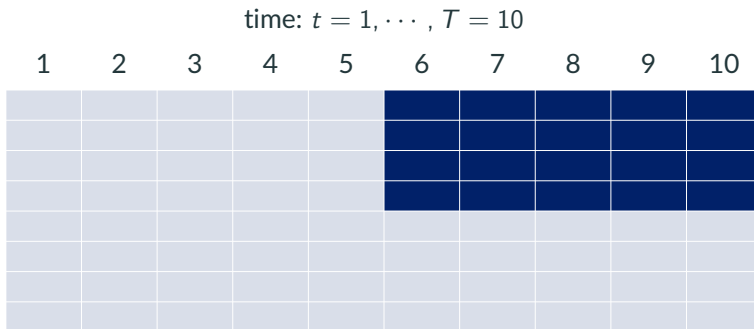
- (i) (S) $\{(Y_{i,t=1}, \dots, Y_{i,t=T}, X'_i, G_i)'\}_{i=1}^n$ is a random sample from $(Y_{t=1}, \dots, Y_{t=T}, X', G)'$.
- (ii) (O) For each $g \in \mathcal{G}$, $\mathbb{E}[G_g|X] \in (0, 1)$ almost surely (a.s.).
- (iii) (NA) For every $g \in \mathcal{G}_{trt}$, and every pre-treatment periods $t < g$, $\mathbb{E}[Y_{i,t}(g)|G = g, X] = \mathbb{E}[Y_{i,t}(\infty)|G = g, X]$ almost surely.

- The maintained Assumption M is not enough to identify ATT or ES type parameters.
- DiD methods impose parallel trends assumptions to identify these parameters.

DiD with Single Treatment Time

No variation in treatment timing

- Single treatment period at time g : $G_i = g$ (treated) or $G_i = \infty$ (untreated).
- $ES(e) = ATT(g, g + e)$.



Parallel Trends Assumption: Post-treatment periods

Assumption (PT in the post-treatment periods)

For each $t \in \{2, \dots, T\}$ such that $t \geq g$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = \infty, X] \text{ a.s.}$$

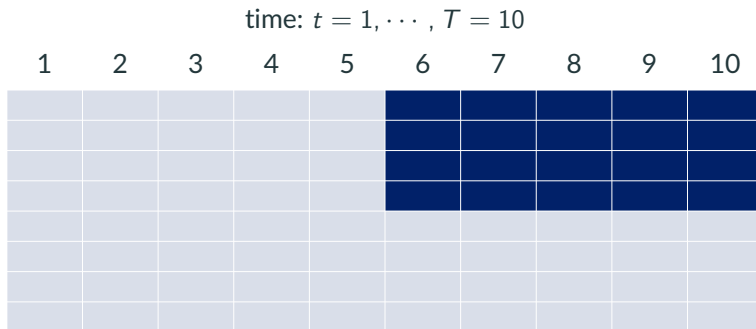
- Impose parallel trends only for post-treatment periods.
- Uses only period $g - 1$ as the baseline.
- Without covariates, easy to show that, for any $t \geq g$,

$$ATT(g, t) = \mathbb{E}[Y_t - Y_{g-1} | G = g] - \mathbb{E}[Y_t - Y_{g-1} | G = \infty].$$

- Limitation: If we were gifted 10 more pre-treatment data periods, the estimand for $ATT(g, t)$ would not be allowed to use any of that.

PT-Post: Implications

- $ES(e) = ATT(g, g + e)$ for any $e \geq 0$



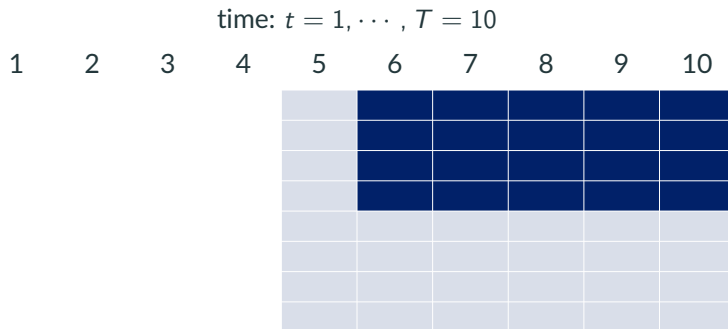
Treated



Untreated

PT-Post: Implications

- $ES(e) = ATT(g, g + e)$ for any $e \geq 0$



Parallel Trends Assumption: A recent window

Assumption (PT-Window- g)

Fix an integer $W \geq 0$. For each $t \in \{2, \dots, T\}$ such that $t \geq g - W$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) \mid G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) \mid G = \infty, X] \text{ a.s.}$$

- Researchers may be willing to impose parallel trends near the treatment date.
- Without covariates, for any $t \geq g$ and $g - W - 1 \leq t' < g$,

$$ATT(g, t) = \mathbb{E}[Y_t - Y_{t'} \mid G = g] - \mathbb{E}[Y_t - Y_{t'} \mid G = \infty].$$

- Same $ATT(g, t)$, more valid baselines.
- What if we are willing to impose parallel trends over the entire pre-treatment history?

Parallel Trends Assumption: All periods

Assumption (PT in all periods)

For each $t \in \{2, \dots, T\}$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = \infty, X] \text{ a.s.}$$

- Impose parallel trends in all periods.
- Allows us to use any pre-treatment period as a baseline.
- Without covariates, easy to show that for any $t \geq g$ and any $t' < g$,

$$ATT(g, t) = \mathbb{E}[Y_t - Y_{t'} | G = g] - \mathbb{E}[Y_t - Y_{t'} | G = \infty].$$

- If we were gifted 10 more pre-treatment periods of data, we could easily use all of them to compute $ATT(g, t)$.
- How to do that efficiently, such that we maximize precision?

Characterization of the DiD model based on seq. conditional moment restrictions

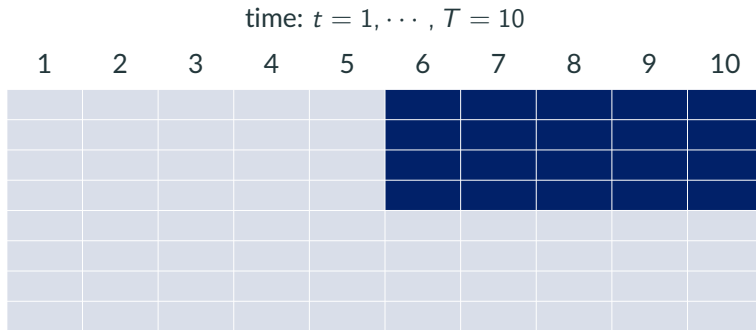
Lemma (Moment-restrictions for over-identified DiD with single treatment time)

The family of prob. dist. of $(Y_{t=1} \cdots, Y_{t=T}, X', G)$ satisfying Assumptions M and PT-All- g are observationally equivalent to the family of prob. dist. of $(Y_{t=1} \cdots, Y_{t=T}, X', G)$ satisfying Assumption M(i)(ii), and the set of moment restrictions: for all $t \in \{g, \dots, T\}$, with prob. one,

$$\begin{aligned}\mathbb{E}[G_g(ATT(g, t) - CATT(g, t, X))] &= 0, \\ \mathbb{E}\left[CATT(g, t, X) - \frac{G_g(Y_t - Y_{g-1})}{p_g(X)} + \frac{G_\infty(Y_t - Y_{g-1})}{p_\infty(X)} \middle| X\right] &= 0, \\ \mathbb{E}\left[\frac{G_g(Y_{t'} - Y_1)}{p_g(X)} - \frac{G_\infty(Y_{t'} - Y_1)}{p_\infty(X)} \middle| X\right] &= 0, \text{ for all } 2 \leq t' \leq g-1, \\ \mathbb{E}[G_g - p_g(X)|X] &= 0.\end{aligned}$$

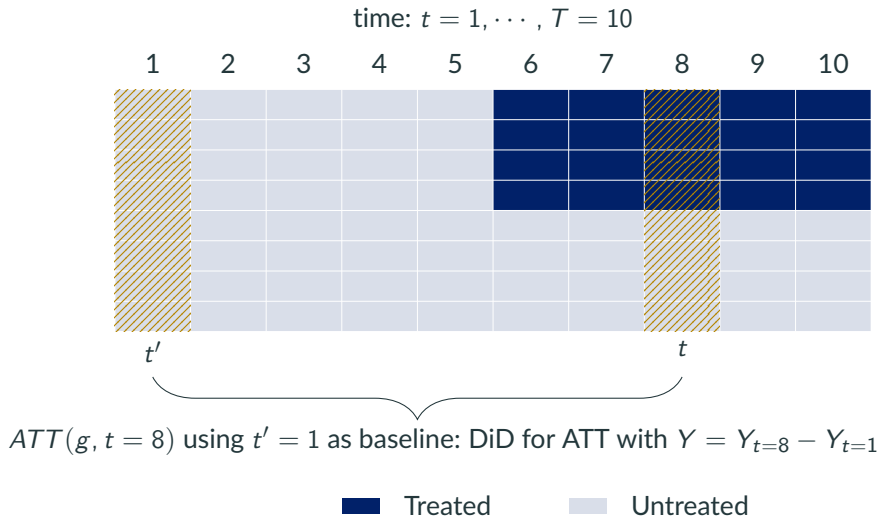
- Semiparametric efficient bound for $ATT(g, t)$: apply Ai and Chen (2012) for seq. moments.

Understanding the sources of over-identification

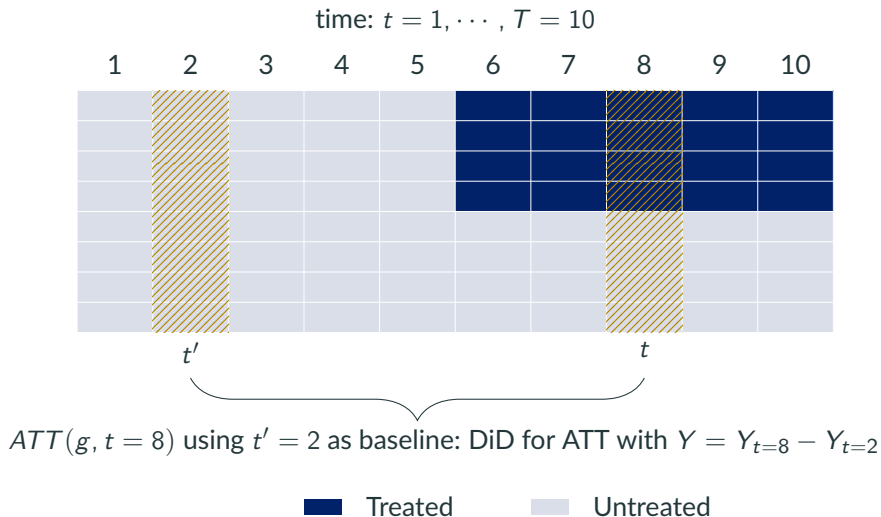


■ Treated ■ Untreated

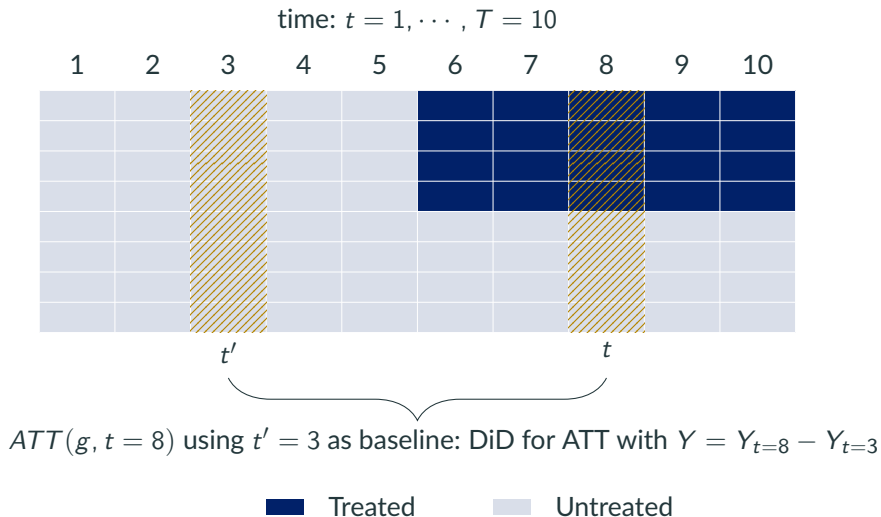
Understanding the sources of over-identification



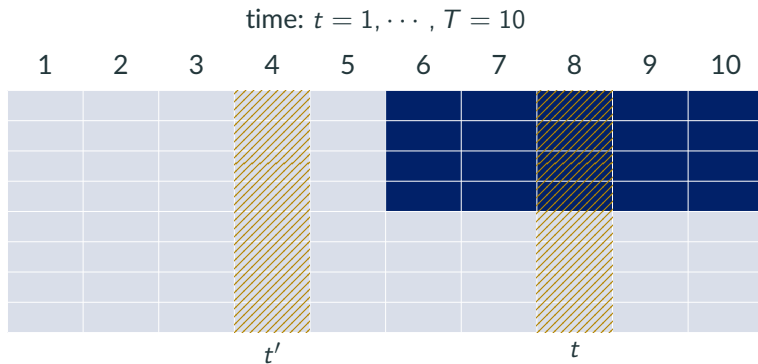
Understanding the sources of over-identification



Understanding the sources of over-identification



Understanding the sources of over-identification



$ATT(g, t = 8)$ using $t' = 4$ as baseline: DiD for ATT with $Y = Y_{t=8} - Y_{t=4}$

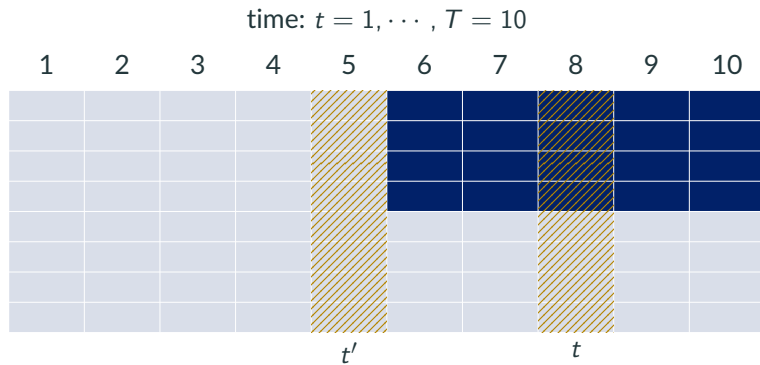


Treated



Untreated

Understanding the sources of over-identification



$ATT(g, t = 8)$ using $t' = 5$ as baseline: DiD for ATT with $Y = Y_{t=8} - Y_{t=5}$

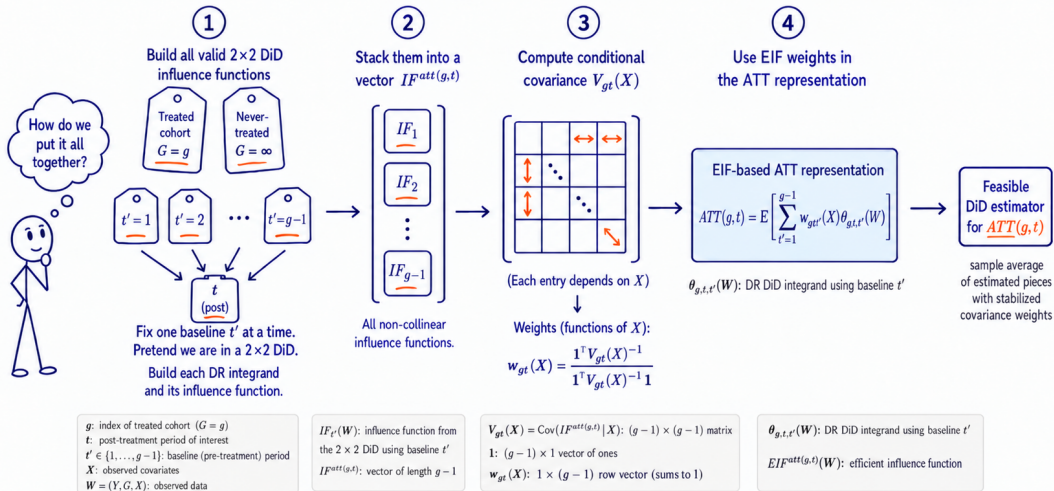


Treated



Untreated

Intuition on how to get semiparametric efficiency



Under stated conditions, optimal weighting attains the semiparametric efficiency bound.

Intuition on how to get semiparametric efficiency

- For each $1 \leq t' \leq g - 1$, we fix the baseline period at t' , and compute the “efficient influence function” for $ATT(g, t)$ as-if there were only 2 groups, $G = g$ and $G = \infty$, and two periods, t (post-treatment) and t' (pre-treatment)
 - ▶ Akin to compute the “DR scores” in DML language.
- Stack all the non-collinear influence functions into a vector, $\mathbb{IF}^{att(g,t)}$.
- Compute the covariance of $\mathbb{IF}^{att(g,t)}$ given covariates, $V_{gt}(X) = \text{Cov}(\mathbb{IF}^{att(g,t)} | X)$.
- Efficient Influence Function for $ATT(g, t)$ is given by

$$EIF^{att(g,t)} = \frac{1' V_{gt}(X)^{-1}}{1' V_{gt}(X)^{-1} 1} \mathbb{IF}^{att(g,t)}.$$

- Next, we explore these results to obtain EIF-based estimands for $ATT(g, t)$, which serve as a blueprint for efficient estimation.

Using EIF as a blueprint for estimating $ATT(g,t)$

- The key is to explore that $\mathbb{E}[EIF^{att(g,t)}] = 0$ to get IF-based estimand:

$$ATT(g, t) = \mathbb{E} \left[\frac{1' V_{gt}(X)^{-1}}{1' V_{gt}(X)^{-1} \mathbf{1}} \theta_{g,t}(W) \right], \quad (1)$$

where $p_g(X) = \mathbb{E}[G_g|X]$, $\theta_{g,t}(W) = (\theta_{g,t,1}(W), \dots, \theta_{g,t,g-1}(W))'$ is a $(g-1) \times 1$ column vector with

$$\theta_{g,t,t'}(W) = \frac{1}{\mathbb{P}(G = g)} \left(G_g - \frac{(1 - G_g)p_g(X)}{1 - p_g(X)} \right) (Y_t - Y_{t'} - \mathbb{E}[Y_t - Y_{t'}|G = \infty, X]).$$

- Here, $\theta_{g,t}(W)$ is a vector of DR DiD “integrands”, each being computed pretending we were in the 2×2 DiD setup of Sant’Anna and Zhao (2020).
- Efficient DiD estimators: estimate the nuisance functions and stabilize noisy covariance inversions before averaging the fitted scores.

DiD with Single Treatment Time

What do these results teach us?

So how should you weight the pre-treatment periods?

The efficient baseline solves a one-line problem: choose weights ω on the pre-periods, summing to one, that **minimize the variance** of the counterfactual you difference against:

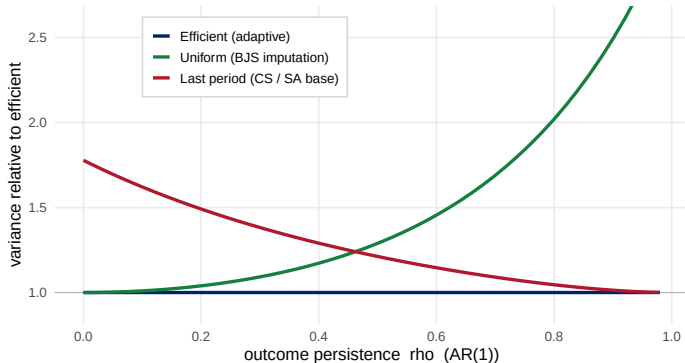
$$\omega^* = \Sigma^{-1} \left[\sigma_{\text{pre,post}} + \frac{1 - 1' \Sigma^{-1} \sigma_{\text{pre,post}}}{1' \Sigma^{-1} 1} 1 \right], \quad \Sigma = \text{Var}(u_{\text{pre}}).$$

The answer turns on one feature of the data – the **serial correlation** of the outcome – with a clean closed form in three canonical cases:

error structure	efficient baseline weights $\omega^*(e)$	matches the <u>baseline</u> of
iid	uniform (long pre-average), any horizon e	BJS, Wooldridge (2025a)
AR(1), ρ	head + tail ; recency fades as e grows	– (interpolates)
unit root	last pre-period Y_{g-1} , any horizon e	CS, SA, MS21, Harmon (2023)

The named estimators each **fix** this baseline choice; only the efficient bound adapts it to the data.

Every fixed baseline is optimal at exactly one persistence



Illustrative ($T_0 = 8$ pre-periods, AR(1)); the gap grows with the number of pre-periods.

On the baseline dimension, today's estimators are two fixed choices:

- **BJS-G-W** $\approx$ uniform: efficient only at $\rho = 0$; pays **up to 3×** near a unit root.
- **CS, SA, dCDH, MS, Harmon** difference against the last period: efficient as $\rho \rightarrow 1$; “waste” **78%** under iid.

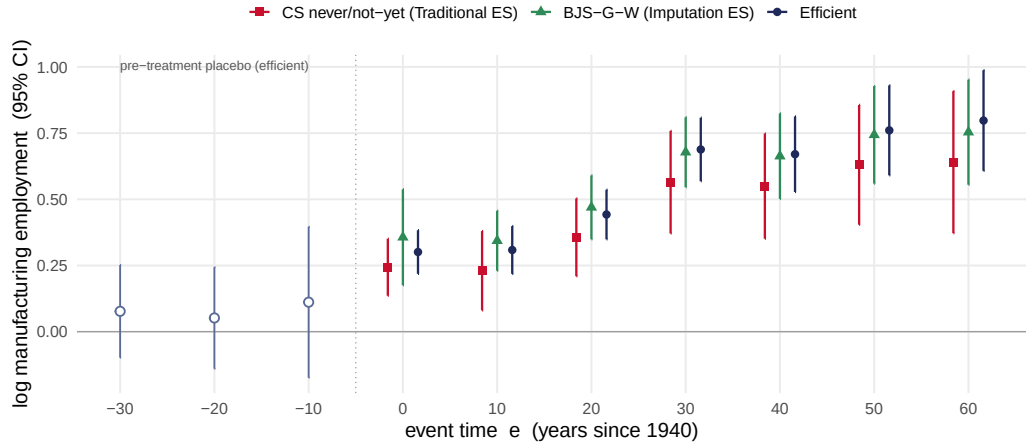
The bound is the lower envelope — optimal at every ρ .

You no longer have to guess.

Empirical Illustration based on Kline and Moretti (2014)

- **The question.** Did the **Tennessee Valley Authority** — the New Deal's flagship place-based “Big Push” (dams, electrification, cheap power, from 1933) — create **lasting agglomeration in manufacturing**?
- **Their finding.** A durable positive effect on **manufacturing** employment (Kline and Moretti, 2014)
- **The data.** **U.S. counties**, decennial census 1900–2000 (11 waves). We use the balanced panel part of KM replication package: 1,420 **counties**, 87 **treated**. Outcome: log county manufacturing employment.
- **Timing.** The TVA Act is 1933, with **73% of federal transfers landed 1940–1958**: so KM take 1940 as their baseline (1900–1940 is their placebo window, 1940–2000 their effect window). We use 1900–1930 as baseline for the efficient estimator (with $G = 1940$).
- **Our spec.** **No covariates**; weights use the covariance of outcome changes clustered by **state**. 47 states (6 treated); four baselines, 1900–1930 (PT-All).
- One date, no covariates: direct and estimate-then-aggregate targeting coincide; gains come from baseline pooling.

Kline and Moretti (2014): efficient DiD vs fixed-weight alternatives



Post: leading normal 95% CIs and ARE; model-qualified MDE. Pre: native t intervals; joint test $p = 0.087$.

Traditional: 0.459 (0.088) · ARE 2.30 · MDE 0.246

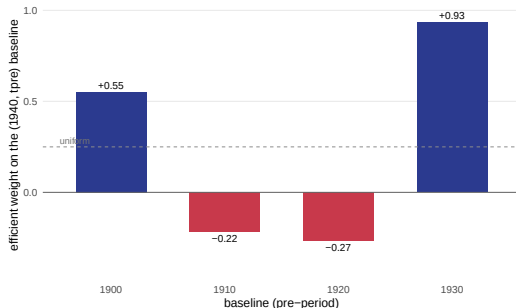
Imputation: 0.573 (0.069) · ARE 1.42 · MDE 0.193

Efficient: 0.567 (0.058) · ARE 1.00 · MDE 0.162

Kline and Moretti (2014): the efficient weights for ES(0) spread

TVA efficient baseline weights for $ES(0) = ATT(1940, 1940)$

Cluster-efficient (47 state clusters), no covariates. Spread 1.20; 2 baselines enter negatively.



Efficiency does not simply favor the recent baseline:

It loads **+0.93** on 1930 and **+0.55** on 1900 while 1910 and 1920 enter **negatively** (-0.22 , -0.27).

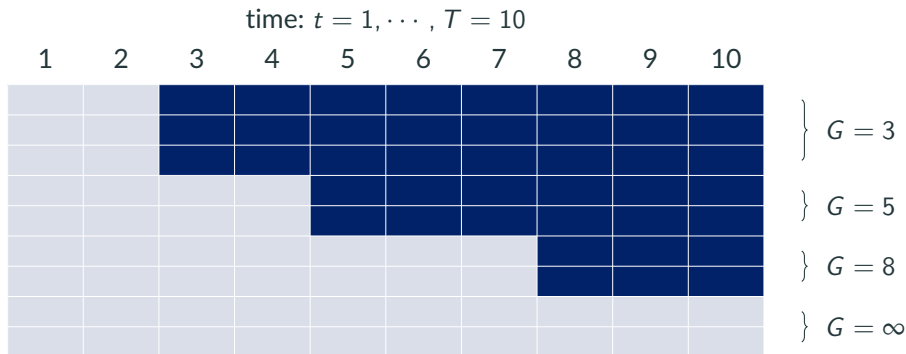
Gain at $ES(0)$ (leading, normal-reference). These efficient weights buy ARE = **1.71×** vs. Trad. ES=last-pre (a 23% shorter CI) and **4.84×** vs. Imputation ES (a 55% shorter CI).

These estimated leading-variance ratios vary with event-time, depending on the (clustered) covariance terms of treated and untreated units: across horizons the traditional ES stays in a **1.7–2.8×** band, while imputation ranges from **4.8×** at impact down to **1.1×** at $e = 60$.

DiD with Staggered Treatment Adoption

Staggered Adoption

- Multiple treatment starting periods, leading to several treatment groups defined by treatment starting date.
- Each group has their $ATT(g, t)$.



Assumption (Parallel Trends for all groups and periods (PT-All))

For each $t \in \{2, \dots, T\}$ and $(g, g') \in \mathcal{G}_{trt} \times \mathcal{G}$,

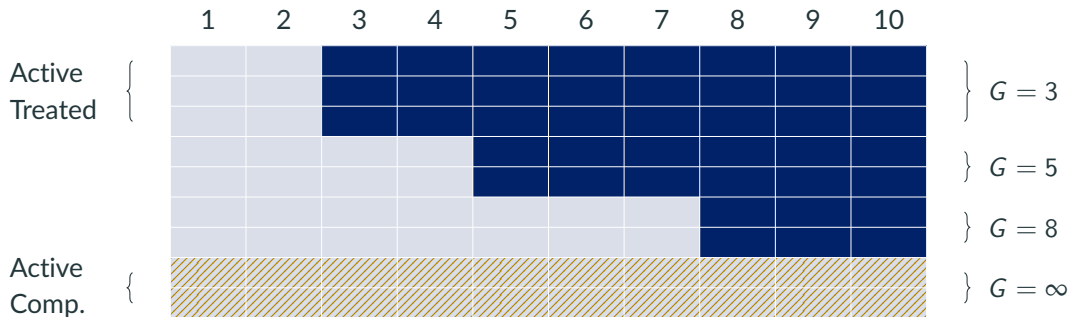
$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = g', X] \text{ a.s.}$$

i.e., conditional on covariates, the average evolution of untreated potential outcomes is the same across treatment groups, in all available periods.

- Two sources of nonparametric over-identification (in the sense of Chen and Santos (2018)): multiple baseline periods and multiple comparison groups.

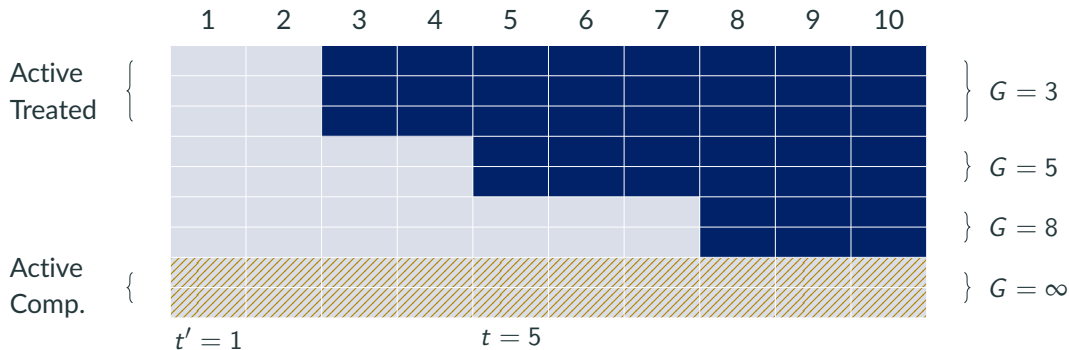
Understanding the sources of over-identification

$ATT(g = 3, t = 5)$, using one active comparison group $G = \infty$:



Understanding the sources of over-identification

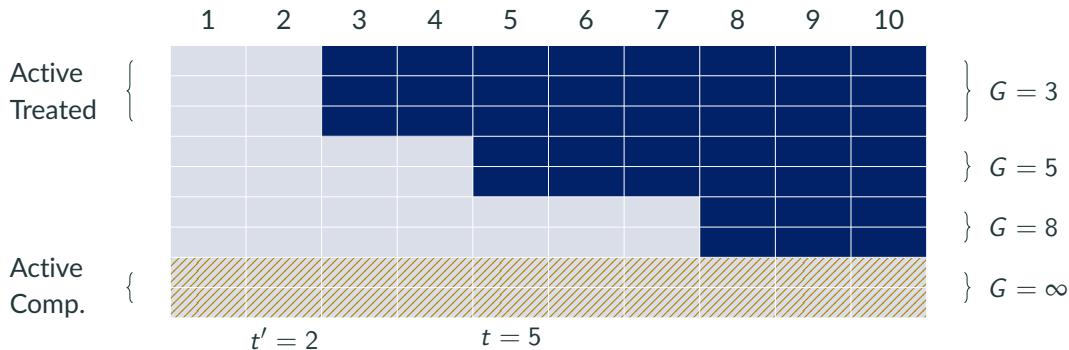
$ATT(g = 3, t = 5)$, using one active comparison group $G = \infty$:



$ATT(g = 3, t = 5)$ using $t' = 1$ and $G = \infty$

Understanding the sources of over-identification

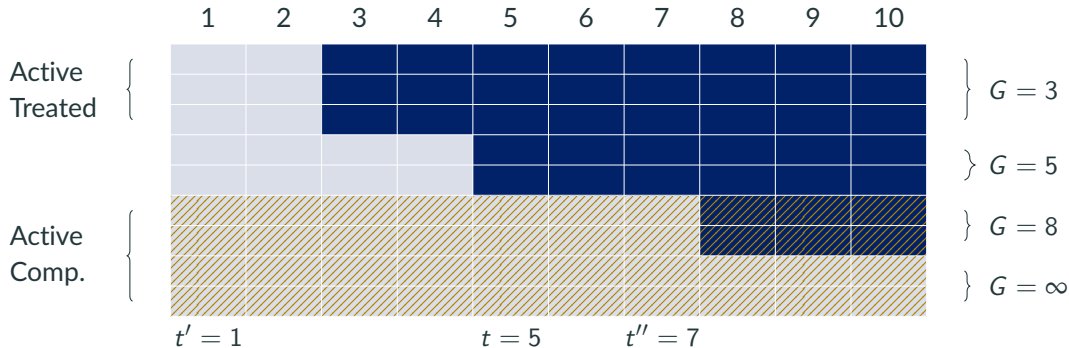
$ATT(g = 3, t = 5)$, using one active comparison group $G = \infty$:



$ATT(g = 3, t = 5)$ using $t' = 2$ and $G = \infty$

Understanding the sources of over-identification

$ATT(g = 3, t = 5)$, using two active comparison groups $G = 8$, and $G = \infty$:



$ATT(g = 3, t = 5)$ using $t' = 1$, $t'' = 7$, and $G = \infty, 8$

Exploring all the content of PT

Lemma (Over-identification in staggered designs)

Under Assumptions M and PT-All, for every group $(g, g') \in \mathcal{G}_{\text{trt}} \times \mathcal{G}_{\text{trt}}$ and time periods $(t, t', t'') \in \mathcal{T} \times \mathcal{T} \times \mathcal{T}$ such that $t \geq g, g > t'$, and $g' > \max\{t', t''\}$, with probability one,

$$CATT(g, t, X) = \underbrace{\mathbb{E}[Y_t - Y_{t'} | G = g, X]}_{\equiv m_{g,t,t'}(X)} - \left(\underbrace{\mathbb{E}[Y_t - Y_{t''} | G = \infty, X]}_{\equiv m_{\infty,t,t''}(X)} + \underbrace{\mathbb{E}[Y_{t''} - Y_{t'} | G = g', X]}_{\equiv m_{g',t'',t'}(X)} \right), \quad (2)$$

and, as a consequence,

$$ATT(g, t) = \mathbb{E}[Y_t - Y_{t'} | G = g] - \mathbb{E}[(m_{\infty,t,t''}(X) + m_{g',t'',t'}(X)) | G = g]. \quad (3)$$

More generally, for any covariate-specific weights $w_{g',t',t''}^{g,t}(X)$ that sum up to one, we have that

$$ATT(g, t) = \mathbb{E} \left[\sum_{(g',t',t'') \in \mathcal{H}^{g,t}} w_{g',t',t''}^{g,t}(X) [m_{g,t,t'}(X) - (m_{\infty,t,t''}(X) + m_{g',t'',t'}(X))] \mid G = g \right].$$

Many valid staggered DiDs for the same ATT(g,t)

Fix baseline $t' = 1$

We fix $t' = 1$ as the baseline for simplicity.



Target:
ATT(g,t)

One valid staggered IF building block



$G = g$

treated DR piece

uses the treated cohort $G = g$ and outcome-regression residuals.

Target cohort g :

$$Y_t - Y_1$$



$G = \infty$

never-treated bridge

uses the ratio $p_g(X)/p_\infty(X)$ to reweight never-treated units.

Never-treated bridge:

$$Y_t - Y_{t''}$$



$G = g'$

comparison-cohort bridge

uses the ratio $p_g(X)/p_{g'}(X)$ to reweight cohort g' .

Comparison cohort g' :

$$Y_{t''} - Y_1$$

DR building block = p-score ratios + outcome regressions

$$p_h(X) = \Pr(G = h \mid X) \quad m_{a,b,c}(X) = E[Y_b - Y_c \mid G = a, X].$$

Card label =
(comparison cohort, pre-period)

cohort g ,
pre 1

cohort g ,
pre 2

cohort $g+1$,
pre 2

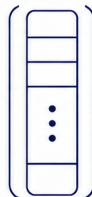
cohort $g+1$,
pre 3

...

one card = one admissible
cohort / pre-period comparison

$IF_{stg}^{att(g,t)} =$

stack of all
non-collinear
IF terms



g = target treated cohort

g' = comparison treated cohort

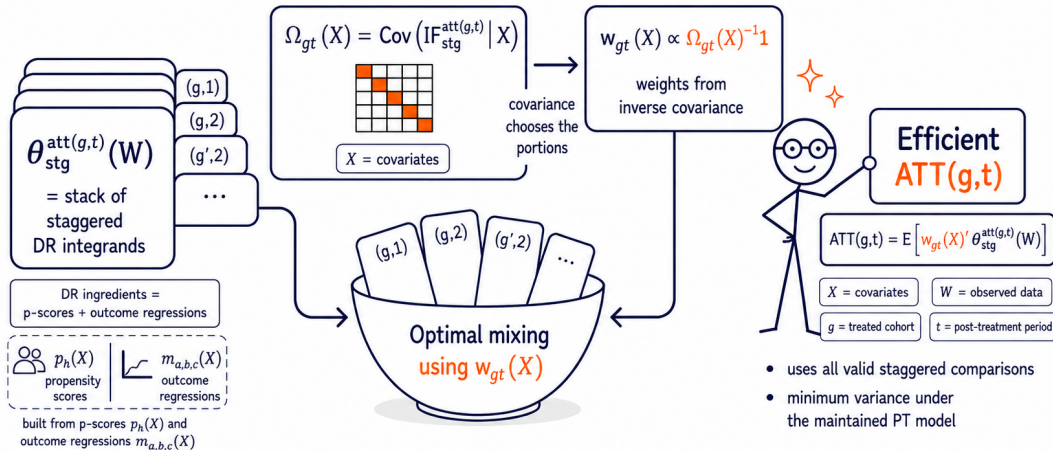
t = post-treatment
period of interest

t'' = admissible pre-treatment
period for cohort g'

$G = \infty$ = never-treated group

One target. **Many** valid staggered comparisons.

Efficient staggered DiD = combine them optimally



Use the covariance to turn many valid staggered pieces into **one efficient DiD estimator**.

Semiparametric Efficient Staggered DiD

■ We will now fix $t' = 1$ for simplicity (no loss of generality).

■ For $g' \in \mathcal{G}_{\text{trt}}$ and $1 \leq t'' \leq g' - 1$, let $\pi_g = \mathbb{E}[G_g]$ and

$$\begin{aligned} \mathbb{IF}_{g',t''}^{\text{att}(g,t)} = & \frac{1}{\pi_g} \left(G_g((m_{g,t,1}(X) - m_{\infty,t,t''}(X) - m_{g',t'',1}(X)) - \text{ATT}(g,t)) \right) \\ & + \frac{G_g}{\pi_g} (Y_t - Y_1 - m_{g,t,1}(X)) \\ & - \frac{G_{\infty}}{\pi_g} \frac{p_g(X)}{p_{\infty}(X)} (Y_t - Y_{t''} - m_{\infty,t,t''}(X)) - \frac{G_{g'}}{\pi_g} \frac{p_g(X)}{p_{g'}(X)} (Y_{t''} - Y_1 - m_{g',t'',1}(X)). \quad (4) \end{aligned}$$

■ We then stack all the non-collinear IF terms to form $\mathbb{IF}_{\text{stg}}^{\text{att}(g,t)}$.

Efficient Staggered DiD

- Here follows the efficient staggered DiD estimand

$$ATT(g, t) = \mathbb{E} \left[\frac{1' \Omega_{gt}(X)^{-1}}{1' \Omega_{gt}(X)^{-1} \mathbf{1}} \theta_{\text{stg}}^{\text{att}(g,t)}(W) \right], \quad (5)$$

where $\Omega_{gt}(X) = \text{Cov}(\mathbb{I}\mathbb{F}_{\text{stg}}^{\text{att}(g,t)} | X)$, $\theta_{\text{stg}}^{\text{att}(g,t)}(W)$ is the column vector

$$\theta_{\text{stg}}^{\text{att}(g,t)}(W) = (\theta_{g'}^{\text{att}(g,t)}(W)', g' \in \mathcal{G}_{\text{trt}})',, \quad (6)$$

such that, for $g' = g$,

$$\theta_{g'}^{\text{att}(g,t)}(W) = (\theta_{g,1}^{\text{att}(g,t)}(W), \dots, \theta_{g,g-1}^{\text{att}(g,t)}(W))',$$

and, for $g' \neq g$,

$$\theta_{g'}^{\text{att}(g,t)}(W) = (\theta_{g',2}^{\text{att}(g,t)}(W), \dots, \theta_{g',g'-1}^{\text{att}(g,t)}(W))',$$

with

$$\begin{aligned} \theta_{g',t''}^{\text{att}(g,t)}(W) = & \frac{1}{\pi_g} G_g(Y_t - Y_1 - m_{\infty,t,t''}(X) - m_{g',t'',1}(X)) \\ & - \left(\frac{G_{\infty}}{\pi_g} \frac{p_g(X)}{p_{\infty}(X)} (Y_t - Y_{t''} - m_{\infty,t,t''}(X)) + \frac{G_{g'}}{\pi_g} \frac{p_g(X)}{p_{g'}(X)} (Y_{t''} - Y_1 - m_{g',t'',1}(X)) \right). \end{aligned} \quad (7)$$

Semiparametric efficiency result

Theorem (Efficient DiD with staggered treatment adoption)

Under Assumptions M and PT-All, the efficient influence function for $ATT(g, t)$, $t \geq g$, is given by

$$\mathbb{E}\mathbb{IF}_{stg}^{att(g,t)} = \frac{1'\Omega_{gt}(X)^{-1}}{1'\Omega_{gt}(X)^{-1}\mathbf{1}}\mathbb{IF}_{stg}^{att(g,t)}.$$

The semiparametric efficient variance bounds are obtained as the second moments of the efficient influence functions, provided they are finite.

For $ES(e)$, $e \geq 0$, DIRECT targets the aggregate score without first estimating the full $ATT(g, t)$ table; alternatively, estimate these effects and then aggregate.

Estimation: Can we just plug in?

- **Yes, in principle.** Estimate the nuisance functions and the covariance, then compute the weights.
- **The problem is the inverse.** Covariance noise can become extreme weights, raising MSE and distorting confidence intervals.
- **Our approach.** Stabilize poorly estimated directions and base inference on the declared sampling design.

Why this matters

Learning weights can be negligible to first order and still matter in finite samples.

From the efficiency bound to feasible weights

1. **Start from the model.** Derive its semiparametric efficiency bound (SEB).
2. **Read off its weights.** The efficient influence function for $ATT(g, t)$ is a covariance-weighted combination of the model's comparison influence functions.
3. **Implement it.** Stabilize the estimated conditional covariance of those IFs, $\hat{\Omega}^S(x)$:

$$\hat{w}(x) = \frac{\hat{\Omega}^S(x)^{-1} \mathbf{1}}{\mathbf{1}' \hat{\Omega}^S(x)^{-1} \mathbf{1}}, \quad \mathbf{1}' \hat{w}(x) = 1.$$

Stabilize: add a diagonal cushion, rescale, and floor small eigenvalues. [Details](#)

A noisy inverse can generate extreme weights (Chaudhuri, 2026).

Stabilize the inverse, retain the comparisons

Pointwise covariance weights with no analysis weights; n panel units. Suppress x .

- **1. Add a diagonal cushion:** $\hat{\Omega}^R = \hat{\Omega} + r_n I$, $r_n = [\text{tr}(\hat{\Omega})/n]_+$.
- **2. Scale by marginal variances:** $\hat{V} = \text{diag}(|\text{diag}(\overline{\Omega})|)$, $\overline{\Omega} = \mathbb{P}_n \hat{\Omega}(X)$.
Form $S = \hat{V}^{-1/2} \hat{\Omega}^R \hat{V}^{-1/2}$.
- **3. Floor the eigenvalues:** if $S = U \text{diag}(\ell_j) U'$, replace ℓ_j by $\max(\ell_j, \tau_n)$; then restore the original units to obtain $\hat{\Omega}^S$.

$$\tau_n = \eta_n \max_j |\ell_j|, \quad \eta_n = \min(n^{-a}, 0.05), \quad a = 0.07\{5 - \min(d, 4)\}.$$

Key: d is the expanded continuous smoothing width; with only discrete covariates, use the expanded exact-support width.

As $n \rightarrow \infty$: with fixed dimension and consistent covariance estimates with a positive-definite limit, $r_n, \tau_n \rightarrow 0$. Efficiency also requires the stated nuisance-rate and regularity conditions.

Estimate the target you want

Estimate, then aggregate

Estimate cohort-time effects.

Then aggregate them into the event-study target.

DIRECT

Build the aggregate score.

Use clean comparisons to improve its precision.

- **The saving:** a short event-study path or average without first estimating the intermediate cohort-time effects.
- **The benchmark:** same target, model and sampling law $\Rightarrow$ same efficiency bound under the stated conditions.
- **The limit:** finite-sample precision can differ. For our supported TVA specification, the two implementations coincide.

The efficient weights — and why existing estimators fall short

- Optimal aggregation across pre-periods and comparison groups is governed by $V_{gt}(X)^{-1}$. The weights:
 - ▶ depend on the **covariance of outcome changes** within each group;
 - ▶ **vary** with the event time t ;
 - ▶ are **covariate-dependent** (unlike GMM) — e.g. optimal weights for men may differ from women;
 - ▶ are **generally not constant** across pre-treatment periods.
- Most DiD / ES estimators are **not** efficient — they either:
 - ▶ use a **single** pre-period t' : Dynamic TWFE (DTWFE), de Chaisemartin and D'Haultfœuille (2020), Callaway and Sant'Anna (2021), Sun and Abraham (2021); or
 - ▶ take a **simple average** over $t' < g$: TWFE, Wooldridge (2025b), Gardner (2021), Borusyak, Jaravel and Spiess (2024).

What changes the efficient weights?

- **Relative noise.** Two equally sized comparison groups need not be equally informative.
 - With mutually uncorrelated control contributions and a shared treated contribution uncorrelated with them, precision weights satisfy $\psi_h \propto N_h / \sigma_h^2$.
- **Predictive value.** A baseline helps when it predicts the post-treatment outcome's noise, even if that baseline is itself noisy.
 - In the random-walk benchmark, use the last pre-period; with iid errors, average the pre-periods.
- **The target horizon.** The useful history can change with the effect we want to estimate.
 - With stationary positive AR(1) errors, the recency advantage fades at longer horizons.

What need not change the efficient weights?

When covariates change only the overall noise level

If the full comparison-IF covariance is $\Omega(x) = a(x)\Omega_0$, with $a(x) > 0$, the factor cancels from the normalized weights.

→ One covariate group can be more volatile and still use **the same combination of comparisons**.

When treatment-effect variation adds shared, uncorrelated noise

If the added component is common to all comparisons and conditionally uncorrelated with their remaining noise given X , it adds $c(x)11'$ to their covariance.

→ **The weights stay the same; uncertainty rises.** The extra common noise cannot be removed by changing these weights.

These are population results for a fixed target and comparison set. Different covariance shapes or correlations with the added effect component can change the weights.

DiD with Staggered Treatment Adoption

What do these results teach us? (staggered)

Staggering adds a second dial: which cohorts to compare to

- **Single date:** only the baseline (time) dial is live — the synthetic-baseline DiD against the one control group.
- **Staggered:** a second dial appears — **which cohorts to compare to** (never- and not-yet-treated $g' > t$). The efficient weight sometimes **factorizes** into a baseline factor $\times$ a comparison factor:

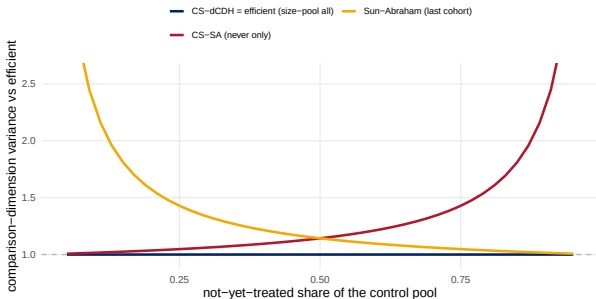
$$w_{i,s}(g, t) \approx \underbrace{\omega_s(g)}_{\text{persistence (time)}} \times \underbrace{\psi_{g(i)}(g, t)}_{\text{precision (groups)}}$$

- **This happens** when the time and group covariances separate (e.g. a unit root); **approximate** otherwise — the not-yet comparison tilts the per-cohort baseline.

The comparison dial: precision, not size

On the comparison dimension the efficient closed form is **precision-weighting**:

$$\psi_{g'} \propto N_{g'} / \sigma_{g'}^2 \quad (\text{a precision-weighted pool of every valid control}).$$



- Callaway and Sant'Anna (2021) (not-yet) and de Chaisemartin and D'Haultfœuille (2020) **pool all valid controls** by size — comparison-efficient.
- Callaway and Sant'Anna (2021) (never) and Sun and Abraham (2021) (last-treated cohort) **discard controls** — the loss grows with the share dropped.
- Imputation-type estimators (BJS-G-W) pool all untreated, but **average the baseline** rather than fixing Y_{g-1} — the iid corner [► Intuition](#).
- The bound pools the comparison by precision and optimizes the baseline.

Empirical Illustration based on Dobkin et al. (2018)

- Dobkin, Finkelstein, Kluender and Notowidigdo (2018) study the effect of hospitalization on out-of-pocket medical spending using a DiD / Event-Study strategy on the timing of hospitalization.
- We follow the sample construction of Sun and Abraham (2021), using the publicly available Health and Retirement Study (HRS) data from the Dobkin et al. (2018) replication package:
 - ▶ Adults hospitalized at ages 50–59, excluding pregnancy-related admissions.
 - ▶ Balanced estimation panel spanning waves 7–10 (2004–2010).
 - ▶ Final sample: 656 individuals in 4 treatment groups — $G_i = 8$ (252), $G_i = 9$ (176), $G_i = 10$ (163), $G_i = \infty$ (only **65** untreated through the window).
- Small comparison group $\Rightarrow$ **power is binding**, and precision matters.

Application: Dobkin et al (2018)

Point estimates are similar across estimators

Estimator	$ATT(8, 8)$	$ATT(8, 9)$	$ATT(8, 10)$	$ATT(9, 9)$	$ATT(9, 10)$	$ATT(10, 10)$	$ES(0)$	$ES(1)$	$ES(2)$	ES_{avg}
EDiD	3071 (805)	1115 (635)	1041 (812)	3067 (690)	95 (636)	2911 (889)	3025 (486)	695 (470)	1041 (812)	1587 (519)
CS-N	2826 (1035)	825 (909)	800 (1008)	3031 (702)	107 (651)	3092 (995)	2960 (539)	530 (585)	800 (1008)	1430 (647)
CS-NYT	3029 (913)	1248 (861)	800 (1008)	3324 (959)	107 (651)	3092 (995)	3134 (536)	779 (570)	800 (1008)	1571 (566)
BJS-G-W	3029 (913)	1285 (765)	1021 (849)	3239 (859)	77 (727)	2758 (955)	3017 (554)	788 (586)	1021 (849)	1609 (580)

EDiD: DIRECT. CS-N / CS-NYT: never / not-yet treated. Parentheses: model-qualified leading SEs, with cohort-share uncertainty retained for all methods.

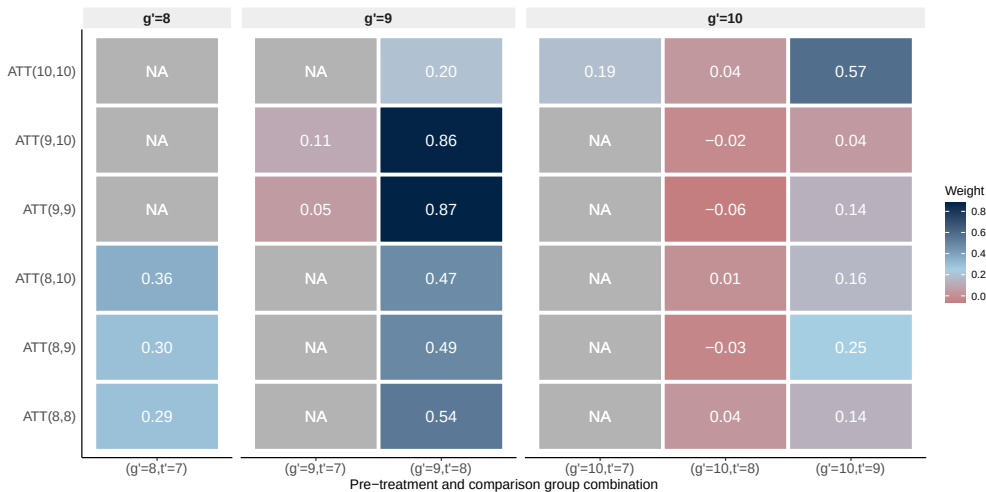
Our efficient DiD delivers substantial gains in efficiency

- Estimated ARE: comparator leading variance divided by DIRECT's — heuristically, the **relative sample size** needed under the maintained model; not a finite-sample coverage guarantee.

Estimator	$ATT(8, 8)$	$ATT(8, 9)$	$ATT(8, 10)$	$ATT(9, 9)$	$ATT(9, 10)$	$ATT(10, 10)$	$ES(0)$	$ES(1)$	$ES(2)$	ES_{avg}
EDiD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CS-N	1.65	2.05	1.54	1.04	1.05	1.25	1.23	1.55	1.54	1.55
CS-NYT	1.29	1.84	1.54	1.93	1.05	1.25	1.22	1.47	1.54	1.19
BJS-G-W	1.29	1.45	1.09	1.55	1.31	1.15	1.30	1.56	1.09	1.25

- Several alternatives need **up to twice the sample size** (e.g. $ATT(8, 9)$: $ARE = 2.05$) for the same precision; the ranking is not fixed across parameters.

Understanding efficient weights for the $ATT(g, t)$ estimates



Takeaways

Takeaways

- **Two design choices are already being made** in every DiD paper — which pre-periods, which comparison cohorts. They are DiD's **over-identifying restrictions**.
- **The efficiency bound says how to set them.** The baseline is a **dial on persistence** (uniform under iid $\rightarrow$ last-period under a random walk, with a closed-form interpolation); the comparison is **precision-weighted**, not size-weighted.
- **The named estimators are corners.** BJS-G-W, CS, SA, dCDH each fix both dials in advance — we recover and explain them, and the bound is the lower envelope. **You no longer have to guess.**
- **Honesty comes from the same moments.** The restrictions that fund the efficiency also test it — the declared ladder, movement tests, and the frontier are read off one object.
- **It matters empirically.** In the lead example, the common alternative needs **1,019** respondents to do what the efficient estimator does with **656**; across eight published designs the fitted history-comparison precision changes run from **-12%** to **+343%**.

Thank you!

overdid — the R implementation

Getting more intuition around imputation-type DiD estimators

◀ Back to talk

At treatment onset, $t = g$, the regression is especially transparent. Let $\mathcal{C}(g) = \{h : h > g\}$ + never be the cohorts still untreated at g .

$$\widehat{\text{ATT}}(g, g) = \sum_{h \in \mathcal{C}(g)} \frac{N_h}{N_{\mathcal{C}(g)}} \frac{1}{g-1} \sum_{s < g} \left\{ [\bar{Y}_{g,g} - \bar{Y}_{g,s}(\infty)] - [\bar{Y}_{h,g}(\infty) - \bar{Y}_{h,s}(\infty)] \right\}.$$

- **Comparisons:** every pre-treatment baseline $s < g$ $\times$ every clean cohort $h \in \mathcal{C}(g)$.
- **Weights:** $1/(g-1)$ across baselines; $N_h/N_{\mathcal{C}(g)}$ across clean cohorts.

At impact: a weighted average of ordinary 2×2 DiDs.

Later horizons: where do the weights come from?

Let

◀ Back to talk

$$B_g \equiv \frac{1}{g-1} \sum_{s < g} \bar{Y}_{g,s}(\infty)$$

be cohort g 's average pre-treatment outcome, and define the clean-control movement at date u as

$$\delta_u^C = \sum_{h \in \mathcal{C}(u)} \frac{N_h}{N_{\mathcal{C}(u)}} \left[\bar{Y}_{h,u}(\infty) - \frac{1}{u-1} \sum_{r < u} \bar{Y}_{h,r}(\infty) \right], \quad \mathcal{C}(u) = \{h : h > u\} + \text{never}.$$

At treatment onset,

$$\hat{\bar{Y}}_{g,g}(\infty) = B_g + \delta_g^C.$$

Before moving to $g+1$, this imputed outcome becomes part of the cohort's untreated history:

$$\frac{(g-1)B_g + \hat{\bar{Y}}_{g,g}(\infty)}{g} = B_g + \frac{1}{g}\delta_g^C.$$

Later horizons: where do the weights come from?

◀ Back to talk

So the next imputation is

$$\hat{Y}_{g,g+1}(\infty) = B_g + \frac{1}{g}\delta_g^C + \delta_{g+1}^C.$$

Repeating once more,

$$\hat{Y}_{g,g+2}(\infty) = B_g + \frac{1}{g}\delta_g^C + \frac{1}{g+1}\delta_{g+1}^C + \delta_{g+2}^C.$$

Why the weights? Earlier clean-control movements get diluted because they enter a growing historical average.

The regression's hidden comparison design

Thus, for any $t \geq g$,

[◀ Back to talk](#)

$$\widehat{\text{ATT}}(g, t) = \underbrace{\bar{Y}_{g,t} - B_g}_{\text{treated long difference}} - \underbrace{\left[\delta_t^C + 1\{t > g\} \sum_{u=g}^{t-1} \frac{1}{u} \delta_u^C \right]}_{\text{weighted collection of clean-control movements}}.$$

Each δ_u^C is itself an average over cohorts still untreated at u and over all earlier baselines:

$$\delta_u^C = \sum_{h \in \mathcal{C}(u)} \sum_{r < u} \underbrace{\frac{N_h}{N_{\mathcal{C}(u)}}}_{\text{cohort weight}} \underbrace{\frac{1}{u-1}}_{\text{baseline weight}} [\bar{Y}_{h,u}(\infty) - \bar{Y}_{h,r}(\infty)].$$

So the regression uses a transparent weighting scheme:

- **Across clean cohorts:** size weight $N_h / N_{\mathcal{C}(u)}$.
- **Across baselines within date u :** equal weight $1 / (u - 1)$.
- **Across clean-control movements:** weight 1 for the current date $u = t$, and $1 / u$ for earlier dates.

DIRECT: begin with the empirical question

HRS example: What is the average effect two survey waves after hospitalization?

Choose the horizon $e \geq 0 \rightarrow$ Keep observed cohorts $\rightarrow$ Average their effects

$$\theta_e = ES(e) = \sum_{g \in \mathcal{G}_e} \underbrace{\frac{P(G = g)}{\pi_e}}_{\text{share of eligible patients}} \underbrace{ATT(g, g + e)}_{\text{effect for cohort } g}.$$

- **Key:** g is the first treatment date; $\mathcal{G}_e$ keeps cohorts with observed outcomes at $g - 1$ and $g + e$.
- $G_g = 1\{G = g\}$; $H_e = \sum_{g \in \mathcal{G}_e} G_g$ marks an eligible patient; $\pi_e = E[H_e]$ is their population share.

Recipe below: independent units, unit analysis weights, no trimming or anticipation, and conditional PT-All given pretreatment covariates X .

Start with a familiar before-after change

For a patient hospitalized at wave 8, the two-wave effect uses $Y_{10} - Y_7$.

Last untreated outcome $\longrightarrow$ Outcome at the requested horizon

$$L_{g,e} = \underbrace{Y_{g+e} - Y_{g-1}}_{\text{observed long change}} - \underbrace{m_{g,e}(X)}_{\text{predicted untreated change}}.$$

- **Remove what would happen without treatment.** $\Delta Y_u = Y_u - Y_{u-1}$; $G = \infty$ means never treated. $\mu_u(X) = E[\Delta Y_u \mid G = \infty, X]$ is the untreated mean change.
- **The same calculation, in long-change form:** $m_{g,e}(X) = \sum_{u=g}^{g+e} \mu_u(X)$, so $L_{g,e} = \sum_{u=g}^{g+e} (\Delta Y_u - \mu_u(X))$.

Fit each untreated increment once $\longrightarrow$ add the shared predictions.

$m_{g,e}(X)$ denotes their sum for this cohort; it is not another regression.

Make the untreated comparison relevant to this cohort

Patients hospitalized at different dates may have different characteristics. Reweight never-treated patients to represent the cohort being compared.

$$\rho_g(X) = \frac{p_g(X)}{p_\infty(X)}, \quad p_g(X) = P(G = g \mid X), \quad G_\infty = 1\{G = \infty\}.$$

$$A_e = \sum_{g \in \mathcal{G}_e} \left\{ \underbrace{G_g L_{g,e}}_{\text{treated contribution}} - \underbrace{G_\infty \rho_g(X) L_{g,e}}_{\text{reweighted comparison}} \right\}.$$

- **English key:** A_e combines all eligible cohorts' comparisons for the one event-time question.
- **Why this answers the question:** $E[A_e] = \pi_e \theta_e \Rightarrow$ divide its average by the eligible share.

These probability ratios adjust composition; the precision adjustment comes next.

Earlier untreated comparisons can help remove noise

Compare cohort h with never-treated patients at a date $u < h$: both groups are still untreated.

$$R_{hu} = \underbrace{(G_h - G_\infty \rho_h(X))}_{\text{composition-adjusted comparison}} \underbrace{(\Delta Y_u - \mu_u(X))}_{\text{untreated change residual}} .$$

Valid parallel trends $\longrightarrow E[R_{hu} \mid X] = 0 \longrightarrow$ An adjustment that preserves the target

- **Intuition:** an unusually high pre-treatment comparison may accompany an unusually high target estimate. Subtracting part can reduce variation.
- **The opportunity:** use several credible comparisons to learn which fluctuations predict the target estimate's noise.

Population zero does not mean sample zero; a pretest cannot establish the restriction.

Learn how much of each comparison helps this target

Stack the clean comparisons in R . Multiply each comparison by chosen covariate functions:

$T = B(X) \otimes R$. **Why?** Let the adjustment vary with characteristics; with $B = 1$, $T = R$. Hats mean fitted quantities; $\mathbb{P}_n$ is the sample average.

$$\tilde{\theta}_e = \frac{\mathbb{P}_n \hat{A}_e}{\mathbb{P}_n H_e}, \quad S_e = \hat{A}_e - \tilde{\theta}_e H_e.$$

Initial estimate $\longrightarrow$ Centred contribution $S_e \longrightarrow$ Regress on $\hat{T}$ (no intercept)

$$\hat{z}_e = (\mathbb{P}_n \hat{T} \hat{T}')^{-1} \mathbb{P}_n [\hat{T} S_e].$$

Centre because the final estimate is a ratio: the eligible cohort mix is estimated too. This is the default full-rank projection. Nuisance-fit penalties and the separate partial-menu loading filter are detailed below.

The estimate: initial comparison minus predicted noise

$$\hat{\theta}_e = \frac{\underbrace{\mathbb{P}_n \hat{A}_e}_{\text{initial target comparison}} - \underbrace{\mathbb{P}_n [\hat{Z}'_e \hat{T}]}_{\text{estimated noise adjustment}}}{\underbrace{\mathbb{P}_n H_e}_{\text{eligible treated share}}}.$$

- **The point estimate can change.** The clean comparisons have population mean zero under valid restrictions. Their sample averages can be nonzero.
- **Why DIRECT?** Requested event-time average $\rightarrow$ one combined score $\rightarrow$ its precision adjustment. No intermediate table of cohort-time effects is required.
- **What buys precision?** Credible additional comparisons that predict noise in this target. More useful information can mean a narrower interval.

The same target and model have the same efficiency bound under either route; Direct targeting need not outperform two-step aggregation in every finite sample.

Share information across cohorts; keep their effects different

- **Untreated outcome changes:** Pool the cohorts licensed by the history at date u . Fit $\mu_u(X)$ once for that date; reuse it across cohorts and horizons.
- **Cohort probabilities:** one joint model $\rightarrow$ the ratios used in each comparison. The ratios do not require a separate model for every cohort-time effect.
- **Treatment effects:** cohort effects remain unrestricted. The requested target averages them using its eligible cohort shares.

Three weights, three jobs

Cohort shares: **who the answer represents.**

Probability ratios: **who provides a comparable untreated change.**

Precision coefficients z_e : **which comparisons help predict noise.**

Uncertainty follows the same calculation

$$\hat{\phi}_e = \frac{\hat{A}_e - \hat{z}_e' \hat{T} - \underbrace{\hat{\theta}_e H_e}_{\text{estimated cohort mix}}}{\hat{\pi}_e}, \quad \hat{\pi}_e = \mathbb{P}_n H_e.$$

Leading SE for n independent units $\rightarrow \widehat{\text{se}}(\hat{\theta}_e) = \sqrt{\mathbb{P}_n[\hat{\phi}_e^2] / n}.$

- **This leading formula:** correct probability and outcome models, with valid rate/regularity conditions. Otherwise retain nonnegligible fitting effects.
- **Efficient limit:** adequate overlap, finite moments, full rank, and a basis rich enough for the optimal adjustment.
- **Clustered observations:** use the cluster sampling construction for both the loading objective and uncertainty.

Applications also account for fitted components; finite-sample coverage needs checking.

The covariance tuning, precisely

Use symmetrized $\hat{\Omega}(x)$ on retained coordinates with positive pooled scale.

$$r(x) = \left[\frac{\text{tr} \hat{\Omega}(x)}{n_{\text{eff}}} \right]_+, \quad \hat{V} = \text{diag} \left(\left| \text{diag} \{ \mathbb{P}_n \hat{\Omega}(X) \} \right| \right), \quad \tau(x) = \eta_n \|S(x)\|_{\text{op}}.$$

- **Pointwise:** floor each local covariance, using the same pooled $\hat{V}$.
 $\eta_n = \min(n^{-a}, 0.05)$, $a = 0.07 \{5 - \min(d, 4)\}$.
- **Averaged:** use one marginal covariance; $\eta_n = n^{-1/3}$.
- **Counts:** n is the number of panel units; $n_{\text{eff}} = n$ without analysis weights (otherwise their Kish size).
- **Dimension:** d is the expanded continuous smoothing width; with only discrete covariates, the expanded exact-support width.

Turning off the ridge leaves the floor active. The efficient limit requires covariance consistency, stable rank and nuisance-rate conditions.

Nuisance fits and clustering are separate choices

Probability ratios: in the fixed-series tilt fit, penalize the fitted log-ratio function $f_\beta(X) = B(X)' \beta$:

$$J_n(\beta) = \frac{\log(n)^2}{2n} \mathbb{P}_n[\varpi(X) f_\beta(X)^2].$$

B is the chosen basis; ϖ is its fit weight. Parametric/growing-series fits omit this penalty; exact-support fits bypass it.

- **Ordinary no- X :** PT-All with multiple comparisons uses a trace ridge; no local correlation-scale floor.
- **Exact clustered weights:** optimize covariance of cluster sums; no ordinary ridge or eigenfloor. Refuse an ill-conditioned contrast system.
- **Continuous- X clustered weights:** optimize within a declared series class; the default loading ridge is zero. A positive declared ridge changes that optimum.

Common principle: respect the target and the actual sampling design.

DIRECT: stabilize cohort-probability predictions

One joint probability model supplies the ratios needed by the target and comparisons. Let p count positive-mass objective groups; set $\lambda = 1/p$.

$$\hat{\beta} = \arg \min_{\beta} \left\{ \underbrace{-\mathbb{P}_n \log p_{G_i}(X_i; \beta)}_{\text{fit observed cohort membership}} + \underbrace{\frac{\lambda}{2} \mathbb{P}_n \sum_g (\eta_g(X_i) - \bar{\eta}(X_i))^2}_{\text{discourage extreme log odds}} \right\}.$$

- **Key:** η_g is the cohort logit index; $\bar{\eta}$ averages over all cohort indices, including the reference.
- **Why this penalty?** It controls separation between predictions without depending on the arbitrary reference cohort.

Analysis weights equal one here. By default, p counts units or physical clusters (as appropriate), before trimming.

DIRECT: stabilize untreated-outcome predictions

For each needed increment, use its permitted untreated pool. Let $B_i = B(X_i)$ be the chosen basis, $\bar{B} = \mathbb{P}_n B_i$, and $\lambda = 1/p$ as above.

$$\hat{\gamma}_u = \arg \min_{\gamma_u} \left\{ \underbrace{\mathbb{P}_n [I_{ui} (\Delta Y_{ui} - B_i' \gamma_u)^2]}_{\text{fit the clean untreated changes}} + \underbrace{\lambda \mathbb{P}_n [\{(B_i - \bar{B})' \gamma_u\}^2]}_{\text{shrink variation in predictions}} \right\}.$$

- **Key:** I_{ui} selects the untreated training pool for date u . The centring $\bar{B}$ uses the full training sample.
- **Why centre the penalty?** It shrinks covariate-driven variation while leaving a constant prediction unpenalized.
- **Reuse the fit:** pool where the history permits; use the same fitted increment in every target that needs it.

These are nuisance penalties; the loading step comes next.

DIRECT: the loading step is a different operation

For the centred target contribution S_e and comparison vector $\hat{T}$:

$$\hat{z}_e = \arg \min_z \mathbb{P}_n[(S_e - z' \hat{T})^2].$$

- **Default full menu:** scaled QR solves this unpenalized projection. It retains all declared directions if full rank; otherwise it refuses the fit.
- **A declared subset of comparisons:** a separate iterated-ridge filter stabilizes its loading inverse. The default full menu does not use it.
- **Eligible single-cohort, no- X requests:** both routes use the same cohort-time construction, as in our supported TVA specification.

Inference follows the fitted map: nuisance predictions, pilot centring, loadings and cohort shares, with their covariance terms and declared cluster law.

The theorem's weights: an equivalent characterization

The SEB derivation comes first. It establishes the efficient influence function in the model's comparison span. Its weights also solve

$$w^*(x) = \arg \min_{1'w=1} w' \Omega(x) w, \quad \Omega(x) = \text{Cov}(H \mid X = x),$$

where H stacks the theorem's comparison influence functions.

Why scale and common noise cancel. For $\Omega_0(x)$ positive definite, $a(x) > 0$ and $c(x) \geq 0$, suppose

$$\Omega(x) = a(x)\Omega_0(x) + c(x)11'.$$

Since $1'w = 1$,

$$w' \Omega(x) w = a(x) w' \Omega_0(x) w + c(x) \implies w^*(x) = \frac{\Omega_0(x)^{-1} 1}{1' \Omega_0(x)^{-1} 1}.$$

Interpretation: a common scale changes the variance level; shared noise cannot be diversified away. Neither changes the optimal relative weights.

A shared IF component must be conditionally orthogonal to the remaining noise for the $c(x)11'$ form. Fix the target and comparison set. Estimated, stabilized weights need not preserve exact finite-sample invariance.

References

- Ai, Chunrong and Xiaohong Chen**, “The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions,” *Journal of Econometrics*, 2012, 170 (2), 442–457.
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess**, “Revisiting Event Study Designs: Robust and Efficient Estimation,” *Review of Economic Studies*, 2024, 91 (6), 3253–3285.
- Callaway, Brantly and Pedro H. C. Sant’Anna**, “Difference-in-Differences with multiple time periods,” *Journal of Econometrics*, 2021, 225 (2), 200–230.
- Chaudhuri, Saraswata**, “Targeted Feasible Generalized Least Squares,” 2026. Working paper, June 14.

References ii

- Chen, Xiaohong and Andres Santos**, “Overidentification in Regular Models,” *Econometrica*, 2018, 86 (5), 1771–1817.
- de Chaisemartin, Clément and Xavier D’Haultfœuille**, “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects,” *American Economic Review*, 2020, 110 (9), 2964–2996.
- Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo**, “The economic consequences of hospital admissions,” *American Economic Review*, 2018, 108 (2), 308–352.
- Gardner, John**, “Two-stage differences in differences,” *Working Paper*, 2021.
- Harmon, Nikolaj A.**, “Difference-in-Differences and Efficient Estimation of Treatment Effects,” *Working Paper*, 2023.
- Kline, Patrick and Enrico Moretti**, “Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority,” *Quarterly Journal of Economics*, 2014, 129 (1), 275–331.

- Sant'Anna, Pedro H. C. and Jun Zhao**, “Doubly robust difference-in-differences estimators,” *Journal of Econometrics*, 2020, 219 (1), 101–122.
- Sun, Liyang and Sarah Abraham**, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 2021, 225 (2), 175–199.
- Wooldridge, Jeffrey M.**, “Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators,” *Empirical Economics*, 2025, pp. 2545–2587.
- , “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” *Empirical Economics*, 2025, 69 (5), 2545–2587.