Causal Inference using Difference-in-Differences

Anzony Quispe[1]

July 19, 2025

[1] This material is based on Pedro Sant'Anna and Gemma Dipoppa's class.

Overview

1. Motivation
2. Basic DiD Setup
3. Assumptions
4. Event Studies
5. TWFE

Causal Effect

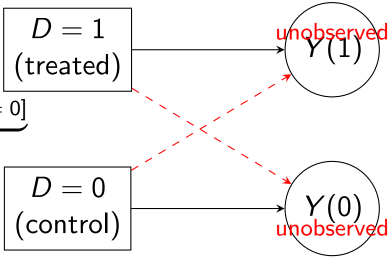
PRIVATE SCHOOL
PUBLIC SCHOOL

Selection Bias: A Formal Illustration

Difference in observed outcomes

$$\underbrace{\mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0]}_{\text{naïve difference}} = \underbrace{\mathbb{E}[Y(1) - Y(0)]}_{\text{Average Treatment Effect (ATE)}} + \underbrace{\mathbb{E}[Y(0) \mid D=1] - \mathbb{E}[Y(0) \mid D=0]}_{\text{Selection Bias}}$$

- $D \in \{0,1\}$ is the treatment indicator.
- $Y(1)$, $Y(0)$ are potential outcomes *with* and *without* treatment.
- The ATE is not directly observed; we only see $Y = D\,Y(1) + (1-D)\,Y(0)$.
- If $\mathbb{E}[Y(0) \mid D=1] \neq \mathbb{E}[Y(0) \mid D=0]$, the naïve difference conflates the treatment effect with pre-treatment differences.



Key takeaway:
Without random assignment
($D \perp Y(0), Y(1)$),
selection into treatment biases simple comparisons.

DiD Trend

- Better computing resources, easy-to-use software.
- Budget constraint for experiments, unavailability of experimental data.
- With observational data, we have no choice but rely on assumptions to conduct causal inference.

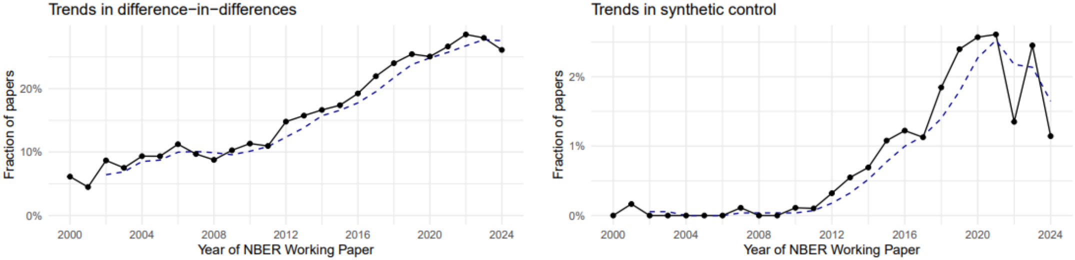


Figure: DiD and Synthetic Control Trend

Observational Data

Assuming **unconfoundedness (either unconditionally or conditional on observed covariates)**, we can estimate the treatment effect using:

- regression
- matching
- reweighting
- double machine learning

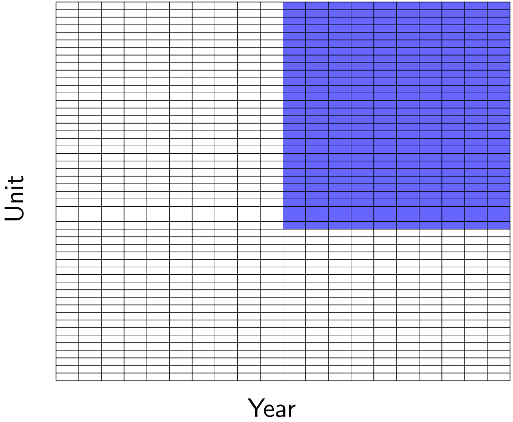
DiD Advantage: Allow for selection on unobservables and time-trends.

Main Assumption

Parallel Trends

Absent the treatment and conditional on covariates (features), the outcome of interest would evolve similarly across treated and control groups.

Panel View: Treatment Adoption (2000–2019)



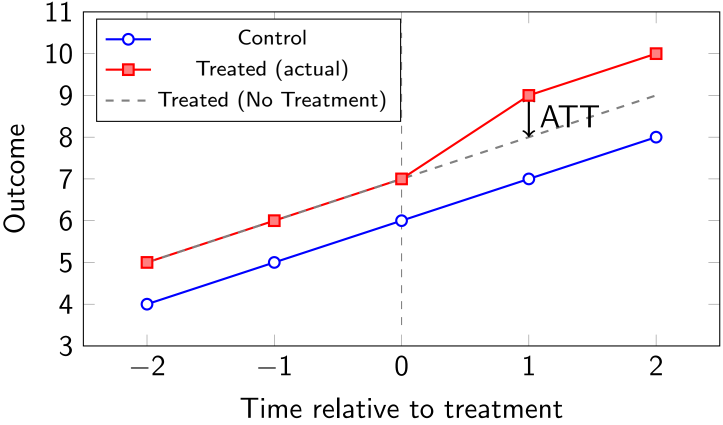
The canonical 2 × 2 DiD estimator

$$\begin{aligned}\hat{\theta}^{\mathrm{DiD}} &= \left(\bar{Y}_{g=\text{treated},\, t=\text{post}} - \bar{Y}_{g=\text{treated},\, t=\text{pre}}\right) \\ &\quad - \left(\bar{Y}_{g=\text{untreated},\, t=\text{post}} - \bar{Y}_{g=\text{untreated},\, t=\text{pre}}\right),\end{aligned}$$

DiD Illustration: Table

	Before	**After**	**Diff (Aft–Bef)**
Control	1.5	4	2.5
Treat	3.5	12	8.5
Diff (T–C)	2	8	**6**

Difference-in-Differences Illustration with Counterfactual



Stable Unit Treatment Value Assumption (SUTVA)

Assumption (SUTVA)

Observed outcomes at time t are realized as

$$Y_{i,t} = \sum_{g \in \mathcal{G}} \mathbf{1}\{G_i = g\}\ Y_{i,t}(g).$$

- Implicitly implies that potential outcomes for unit i are not affected by the treatment of unit j.
  - Rules out interference across units
  - Rules out spillover effects
  - Rules out general equilibrium effects

No-Anticipation Assumption

Assumption (No-Anticipation)

For all units i,

$$Y_{i,t}(g) \;=\; Y_{i,t}(\infty) \quad \textit{for all } t < g$$

i.e. in every pre-treatment period.

- Replace all "untreated" (or "not-yet-treated") potential outcomes by $Y_{i,t}(\infty)$.
- Many times, this assumption is already "baked" into the potential-outcome notation (replace $Y_{i,t}(\infty)$ with $Y_{i,t}(0)$ in all pre-treatment periods).
- Potential outcome of control Group is equivalent to treatment group pre treatment.

Parallel Trends Assumption – 2×2 setup

Since a simple comparison of means at time $t = 2$ does not recover a parameter of interest (ATT), we can take a different route.

Assumption (Parallel Trends Assumption)

$$\mathbb{E}\big[Y_{i,t=2}(\infty) \mid G_i = 2\big] - \mathbb{E}\big[Y_{i,t=1}(\infty) \mid G_i = 2\big] \; = \; \mathbb{E}\big[Y_{i,t=2}(\infty) \mid G_i = \infty\big] - \mathbb{E}\big[Y_{i,t=1}(\infty) \mid$$

The parallel trends (PT) assumption states that, in the absence of treatment, the evolution of the outcome among the treated units is, on average, the same as the evolution among the untreated units.

Parallel Trends and the ATT

- We will start from the perspective that the *ATT* at time $t = 2$ is the target parameter.
- From the definition of the ATT and SUTVA, we have

$$\begin{aligned} \mathrm{ATT} &\equiv \mathbb{E}\big[Y_{i,t=2}(2) \mid G_i = 2\big] - \mathbb{E}\big[Y_{i,t=2}(\infty) \mid G_i = 2\big] \\ &= \underbrace{\mathbb{E}\big[Y_{i,t=2} \mid G_i = 2\big]}_{\text{by SUTVA}} - \mathbb{E}\big[Y_{i,t=2}(\infty) \mid G_i = 2\big] \end{aligned}$$

- Green object is estimable from data (under SUTVA).
- Red object still depends on potential outcomes, and we aim to find ways to "impute" it.
- This is where PT comes into play!

Parallel Trends and the ATT

1. First, recall the PT assumption:

$$\mathbb{E}\big[Y_{i,t=2}(\infty)\,|\,G_i=2\big]-\mathbb{E}\big[Y_{i,t=1}(\infty)\,|\,G_i=2\big]=\mathbb{E}\big[Y_{i,t=2}(\infty)\,|\,G_i=\infty\big]-\mathbb{E}\big[Y_{i,t=1}(\infty)\,|\,G_i=\infty\big].$$

2. By simple manipulation, we can write it as

$$\mathbb{E}\big[Y_{i,t=2}(\infty)\,|\,G_i=2\big]=\mathbb{E}\big[Y_{i,t=1}(\infty)\,|\,G_i=2\big]+\Big(\mathbb{E}\big[Y_{i,t=2}(\infty)\,|\,G_i=\infty\big]-\mathbb{E}\big[Y_{i,t=1}(\infty)\,|\,G_i=\infty\big]\Big)$$

3. Now, exploiting No-Anticipation and SUTVA:

$$\mathbb{E}\big[Y_{i,t=2}(\infty)\,|\,G_i=2\big]=\underbrace{\mathbb{E}\big[Y_{i,t=1}(2)\,|\,G_i=2\big]}_{\text{by No-Anticipation}}+\Big(\mathbb{E}\big[Y_{i,t=2}(\infty)\,|\,G_i=\infty\big]-\mathbb{E}\big[Y_{i,t=1}(\infty)\,|\,G_i=\infty\big]\Big)$$

$$\mathbb{E}\big[Y_{i,t=2}(\infty)\,|\,G_i=2\big]=\underbrace{\mathbb{E}[Y_{i,t=1}\,|\,G_i=2]}_{\text{by SUTVA}}+\Big(\mathbb{E}[Y_{i,t=2}\,|\,G_i=\infty]-\mathbb{E}[Y_{i,t=1}\,|\,G_i=\infty]\Big).$$

DiD Main Result

- Combining these results, we have that under SUTVA+No-Anticipation+PT assumptions:

$$\begin{aligned} \mathrm{ATT} &= \mathbb{E}[Y_{i,t=2} \mid G_i = 2] - (\mathbb{E}[Y_{i,t=1} \mid G_i = 2] + (\mathbb{E}[Y_{i,t=2} \mid G_i = \infty] - \mathbb{E}[Y_{i,t=1} \mid G_i = \infty])) \\ &= (\mathbb{E}[Y_{i,t=2} \mid G_i = 2] - \mathbb{E}[Y_{i,t=1} \mid G_i = 2]) - (\mathbb{E}[Y_{i,t=2} \mid G_i = \infty] - \mathbb{E}[Y_{i,t=1} \mid G_i = \infty]) \end{aligned}$$

- This is "the birth" of the DiD estimand!

Event Study Setup: Medicaid Expansion

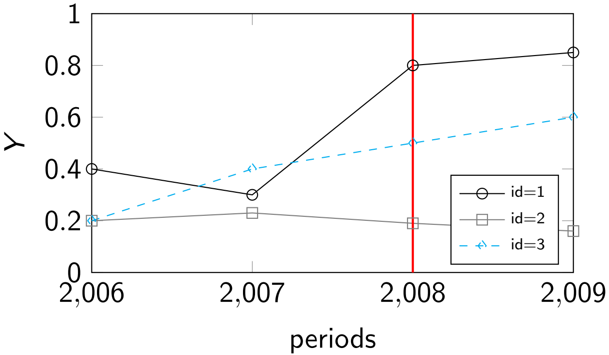
- Compare individuals where treatment took place vs control.
- For individual i at time t, and relative time with respect to treatment $r = (-q, +R)$:

$$Y_{it} = \gamma_s + \theta_t + \sum_{r=-q}^{R} \beta_r D_{st} + \epsilon_{it}$$

- We generate one dummy variable for each relative time period before and after treatment, and assign value 1 only when the unit is in that relative time (e.g. $+1$ in the year after the treatment (lags), -2 two years before the treatment (leads)).

EV: Example

id	time	treat	period	Y
1	2006	0	-2	0.4
1	2007	0	-1	0.3
1	2008	1	0	0.8
1	2009	1	1	0.85
2	2006	0	-2	0.2
2	2007	0	-1	0.23
2	2008	0	0	0.19
2	2009	0	1	0.16
3	2006	0	-2	0.2
3	2007	0	-1	0.4
3	2008	1	0	0.5
3	2009	1	1	0.6



TWFE

- Combining these results, we have that under SUTVA+No-Anticipation+PT assumptions:

$$\begin{aligned}\mathrm{ATT} &= \mathbb{E}[Y_{i,t=2} \mid G_i = 2] - (\mathbb{E}[Y_{i,t=1} \mid G_i = 2] + (\mathbb{E}[Y_{i,t=2} \mid G_i = \infty] - \mathbb{E}[Y_{i,t=1} \mid G_i = \infty])) \\ &= (\mathbb{E}[Y_{i,t=2} \mid G_i = 2] - \mathbb{E}[Y_{i,t=1} \mid G_i = 2]) - (\mathbb{E}[Y_{i,t=2} \mid G_i = \infty] - \mathbb{E}[Y_{i,t=1} \mid G_i = \infty])\end{aligned}$$

- This is “the birth” of the DiD estimand!

Two-Way Fixed Effects (TWFE)

- Standard Difference-in-Differences estimator when treatment timing is **uniform**:

$$Y_{it} = \alpha + \beta\, D_{it} + \delta_i + \gamma_t + \varepsilon_{it}$$

- δ_i absorbs **time-invariant heterogeneity** across units (states, firms, etc.).
- γ_t absorbs shocks **common to all units** in each period.
- β captures the *average treatment effect on the treated* (ATT) *under parallel trends*.

Assumptions for 2WFE

$$Y_{it} = \tau D_{it} + X'\beta + \alpha_i + \xi_t + \varepsilon_{it}$$

in which D_{it} is dichotomous

1. **Functional form**
   - Additive fixed effect: no interaction of ommited variables.
   - *Constant* (heterogeneity) and *contemporaneous* (no anticipation) treatment effect
   - Linearity in covariates
2. **Strict exogeneity**

$$\varepsilon_{it} \perp\!\!\!\perp D_{js}, X_{js}, \alpha_j, \xi_s \quad \forall i, j, t, s$$

$$\Rightarrow \quad \{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{js} \mid X, \alpha, \xi \quad \forall i, j, t, s$$

If only two groups, parallel trends implies:

$$\mathbb{E}[Y_{it}(0) - Y_{i't'}(0) \mid X] = \mathbb{E}[Y_{jt}(0) - Y_{j't'}(0) \mid X] \quad i \in \mathcal{T},\ j \in \mathcal{C},\ \forall t, t'$$

Fixed Effects & Linear Trends

- **Fixed Effects only (FE):**

$$Y_{it} = \alpha + \beta\, D_{it} + \delta_i + \gamma_t + \varepsilon_{it}$$

  - Controls for unobserved, time-invariant unit factors (δ_i).
  - Controls for period shocks common to all units (γ_t).

- **Adding Unit-Specific Linear Trends:**

$$Y_{it} = \alpha + \beta\, D_{it} + \delta_i + \gamma_t + \lambda_i\, t + \varepsilon_{it}$$

  - λ_i allows each unit to follow its own pre-treatment trajectory.
  - Helps when parallel trends hold *after detrending*.
  - Costs degrees of freedom and may absorb some treatment variation.

When to Include Trends?

Use diagnostic plots / pre-trend tests; include trends only if parallel-trends plausibly fails without them.

Event-Study (Leads/Lags) with Staggered Adoption in TWFE

Step 1: Define Event-Time Dummies

$$E_{it}^{(k)} = \mathbf{1}\{\text{unit } i \text{ is } k \text{ periods from first treatment at } t\}, \quad k = -K, \dots, -1, 0, 1, \dots, L$$

- $k < 0 \Rightarrow$ *lead* (placebo / pre-trend test)
- $k = 0 \Rightarrow$ *contemporaneous treatment*
- $k > 0 \Rightarrow$ *lag* (dynamic effect)
- Omit one event time (e.g. $k = -1$) as the reference period.

Step 2: TWFE Event-Study Specification

$$Y_{it} = \alpha + \sum_{k \neq -1} \beta_k \, E_{it}^{(k)} + \delta_i + \gamma_t + \varepsilon_{it}$$

- δ_i = unit fixed effects, γ_t = time fixed effects.
- Coefficients β_k trace the dynamic path relative to treatment.

What TWFE Assumptions Entail

Functional Form and Exogeneity

$$Y_{it} = \delta^{\text{TWFE}} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \varepsilon_{it}$$

$$D_{it} \perp\!\!\!\perp \varepsilon_{js} \,\big|\, X^{1:T}, \alpha, \xi^{1:T}, \quad \forall i, j, t, s$$

- **On treatment assignment**
  - Additive unobserved confounding
  - No “feedback”
- **On interference (SUTVA)**
  - No spatial spillover
  - No anticipation effects
  - No carryover effects
- **On HTE**
  - Constant treatment effect

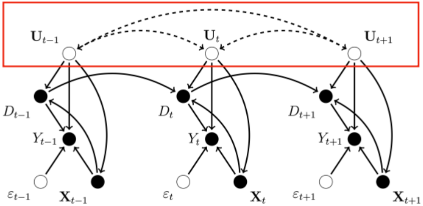
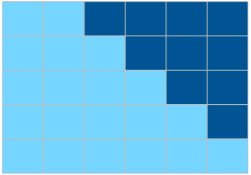


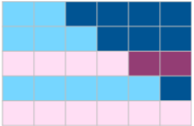
Figure: *

Illustration: Time-varying additive unobserved confounding (U_t)

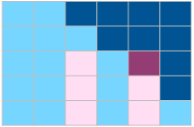
What TWFE Assumptions Entail



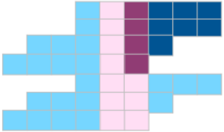
Original Data



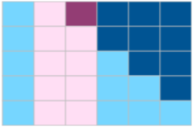
Interaction Weighted
& Stacked DID



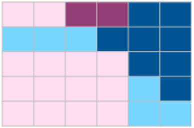
CSDID



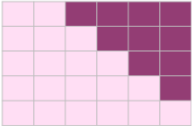
Stacked DID



DID multiple



PanelMatch

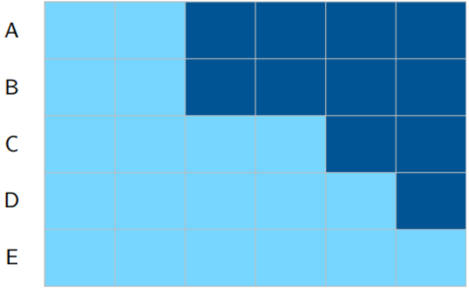


Imputation Method

Sun & Abraham (2021)

Interaction Weighted (IW)

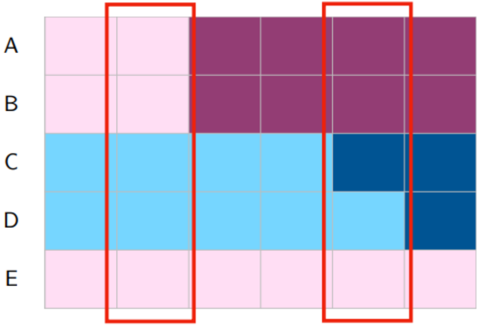
- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using 2×2 DID for each cohort g and period since treatment l
- ATT = average CATT, weighted by cohort size



Sun & Abraham (2021)

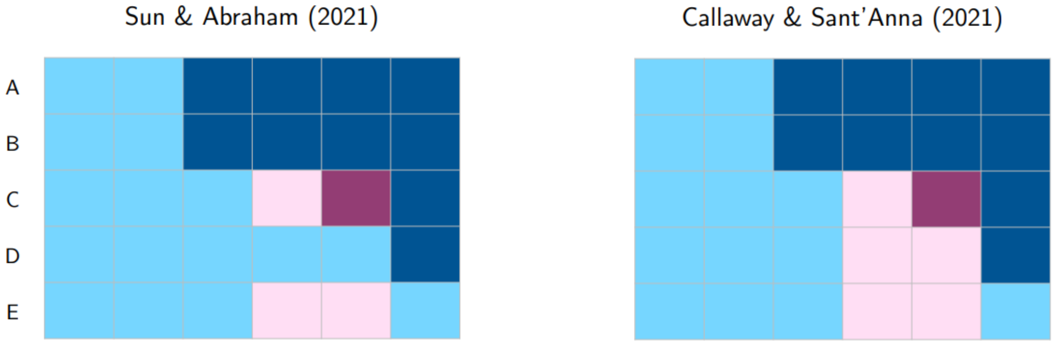
Interaction Weighted (IW)

- Comparison group: never-treated
- Estimate Cohort ATT (CATT) using 2×2 DID for each cohort g and period since treatment l
- ATT $=$ average CATT, weighted by cohort size



Callaway & Sant'Anna (2021)

- Comparison group: **not-yet-treated** (in addition to never treated)
- "Doubly robust" with covariates



Sun & Abraham (2021)

Callaway & Sant'Anna (2021)

Figure: Enter Caption

Motivation: Why a New DiD?

- Staggered treatment timing **breaks** the canonical 2×2 DiD set-up.
- Two–way fixed effects (TWFE) can assign negative weights and mask dynamic/heterogeneous effects.
- **Callaway & SantÁnna (2021) (CS)** propose a *divide–and–conquer* approach:
  1. Split the panel into many honest 2×2 comparisons.
  2. Estimate each group–time ATT under familiar DiD assumptions.
  3. Aggregate with user-chosen weights to answer specific policy questions.

Set-up and Notation

- Panel $\{Y_{it}, D_{it}\}$ for units $i = 1, \ldots, n$ and periods $t = 1, \ldots, T$.
- First treatment time $G_i \in \{2, \ldots, T, \infty\}$ creates **cohorts**. Let $G_g = \mathbf{1}\{G_i = g\}$.
- Potential outcomes $Y_{it}(g)$: outcome at t if first treated in g; $Y_{it}(\infty)$ if never treated.
- **Building block**: cohort-time average treatment effect

$$\mathrm{ATT}(g, t) = \mathbb{E}\big[Y_t(g) - Y_t(\infty) \mid G_g = 1\big], \quad t \geq g.$$

Identification Assumptions

No Anticipation

$Y_{it}(g) = Y_{it}(g') \; \forall \, t < \min\{g, g'\}$. Treatment cannot influence pre-treatment outcomes.

Conditional Parallel Trends

For $t \geq g$,

$$\mathbb{E}\big[\Delta Y_t(\infty) \mid G_g = 1, X\big] = \mathbb{E}\big[\Delta Y_t(\infty) \mid C, X\big],$$

where C denotes either

- *Never-treated* units, or
- *Not-yet-treated* units ($D_{st} = 0$).

Identification via Long Differences

With the assumptions:

$$\mathrm{ATT}(g,t) = \underbrace{\mathbb{E}[\Delta_{g-1,t}Y \mid G_g = 1]}_{\text{treated}} - \underbrace{\mathbb{E}[\Delta_{g-1,t}Y \mid C]}_{\text{control}}, \quad \Delta_{g-1,t}Y = Y_t - Y_{g-1}.$$

- Choice of control group determines estimator (never vs. not-yet).
- Can incorporate covariates via

  IPW inverse-probability weighting with group-specific propensity scores $p_g(X)$.

  REG outcome regression for controls $m_{g,t}(X)$.

  DR doubly robust combination of both (preferred).

Stacked DID: Cengiz et al. (2019)

- Duplicate the pure control group for each cohort
- “Stack” on top of each other, align by relative time to treatment onset
- Run saturated regression
- Similar to IW with disproportionate weights

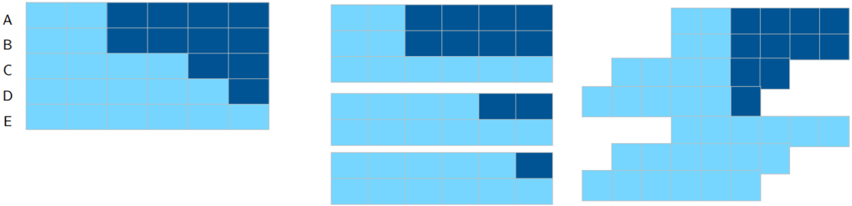
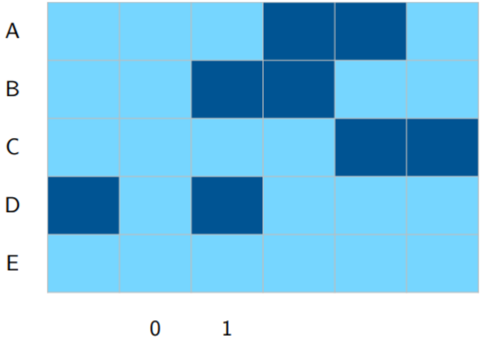


Figure: Enter Caption

De Chaisemartin and D'Haultfoeuille (2020)

- No cohorts — estimates a single average effect
- Effect for switchers (not ATT)
- Match treated to control with shared treatment status in previous period
  - Switchers $(i, t) : D_{it} \neq D_{it-1}$
  - Stable group $(i, t) : D_{it} = D_{it-1}$



The End