Efficient Difference-in-Differences and Event Study Estimators

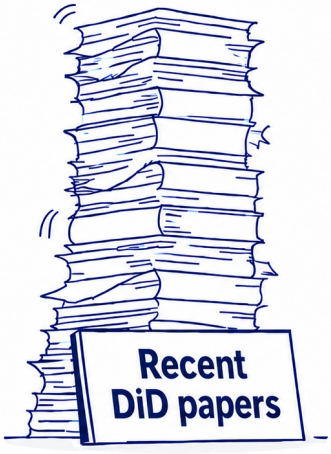
Xiaohong Chen
Yale University

Pedro H. C. Sant'Anna
Emory University

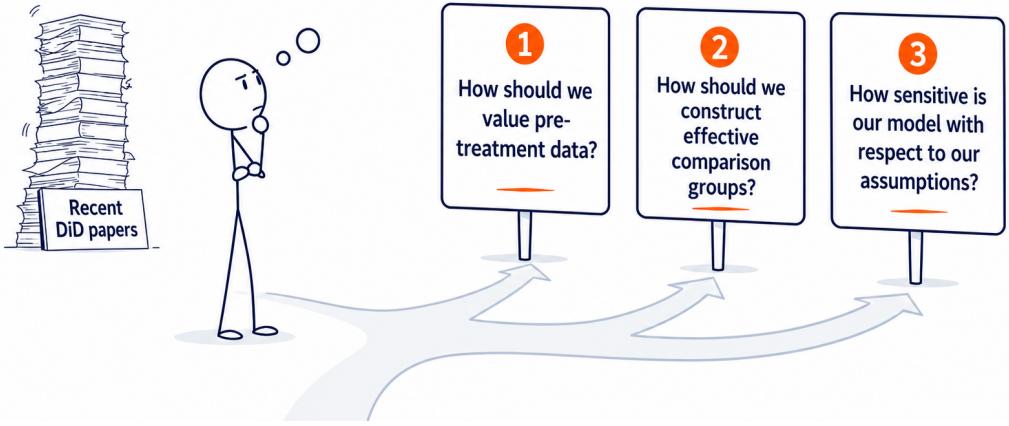
Haitian Xie
Peking University

EEA-ESEM 2026
Dublin
August 2026

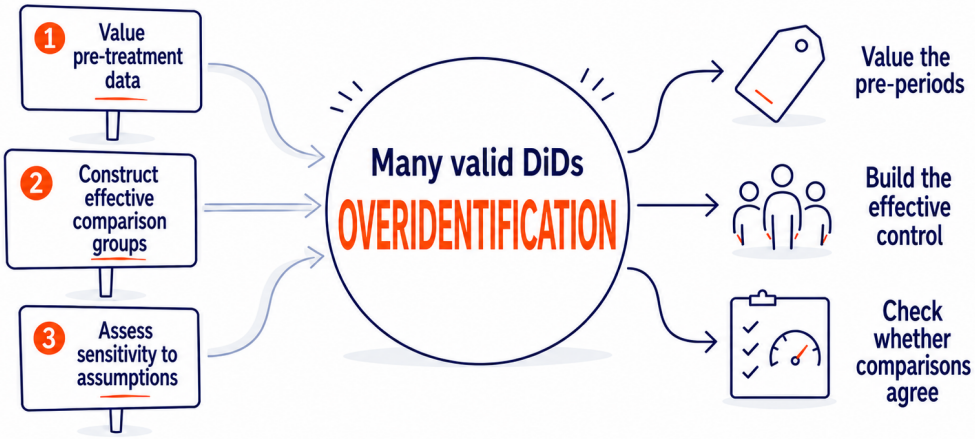
DiD has evolved so much in recent years...



But so many basic questions remain



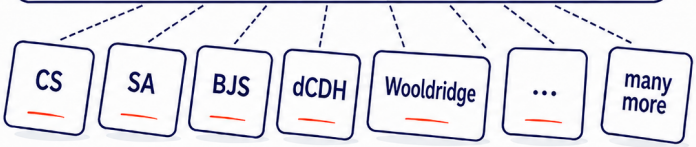
DiD is often nonparametrically overidentified



This paper leverages NP overidentification to better understand DiD

What does this mean?

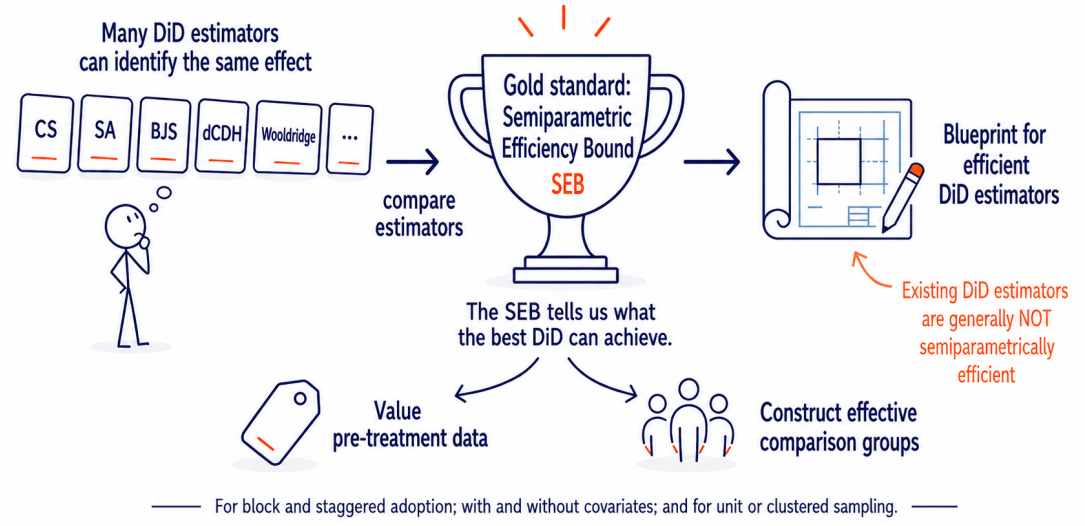
Under (conditional) PTs, many different DiD estimators can identify the same effect



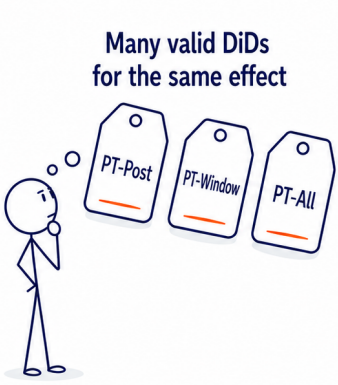
The familiar procedures are only some of the possibilities

How should we choose between the many options?

We need a gold standard for comparison



Overidentification also gives us diagnostics



1 **Restriction-aligned pre-treatment ES**
Do the comparisons agree before treatment?

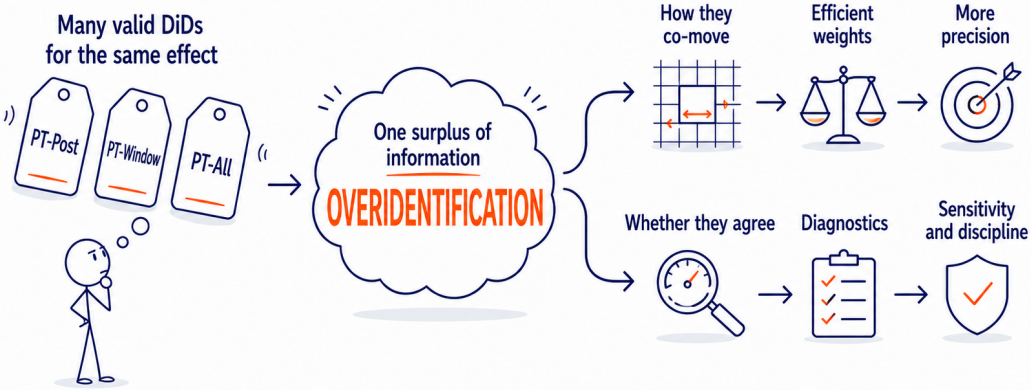
2 **Directed comparison**
Does added history move
the reported effect?
H_{RU}, J-statistic, incremental Sargan

3 **Precision–sensitivity frontier**
What is the trade-off?
Andrews et al. (2025)

Overidentification helps us assess the coherence of our design.

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

One surplus, two uses



Declare the history first. Then use it efficiently—and show its sensitivity.

Efficiency does not choose the model. Diagnostics do not validate parallel trends.

What the 9 applications reveal

ATT(g,t) • ES(e) • scalar summaries

1 Large precision gains

1.11–3.39×
lower variance
across 9 designs

or 11–51% lower MDE

2 Added history can help — or change the answer

7 of 9
stable across
the ladder

2 of 9
meaningfully
history-sensitive

Precision and credibility are not the same thing.

3 Diagnostics keep them apart


Pre-treatment ES


Directed comparisons


Precision–sensitivity frontier

They show when added history matters.

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

A taste of the econometrics

DiD framework

- "Short" panel $\{W_i\}_{i=1}^n = \{(Y_{i,1}, \ldots, Y_{i,T}, X_i', G_i)'\}_{i=1}^n$: 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; $G_i \in \mathcal{G} \subseteq \{2, \ldots, T, \infty\}$.
  - Single date: $G_i = g$ (treated) or $G_i = \infty$ (untreated). Let $G_g = \mathbf{1}\{G = g\}$.

Target parameters

- 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 ($e = t - g$):

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

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

Assumption (Maintained Assumption (M))

(i) *(S)* $\{(Y_{i,t=1}, \ldots, Y_{i,t=T}, X_i', G_i)'\}_{i=1}^n$ *is a random sample from* $(Y_{t=1}, \ldots, Y_{t=T}, X', G)'$.

(ii) *(O) For each* $g \in \mathcal{G}$, $\mathbb{E}[G_g|X] \in (0,1)$ *almost surely (a.s.).*

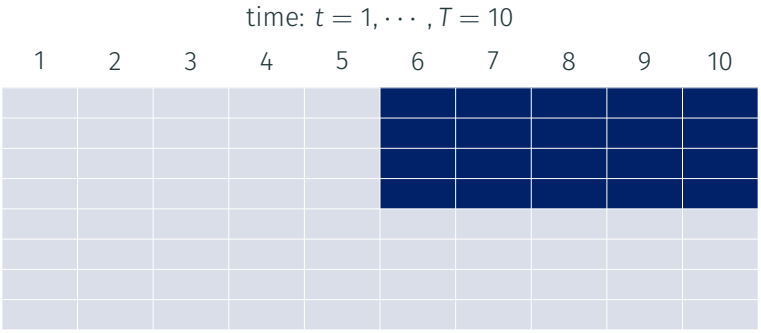
(iii) *(NA) For every* $g \in \mathcal{G}_{trt}$, *and every pre-treatment periods* $t < g$, $\mathbb{E}[Y_{i,t}(g)|G = g, X] = \mathbb{E}[Y_{i,t}(\infty)|G = g, X]$ *almost surely.*

- 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 $t \in \{2, \ldots, T\}$ such that $t \geq g$,

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

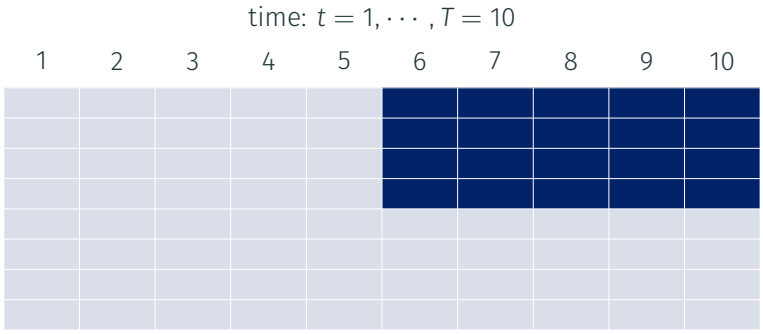
- Impose parallel trends only for post-treatment periods.
- Uses only period $g - 1$ as the baseline.
- Without covariates, easy to show that, for any $t \geq g$,

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

- 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.

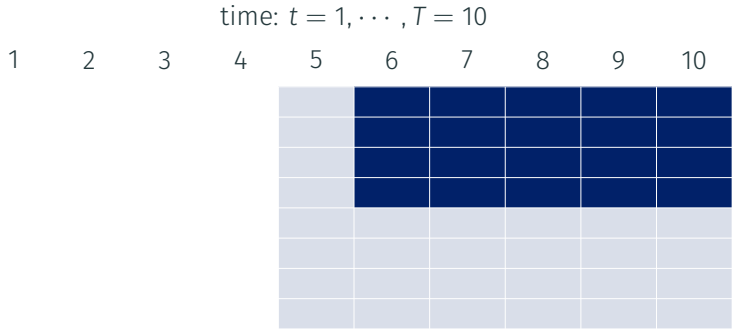
PT-Post: Implications

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



PT-Post: Implications

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



Parallel Trends Assumption: All periods

Assumption (PT in all periods)

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

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

- 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* $t \in \{g, \ldots, T\}$*, with prob. one,*

$$\mathbb{E}[G_g(ATT(g,t) - CATT(g,t,X))] = 0,$$

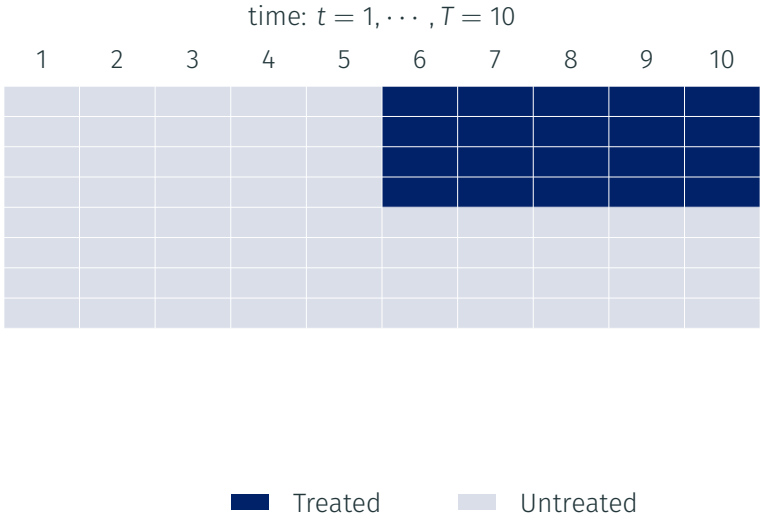
$$\mathbb{E}\left[CATT(g,t,X) - \frac{G_g(Y_t - Y_{g-1})}{p_g(X)} + \frac{G_\infty(Y_t - Y_{g-1})}{p_\infty(X)} \middle| X\right] = 0,$$

$$\mathbb{E}\left[\frac{G_g(Y_{t'} - Y_1)}{p_g(X)} - \frac{G_\infty(Y_{t'} - Y_1)}{p_\infty(X)} \middle| X\right] = 0, \text{ for all } 2 \le t' \le g-1,$$

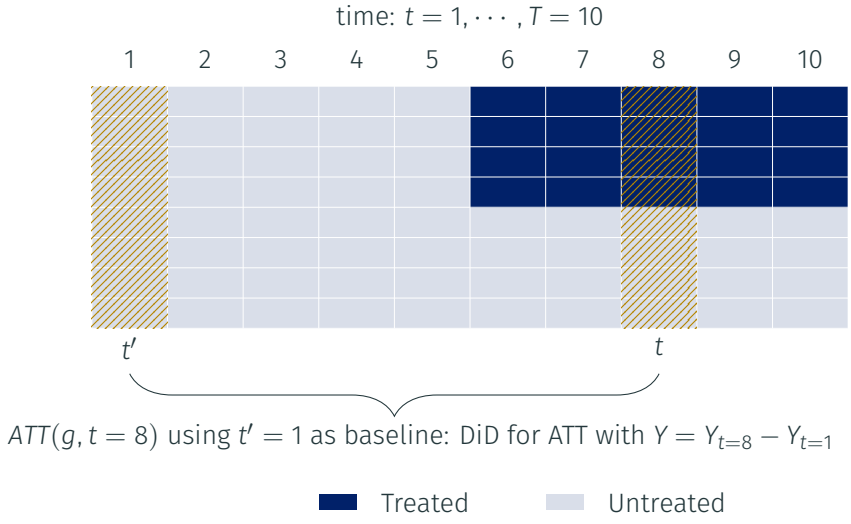
$$\mathbb{E}[G_g - p_g(X)|X] = 0.$$

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

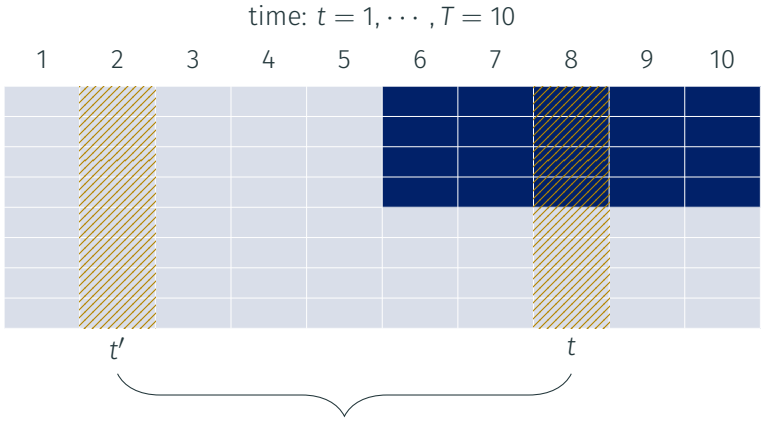
Understanding the sources of over-identification



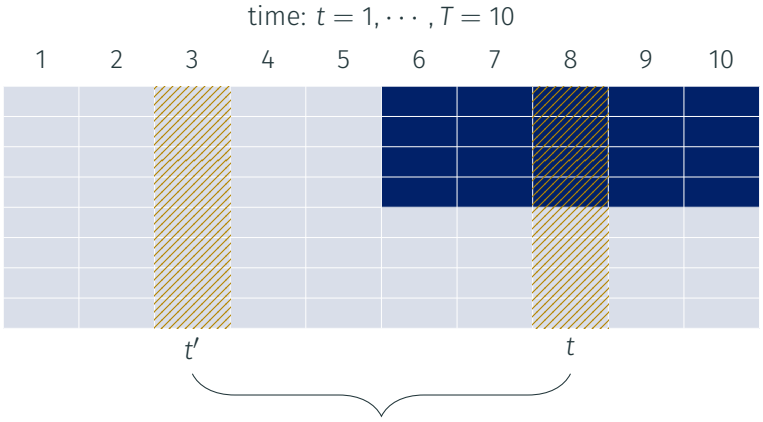
Understanding the sources of over-identification



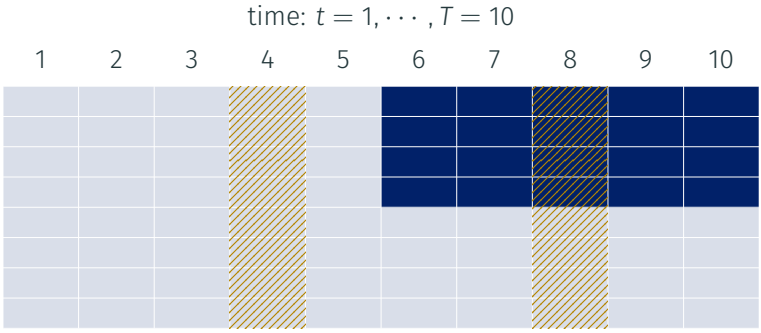
Understanding the sources of over-identification



Understanding the sources of over-identification



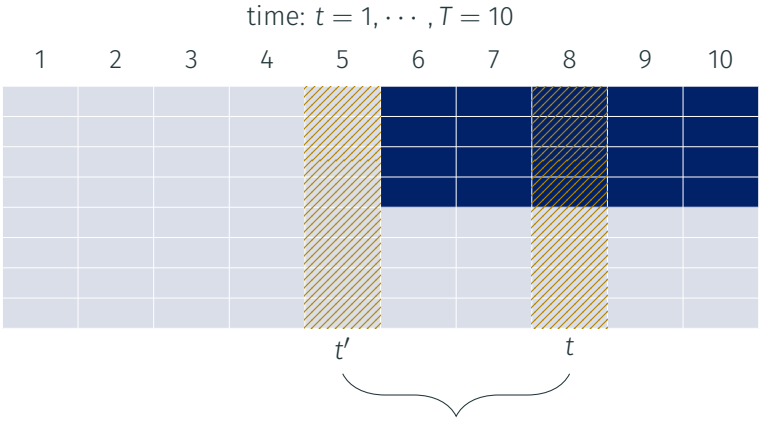
Understanding the sources of over-identification



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

Treated Untreated

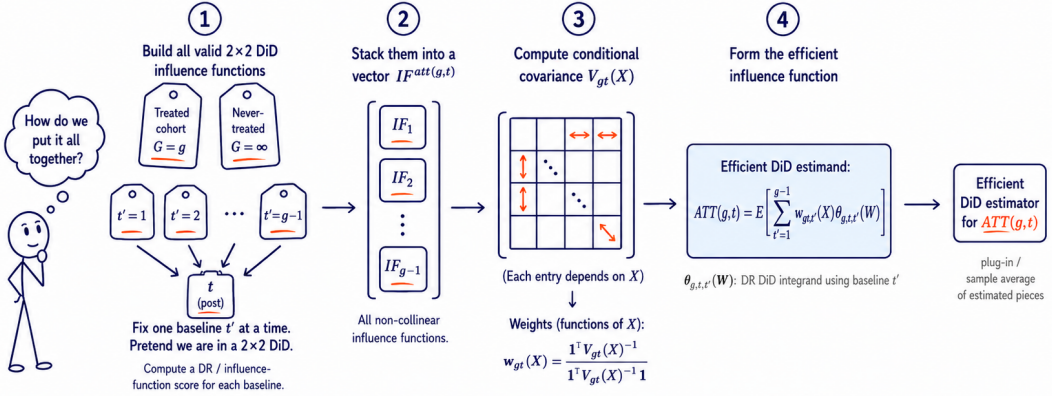
Understanding the sources of over-identification



$ATT(g, t = 8)$ 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



g: index of treated cohort $(G = g)$
t: post-treatment period of interest
$t' \in \{1, \ldots, g-1\}$: baseline (pre-treatment) period
X: observed covariates
$W = (Y, G, X)$: observed data

$IF_{t'}(W)$: influence function from the 2 × 2 DiD using baseline t'
$IF^{att(g,t)}$: vector of length $g-1$

$V_{gt}(X) = \mathrm{Cov}(IF^{att(g,t)} | X)$: $(g-1) \times (g-1)$ matrix
1: $(g-1) \times 1$ vector of ones
$w_{gt}(X)$: $(g-1) \times 1$ weight vector (sums to 1)

$\theta_{g,t,t'}(W)$: DR DiD integrand using baseline t' $(= IF_{t'}(W)$ up to a known scale$)$
$EIF^{att(g,t)}(W)$: efficient influence function

Semiparametric efficiency is achieved by optimally weighting all valid 2×2 DiD influence functions.

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:

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

where $p_g(X) = \mathbb{E}[G_g|X]$, $\theta_{g,t}(W) = (\theta_{g,t,1}(W), \ldots, \theta_{g,t,g-1}(W))'$ 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: apply plug-in principle, or do DML.

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:

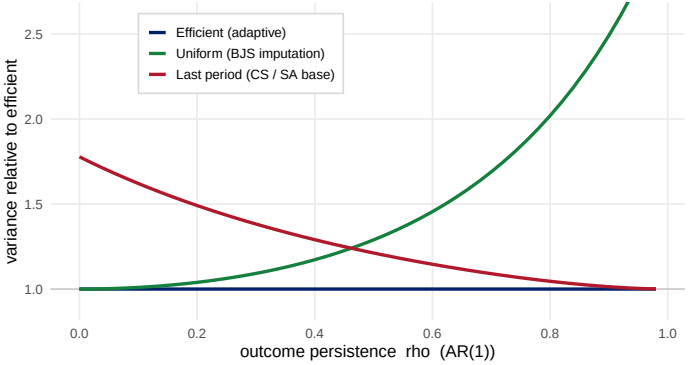
$$\omega^\star \;=\; \Sigma^{-1}\Big[\sigma_{\text{pre,post}} \;+\; \frac{1-\mathbf{1}'\Sigma^{-1}\sigma_{\text{pre,post}}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}\,\mathbf{1}\Big], \qquad \Sigma = \mathrm{Var}(u_{\text{pre}}).$$

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 $\omega^\star(e)$	matches the baseline 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 $3\times$ near a unit root.
- **CS, SA, dCDH, MS, Harmon** difference against the last period: efficient as $\rho \to 1$; "waste" 78% under iid.

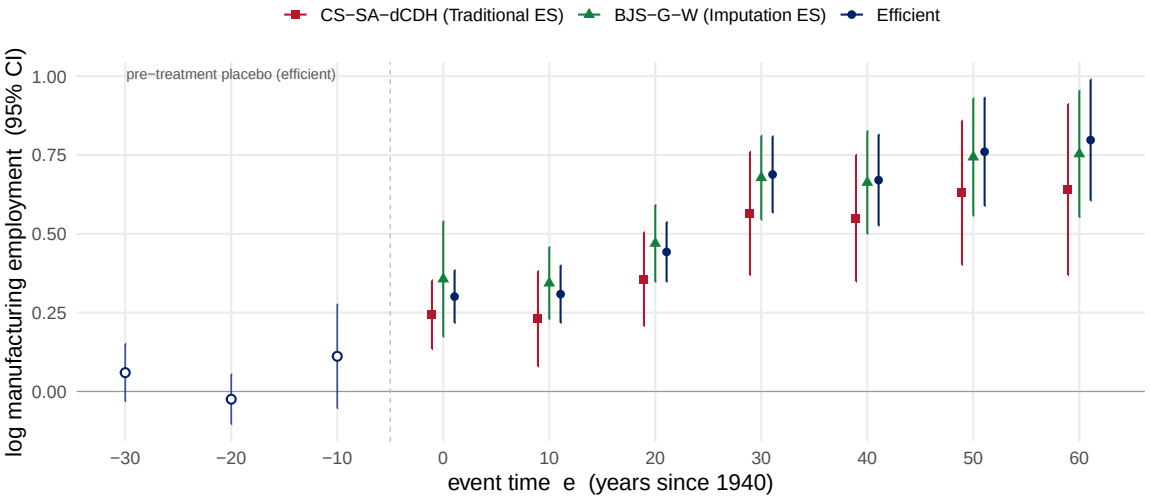
The bound is the lower envelope — optimal at every ρ.

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** — the efficient weights use the unconditional cluster-covariance of the pre-period changes (saturated cells, no working models to defend). SEs **state**-clustered, as in K–M. Four decennial baselines, 1900–1930.
- One date + a long pre-window $\Rightarrow$ baseline pooling is the whole efficiency story.

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



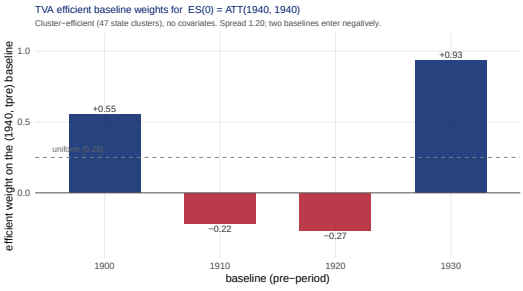
Post−treatment average (s.e.), asymptotic relative efficiency, and 80%−power MDE. Pre−treatment placebo path jointly consistent with zero (p = 0.269).

Traditional: 0.459 (0.089) · ARE 2.30 · MDE 0.248

Imputation: 0.573 (0.070) · ARE 1.42 · MDE 0.195

Efficient: 0.567 (0.058) · ARE 1.00 · MDE 0.164

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



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).

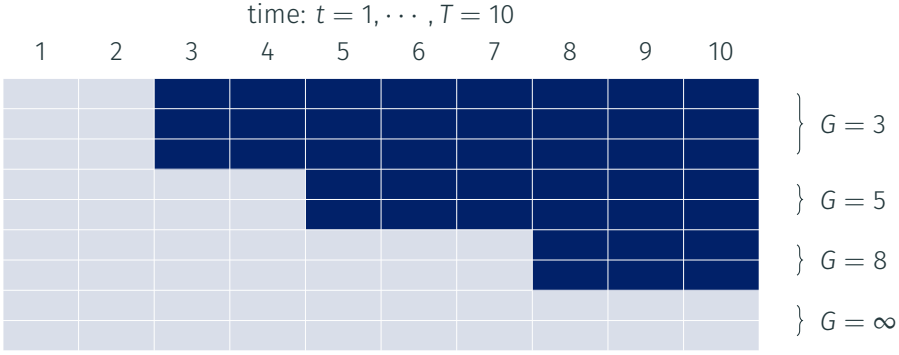
Gain at $ES(0)$**.** These efficient weights buy ARE $= 1.71\times$ vs. Trad. ES=last-pre (a 23% shorter CI) and $4.84\times$ vs. Imputation ES (a 55% shorter CI). horizon.

These ARE gains 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–$2.8\times$ band, while imputation ranges from $4.8\times$ at impact down to $1.1\times$ 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 $ATT(g, t)$.



PT for all periods

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

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

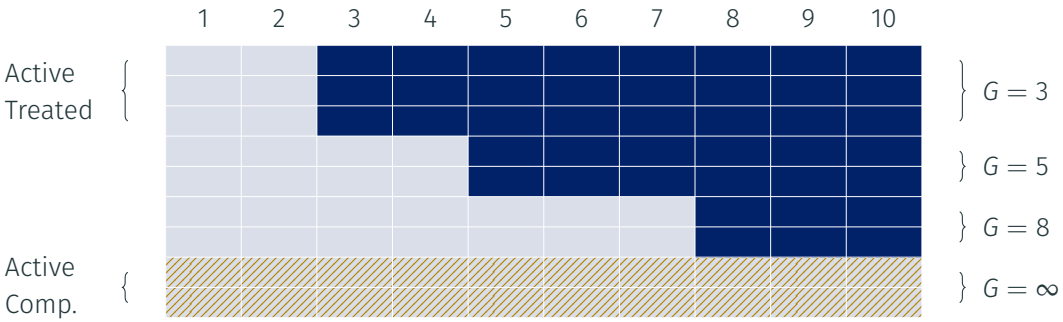
$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = g', X] \text{ a.s.}$$

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.

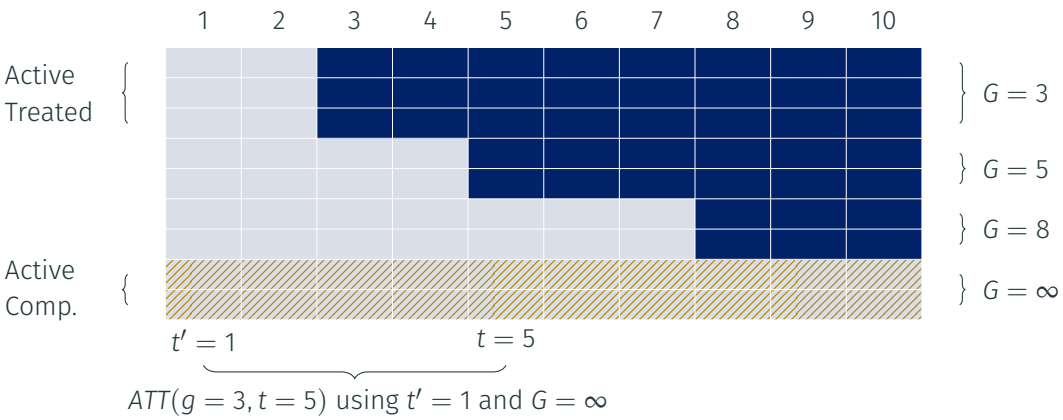
Understanding the sources of over-identification

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



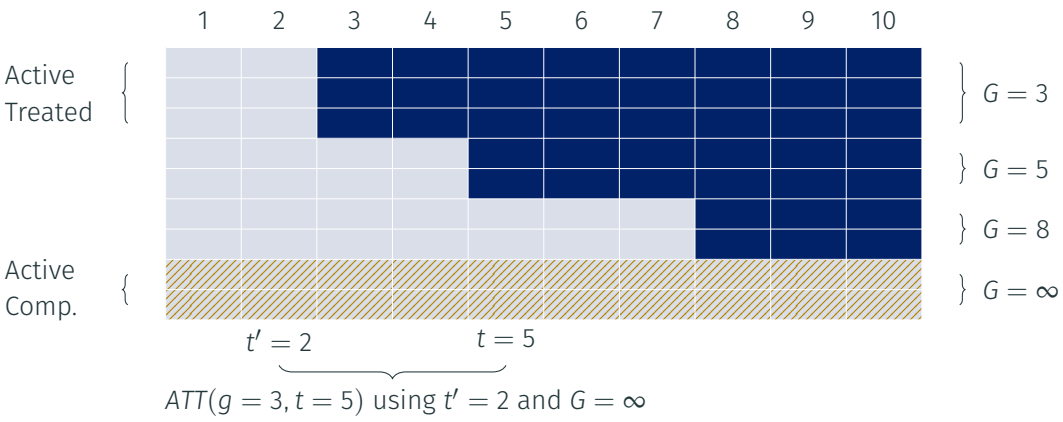
Understanding the sources of over-identification

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



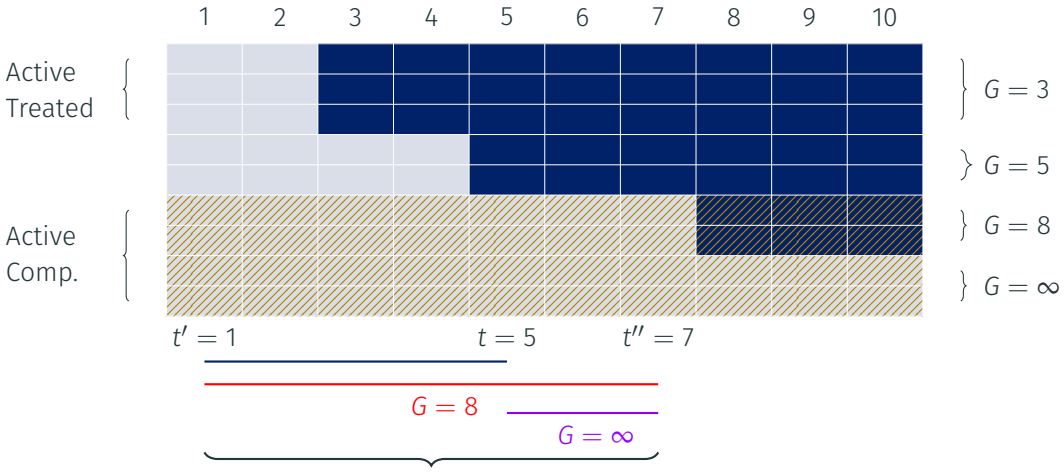
Understanding the sources of over-identification

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



Understanding the sources of over-identification

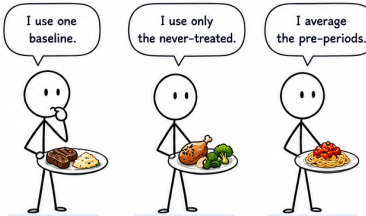
$ATT(g = 3, t = 5)$, using two active comparison groups $G = 8$, and $G = \infty$:



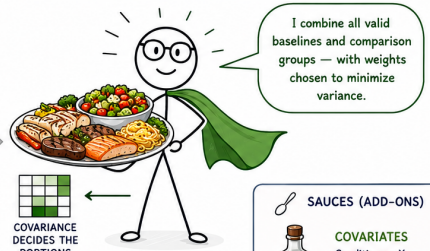
The DiD Buffet

Same ATT(g,t). More information. More precision.

PICK ONE. LEAVE THE REST.



USE EVERY VALID COMPARISON. NOT IN EQUAL PORTIONS.



BASELINE OPTIONS

Earlier baselines

Recent baselines

COMPARISON GROUP OPTIONS

Never-treated units

Not-yet-treated cohorts

SAUCES (ADD-ONS)

COVARIATES

Condition on X. Change the recipe by subgroup.

CLUSTERING

Changes the covariance used to choose weights. Can affect point estimates.

ONE TARGET (ATT(g,t)). MANY VALID COMPARISONS. USE THEM ALL — OPTIMALLY.

Lemma (Over-identification in staggered designs)

Under Assumptions M and PT-All, for every group $(g, g') \in \mathcal{G}_{trt} \times \mathcal{G}_{trt}$ *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,*

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

and, as a consequence,

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

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

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

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

treated DR piece ($G = g$)
uses the treated cohort $G = g$ and outcome-regression residuals.
Target cohort g: $Y_t - Y_1$

never-treated bridge ($G = \infty$)
uses the ratio $p_g(X)/p_\infty(X)$ to reweight never-treated units.
Never-treated bridge: $Y_t - Y_{t''}$

comparison-cohort bridge ($G = g'$)
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

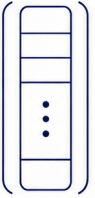
$$p_h(X) = \Pr(G = h \mid X) \qquad m_{a,b,c}(X) = \mathbb{E}[Y_c - Y_a \mid G = a, X]$$

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

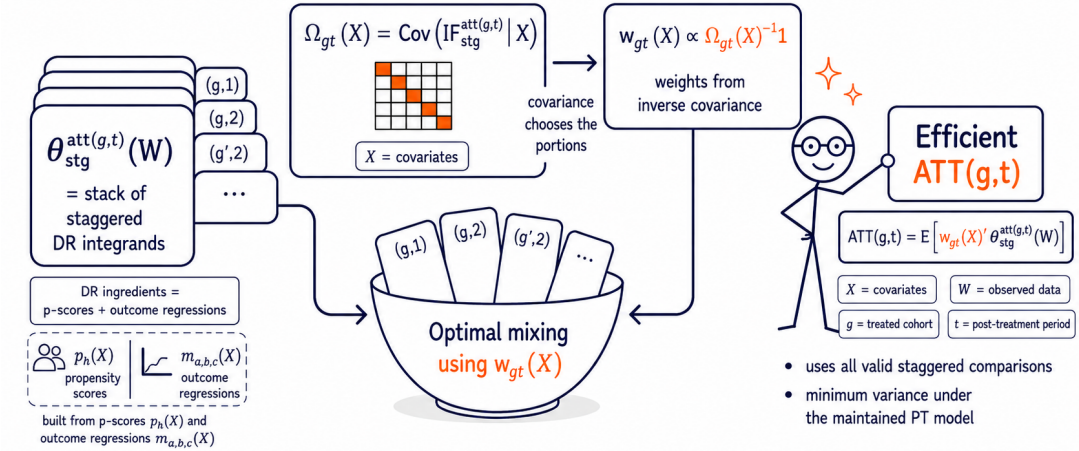
$IF_{stg}^{att(g,t)}$ = **stack of all non-collinear IF terms**



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

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

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

Efficient Staggered DiD

- Here follows the efficient staggered DiD estimand

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

where $\Omega_{gt}(X) = \text{Cov}(\mathbb{1}\mathbb{F}_{\text{stg}}^{att(g,t)}|X)$, $\theta_{\text{stg}}^{\text{att(g,t)}}(W)$ is the column vector

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

such that, for $g' = g$,

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

and, for $g' \neq g$,

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

with

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

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

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

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

See our paper for the efficient influence function of a $ES(e)$, $e \geq 0$.

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).

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:

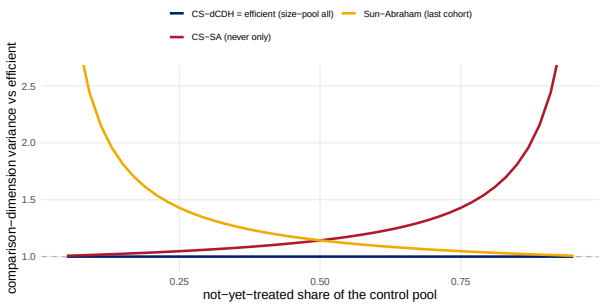
$$w_{i,s}(g,t) \approx \underbrace{\omega_s(g)}_{\text{persistence (time)}} \times \underbrace{\psi_{g(i)}(g,t)}_{\text{precision (groups)}}$$

- 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**:

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



- 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 (next).
- The bound pools the comparison by precision and optimizes the baseline.

Getting more intuition around imputation-type DiD estimators

For any $\mathrm{ATT}(g,t)$, imputation-type estimators can be written in closed form — the base case with one clean time-jump per extra period, no inverse, no sweep:

$$\widehat{\mathrm{ATT}}(g,t) = \underbrace{\left[\bar{Y}_{g,t} - \tfrac{1}{g-1}\sum_{s<g}\bar{Y}_{g,s}\right]}_{\text{treated own long-difference}} - \underbrace{\left[\Delta_{t-1} + 1\{t>g\}\sum_{j=g-1}^{t-2}\tfrac{1}{j+1}\Delta_j\right]}_{\text{untreated time movement (controls)}}.$$

The **clean time-jump** Δ_j is a size-weighted average over the cohorts still untreated at $j{+}1$ — pool $\mathcal{C}(j{+}1){=}\{g'{>}j{+}1\}{+}$never:

$$\Delta_j = \sum_{g'\in\mathcal{C}(j+1)} \frac{N_{g'}}{N_{\mathcal{C}(j+1)}}\left(\bar{Y}_{g',j+1} - \tfrac{1}{j}\sum_{r\le j}\bar{Y}_{g',r}\right).$$

It measures how far period $j{+}1$ sits above the pool's own 1:j average

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 panel spanning waves 7–11 (2004–2012).
  - Final sample: 652 individuals in 4 treatment groups — $G_i = 8$ (252), $G_i = 9$ (176), $G_i = 10$ (163), $G_i = \infty$ (only **65** never-treated).

- Small comparison group $\Rightarrow$ **power is binding**, and precision matters.

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	3072	1112	1038	3063	90	2908	3024	692	1038	1585
	(806)	(637)	(817)	(690)	(641)	(894)	(486)	(471)	(816)	(521)
CS-SA	2826	825	800	3031	107	3092	2960	530	800	1430
	(1035)	(909)	(1008)	(702)	(651)	(995)	(539)	(585)	(1008)	(647)
CS-dCDH	3029	1248	800	3324	107	3092	3134	779	800	1571
	(913)	(861)	(1008)	(959)	(651)	(995)	(536)	(570)	(1008)	(566)
BJS-G-W	3029	1285	1021	3239	77	2758	3017	788	1021	1609
	(916)	(767)	(851)	(862)	(729)	(957)	(555)	(587)	(851)	(582)

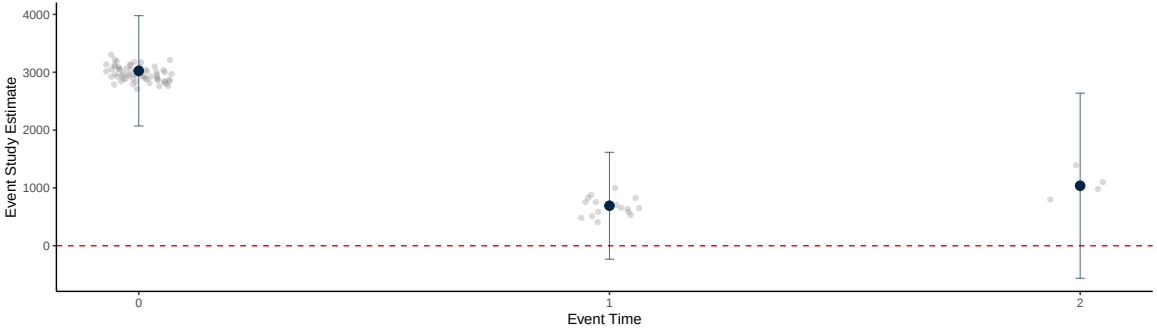
Our efficient DiD delivers substantial gains in efficiency

- Asymptotic relative efficiency (ARE) of EDiD vs. the other estimators — heuristically, the **relative sample size** they would need to match EDiD's precision.

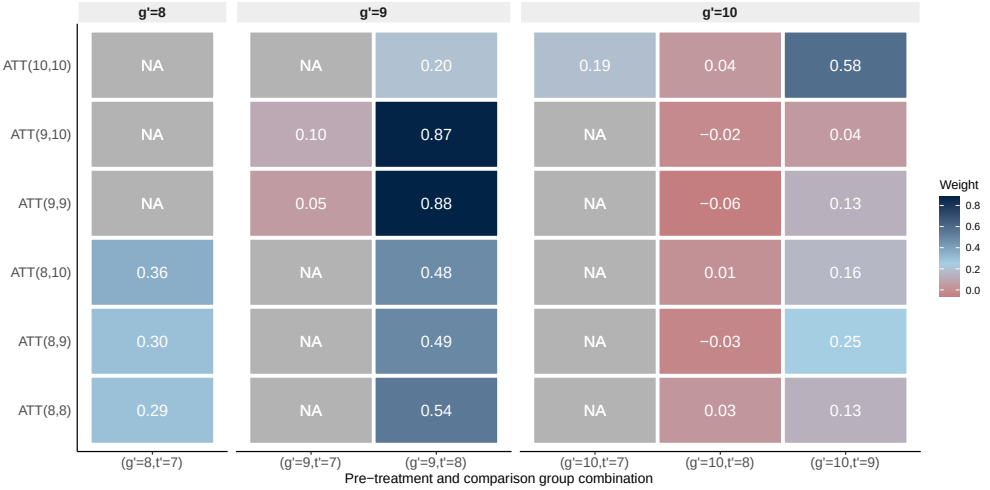
Estimator	$ATT(8,8)$	$ATT(8,9)$	$ATT(8,10)$	$ATT(9,9)$	$ATT(9,10)$	$ATT(10,10)$	$ES(0)$	$ES(1)$	$ES(2)$	ES_{avg}
EDiD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CS-SA	1.65	2.04	1.52	1.04	1.03	1.24	1.23	1.54	1.52	1.54
CS-dCDH	1.28	1.83	1.52	1.93	1.03	1.24	1.21	1.46	1.52	1.18
BJS-G-W	1.29	1.45	1.09	1.56	1.29	1.15	1.30	1.55	1.09	1.25

- Several alternatives need **up to twice the sample size** (e.g. $ATT(8,9)$: ARE $= 2.04$) for the same precision; the ranking is not fixed across parameters.

Visualizing event-study stability



Understanding the efficient weights for the $ATT(g,t)$'s



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 1,010 respondents to do what the efficient estimator does with 656; across eight published designs the equal-precision gains run from $+1\%$ to $+735\%$.

Thank you!

overdid — the R implementation

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.

Chen, Xiaohong and Andres Santos, "Overidentification in Regular Models," *Econometrica*, 2018, *86* (5), 1771–1817.

References ii

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.

References iii

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.