Week 7:
Difference-in-Differences

Dr Christian Engels
ce50@st-andrews.ac.uk

FI5699 Dissertation Module

Department of Finance
University of St Andrews Business School

Introduction

What We Will Cover Today

This session covers the difference-in-differences revolution: why standard TWFE fails with staggered treatment timing, and how modern estimators fix it.

1. **DiD Refresher** — the canonical estimator, graphical intuition, and parallel trends
2. **What's Wrong with TWFE?** — Goodman-Bacon decomposition and negative weights
3. **The Forward-Engineering Approach** — define targets before choosing estimators
4. **Modern DiD Estimators** — CS, SA, BJS, stacked, DR-DiD
5. **Pre-trends and Sensitivity Analysis** — Roth (2022), Rambachan & Roth (2023)
6. **Python Implementation** — the `diff-diff` package
7. **Empirical Lessons** — what happens when you re-estimate published results
8. **Practical Recommendations** — decision tree and checklist for your dissertation

By the end of today, you will be able to diagnose TWFE bias, estimate robust DiD models, and run sensitivity analysis in Python.

Recap: What You Learned in Week 6

Concept	What you learned
Instrumental variables	IV solves time-varying confounders; requires relevance, independence, exclusion
2SLS estimation	Always check first-stage $F > 10$; IV identifies the LATE for compliers
Classical event studies	Market model $\rightarrow$ abnormal returns $\rightarrow$ CAR; use daily data, short windows
`pyfixest` IV syntax	`"y ~ x | fe | endo ~ z"`

Today we tackle staggered policy adoption. The standard two-way fixed effects regression breaks down — and we learn why and how to fix it.

55% of DiD Papers in Top Finance Journals Use Staggered Designs

Baker, Larcker & Wang (2022) audit 744 DiD papers in top-5 finance and accounting journals (2000–2019).

The problem: standard TWFE regressions are biased in most of these settings.

This week covers the diagnosis, the cure, and how to implement it in Python.

DiD Refresher

DiD Exploits Two Sources of Variation to Identify Causal Effects

DiD methods exploit variation in **time** (before vs. after) AND across **groups** (treated vs. untreated).

DiD combines two imperfect approaches to avoid their pitfalls:

1. **Pre–post comparison alone fails** because it ignores time trends that affect the outcome regardless of treatment.
2. **Cross-sectional comparison alone fails** because treated and untreated groups may differ systematically at baseline.

DiD differences out *both* problems: it removes baseline group differences *and* common time trends in a single step.

Not magic: we must assume that, absent treatment, the outcome would have evolved similarly across groups — the Parallel Trends assumption.

The Canonical DiD Estimator

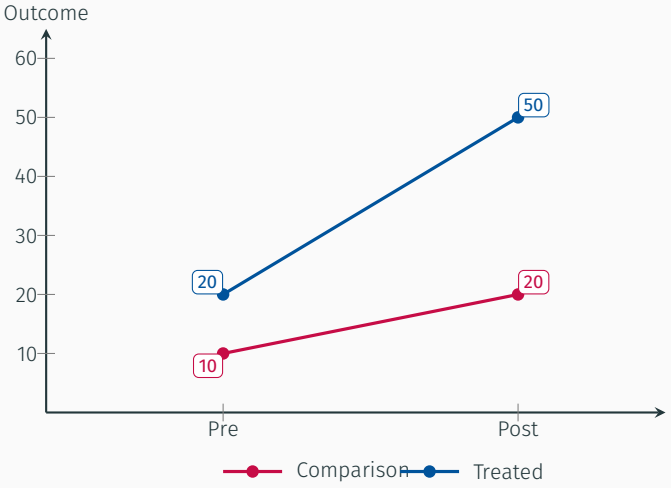
The difference-in-differences estimator is:

$$\hat{\theta}^{\text{DiD}} = \left(\bar{Y}_{\text{Treated,Post}} - \bar{Y}_{\text{Treated,Pre}}\right) - \left(\bar{Y}_{\text{Untreated,Post}} - \bar{Y}_{\text{Untreated,Pre}}\right)$$

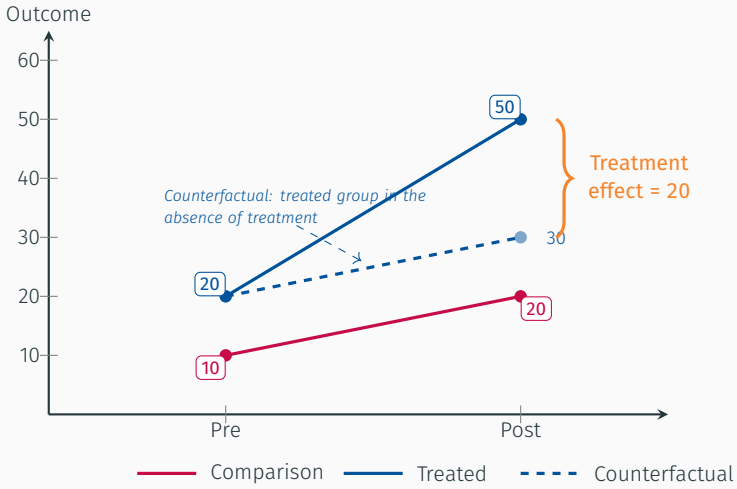
Component	Interpretation
First difference	How did the **treated** group change over time?
Second difference	How did the **untreated** group change over time?
The DiD	Subtract the second from the first to remove common trends

The second difference removes any time trend that affects both groups equally. What remains is the treatment effect — if parallel trends holds.

DiD via Graphs: The Raw Data



DiD via Graphs: The Counterfactual



The Classic Example: Card and Krueger (1994)

- **Question:** Does increasing the minimum wage reduce employment?
- **Setting:** New Jersey raised its minimum wage; neighbouring Pennsylvania did not.
- **Treated:** Fast-food restaurants in New Jersey (before and after the increase).
- **Control:** Fast-food restaurants in Pennsylvania (same time period).
- **Finding:** No significant reduction in employment — challenged the conventional wisdom.

The Pennsylvania restaurants control for any economy-wide trends affecting fast food. The DiD strips out these common shocks, isolating the minimum wage effect.

From Simple DiD to the Modern Problem

The basic 2×2 DiD (one treatment group, one control group, one time period) is clean and interpretable.

But most real applications have staggered treatment timing: units are treated at different times.

Standard practice: run a TWFE regression and interpret $\hat{\beta}$ as the ATT:

$$y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it}$$

- This works perfectly with a **single treatment date**.
- But with staggered timing and heterogeneous effects, $\hat{\beta}$ is **not** what you think.

The rest of this lecture shows why TWFE breaks down with staggered timing, and what to do about it.

What's Wrong with TWFE?

TWFE Is a Weighted Average of All Pairwise 2 × 2 DiDs

Goodman-Bacon (2021, *J. Econometrics*)

The DD Decomposition Theorem:

$$\hat{\beta}^{DD} = \underbrace{\sum_{k \neq U} s_{kU} \, \hat{\beta}^{2\times2}_{kU}}_{\text{Treated vs. untreated}} + \underbrace{\sum_{k} \sum_{\ell > k} \left[s^{k}_{k\ell} \, \hat{\beta}^{2\times2,k}_{k\ell} + s^{\ell}_{k\ell} \, \hat{\beta}^{2\times2,\ell}_{k\ell} \right]}_{\text{Timing comparisons (problematic)}}$$

Weights depend on:

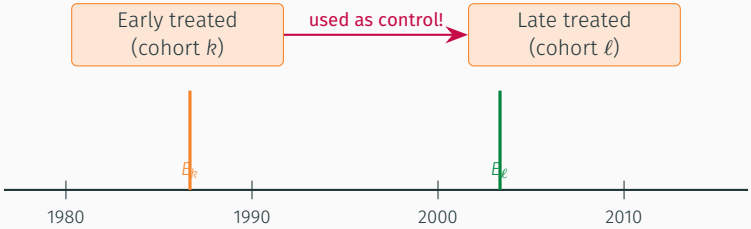
- Squared subsample size
- Variance of treatment dummy
- Units treated mid-panel get most weight

Estimand:

$$\text{plim } \hat{\beta}^{DD} = \underbrace{\text{VWATT}}_{\text{target}} + \underbrace{\text{VWCT}}_{\substack{=0 \text{ under} \\ \text{par. trends}}} - \underbrace{\Delta\text{ATT}}_{\substack{\text{bias from} \\ \text{dyn. eff.}}}$$

The Bias Comes from Using Already-Treated Units as Controls

Goodman-Bacon (2021)



Between E_k and E_ℓ, cohort k's **changing treatment effects** contaminate the control trend.

If effects grow over time $\Rightarrow \Delta\text{ATT} \neq 0 \Rightarrow$ TWFE is biased.

Staggered + Static Effects: TWFE Still OK (Simulations 1–3)

Baker, Larcker & Wang (2022, *JFE*) — Fig. 1 — Problems emerge only when timing is staggered *and* effects are dynamic

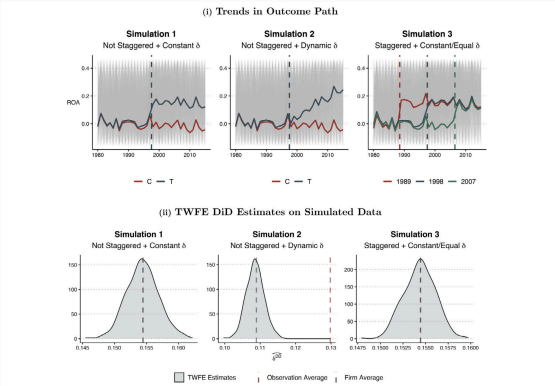


Fig. 1. Simulation: TWFE DiD Estimates Under Uniform Treatment Timing or Treatment Effect Homogeneneity.
Figure 1, panel (i), plots the firm-level outcome path (the gray lines) and the average outcome path by treatment groups (the bold lines) in one of the simulated Compustat datasets for Simulations 1, 2, and 3. To construct a simulated panel dataset, for each year, firm, and observation in the sample, we draw year-fixed effects, firm-fixed effects, and ROA residuals, respectively, from the empirical distribution. We then randomly draw states of incorporation for each firm and randomly assign states into treatment (T) and control groups (C) (i.e., in Simulations 1 and 2) or different treatment timing groups (i.e., in Simulation 3). Finally, we introduce treatment effects to the firms incorporated in treated states. Simulation 1 introduces a single treatment with a static effect (Eq. (5)). Simulation 2 introduces a single treatment with a dynamic effect (Eq. (6)). Simulation 3 introduces three treatments—to firms assigned to the 1989, 1998, or 2007 treatment-timing groups—each with static effects of the same magnitude (Eq. (7)). Panel (ii) plots the distribution of the static TWFE DiD treatment effect estimate ($\hat{\delta}^{DD}$ from Eq. (2)) from 500 Monte Carlo simulations of our three different data generating processes. The curve represents the distribution of the TWFE estimates, while the dashed vertical lines represent the observation-level or firm-level average ATT.

Staggered + Dynamic Effects: TWFE Fails (Simulations 4–6)

Baker, Larcker & Wang (2022, *JFE*) — Fig. 2 — Sim. 6: every cohort has positive ATT, yet TWFE is below zero

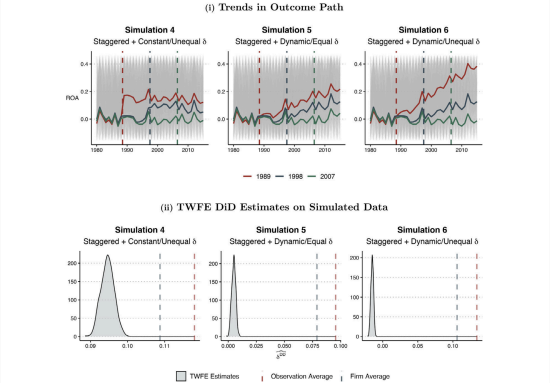


Fig. 2. Simulation: TWFE DiD Estimates Under Uniform Treatment Timing or Treatment Effect Homogeneneity.
Figure 2, panel (i), plots the firm-level outcome path (the gray lines) and the average outcome path by treatment groups (the bold lines) in one of the simulated Compustat datasets for Simulations 4, 5, and 6. To construct a simulated panel dataset, for each year, firm, and observation in the sample, we draw year-fixed effects, firm-fixed effects, and ROA residuals, respectively, from the empirical distribution. We then randomly draw states of incorporation for each firm and randomly assign states into different treatment timing groups: 1989, 1998, or 2007. Finally, we introduce treatment effects to the firms incorporated in treated states. Simulation 4 introduces static treatment effects, where the effect magnitudes differ across treatment-timing groups (Eq. (8)). Simulation 5 introduces dynamic treatment effects, where the dynamics are the same across treatment-timing groups (Eq. (9)). Simulation 6 introduces dynamic treatment effects, where the dynamics differ across treatment-timing groups (Eq. (10)). Panel (ii) plots the distribution of the static TWFE DiD treatment effect estimate ($\hat{\delta}^{DD}$ from Eq. (2)) from 500 Monte Carlo simulations of the three different data generating processes. The curve represents the distribution of the TWFE estimates, while the dashed vertical lines represent the observation-level or firm-level average ATT.

TWFE Event Studies Fabricate Pre-trends That Do Not Exist

Baker, Larcker & Wang (2022, *JFE*) — Fig. 5 — No real pre-trends exist, yet TWFE reports them

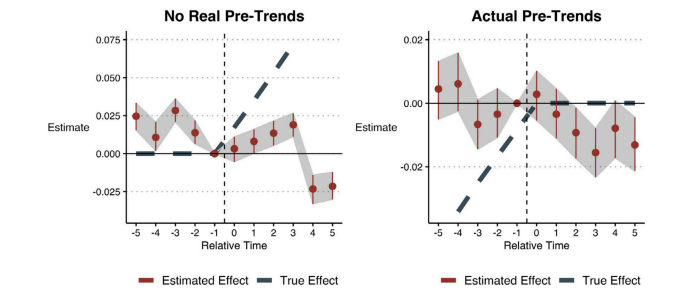


Fig. 5. Simulation: TWFE Dynamic Treatment Effect Estimates with No Real Pre-Trends and Actual Pre-Trends.
Figure 5, left-hand panel, plots the distribution of event-study estimates based on a variant of Simulation 6 (Eq. (10)) in which there are no pre-period trends and post-treatment trend-breaks of the three different cohorts are $\delta_{1989} = 0.10\sigma_{ROA}$, $\delta_{1998} = 0.05\sigma_{ROA}$, and $\delta_{1998} = 0.01\sigma_{ROA}$, where σ_{ROA} is the empirical *ROA* standard deviation in Compustat. For each of the 500 simulated Compustat panel datasets, we estimate a TWFE event-study specification (Eq. (12)) that includes relative-time indicators for the five years before and after the year of treatment (Relative Time = 0). We exclude the relative-time indicator for the year prior to treatment (Relative Time = -1). Moreover, we combine relative-time periods more than five years before treatment into one bin and relative-time periods more than five years after treatment into another bin. For each relative-time period from -5 to 5, we plot the point estimate (the solid circle), the 95% confidence interval (the vertical lines intersecting the solid circles), and the observation-average ("true") ATT for each relative-time period (the dashed line). The right-hand panel plots the distribution of event-study estimates based on Simulation 7 (Eq. (15)), which has pre-treatment trends and but no treatment effects. Aside from the data generating function, all other aspects of this simulation are the same as the left-hand-side panel.

de Chaisemartin & D'Haultfœuille: TWFE Weights Can Be Negative

de Chaisemartin & D'Haultfœuille (2020, *AER*)

Under heterogeneous effects, $\hat{\beta}^{fe} = E\left[\sum_{(g,t):\, D_{g,t}=1} W_{g,t} \cdot \Delta_{g,t}\right]$ where $\sum W_{g,t} = 1$.

The key result: Some weights $W_{g,t}$ can be **strictly negative**. Even if every $\Delta_{g,t} > 0$, TWFE can estimate $\hat{\beta}^{fe} < 0$.

Diagnostic: compute the weights — if many are negative, TWFE is unreliable.
Remedy: DID_M uses only "switchers" and clean controls.

Stata	`twowayfeweights, did_multiplegt`
R	`DIDmultiplegt`
Python	`diff_diff.TwoWayFixedEffects`

The Forward-Engineering Approach

All DiD Designs Reduce to 2 × 2 Building Blocks

Baker et al. (2025, *J. Econ. Lit.*)

The central insight:

Any DiD study — however complex — decomposes into clean 2 × 2 comparisons between units whose treatment *changes* and units whose treatment does *not*.

	Comparison	Treated
Pre	$\bar{Y}_{C,\text{pre}}$	$\bar{Y}_{T,\text{pre}}$
Post	$\bar{Y}_{C,\text{post}}$	$\bar{Y}_{T,\text{post}}$
Δ	$\hat{\delta}_C$	$\hat{\delta}_T$

$$\hat{\tau} = \hat{\delta}_T - \hat{\delta}_C$$

Estimate each building block separately, then aggregate.

Weighting Choices Change the Answer, Not Just the Precision

Baker et al. (2025) — ACA Medicaid expansion example

2,604 counties $\times$ 2009–2019. Treatment staggered: 2014, 2015, 2016, 2019 cohorts.

The same design, two answers:

Unweighted 2×2 DiD: $+0.1$ deaths/100k (each county counts equally)

Weighted 2×2 DiD: -2.6 deaths/100k (each adult counts equally)

Different parameters, not different estimates. The weighting scheme defines the causal quantity you are estimating.

Note: These are simple 2×2 DiDs for the 2014 cohort only. With staggered timing, the TWFE and CS estimates will differ — we reconcile all three numbers in the Python section.

Define Your Target Before Choosing Your Estimator

Baker et al. (2025) — the 8-step framework

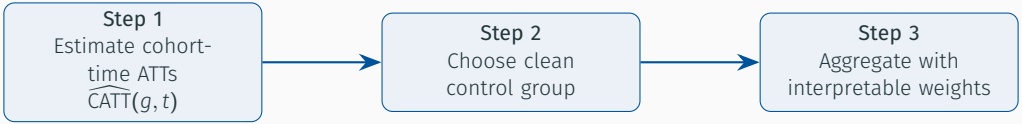
1. **Define** target parameters in potential outcomes notation
2. **State** identification assumptions (PT, NA, overlap)
3. **Choose** estimation method (RA, IPW, DR, or regression)
4. **Discuss** inference framework (sampling vs. design-based)
5. **Estimate**
6. **Sensitivity analysis** (Rambachan-Roth, functional form)
7. **Heterogeneity analysis** (sub-group ATTs)
8. **Keep learning** — if DiD assumptions look implausible, consider alternative designs

Old approach: run TWFE, reverse-engineer its interpretation.
New approach: define the parameter you want, then derive the estimator that delivers it.

Modern DiD Estimators

The Core Idea: Estimate Clean 2×2 DiDs, Then Aggregate

All modern estimators share a common logic:



Callaway & Sant'Anna: The Most Flexible Framework

Callaway & Sant'Anna (2021, *J. Econometrics*)

Group-time ATT: For cohort g at time t:

$$\text{ATT}(g,t) = E[Y_t(g) - Y_t(0) \mid G_g = 1]$$

Estimated via clean 2×2 DiDs under **conditional parallel trends** + **no anticipation**.

Three design choices make CS the most flexible framework:

1. **Control group:** not-yet-treated (preferred), never-treated, or last-treated
2. **Estimation:** outcome regression, IPW, or doubly robust
3. **Aggregation:** by event time, cohort, calendar time, or overall

CS Provides Simultaneous Inference and Multiple Aggregation Schemes

Aggregation options:

- By relative time (event study)
- By cohort (group-specific effects)
- Overall (single summary)
- Calendar time

Inference: simultaneous confidence bands, not just pointwise — accounts for multiple testing across event-time estimates.

R	`did` package
Python	`diff_diff.CallawaySantAnna`

Default: doubly robust estimation.

Switch `control_group` between `'not_yet_treated'` and `'never_treated'` as a robustness check.

Sun & Abraham: Interaction-Weighted Estimator

Sun & Abraham (2021, *J. Econometrics*)

Step 1: Saturated interacted regression: $y_{it} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{\ell \neq -1} \delta_{e,\ell} \left(\mathbb{1}\{E_i = e\} \cdot D_{it}^{\ell} \right) + \varepsilon_{it}$
— each $\hat{\delta}_{e,\ell}$ consistently estimates CATT(e, ℓ).

Step 2: Aggregate: $\hat{\nu}_{\ell} = \sum_e \hat{w}_{e,\ell} \cdot \hat{\delta}_{e,\ell}$, with $\hat{w}_{e,\ell} \geq 0$. Weights are **convex**.

Advantage: Single regression, fast, direct SEs.
Limitation: No covariate adjustment.
R: `fixest::sunab()`. Python: `diff_diff.SunAbraham`.

Key difference from TWFE:
TWFE: non-convex, contaminated.
SA: convex weights, clean.

Borusyak, Jaravel & Spiess: Imputation Approach

Borusyak et al. (2024, *ReStud*)

Three steps:

1. Estimate $\hat{\alpha}_i, \hat{\lambda}_t$ from untreated obs only
2. Impute $\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\lambda}_t$ for treated
3. $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$, aggregate

Strengths:

- Most efficient among heterogeneity-robust estimators
- Handles covariates naturally
- Equivalent to OLS on untreated data

R: `didimputation`

Python: `diff_diff.EfficientDiD`

Stacked Regression: The Pragmatist's Solution

1. For each cohort, build a clean 2×2 dataset with the cohort and its clean controls
2. Add a dataset identifier to each sub-experiment
3. **Stack** all datasets together
4. Run TWFE with dataset-specific unit and time FE: α_{ig} and λ_{tg}

Pros:

- Familiar OLS workflow
- Efficient (variance-weighted)
- Handles covariates and clustering

Cons:

- May be inconsistent for sample-average ATT
- Less flexible aggregation
- Duplicate observations across stacks

Used in: Cengiz et al. (2019, *QJE*); Deshpande & Li (2019, *AEJ:EP*).

Doubly Robust DiD: Insurance Against Misspecification

Sant'Anna & Zhao (2020, *J. Econometrics*)

When parallel trends holds only **conditionally on covariates**, model two nuisance functions:

Outcome regression

Model $m_{0,t}(X)$

Fails if OR misspecified

IPW (Abadie 2005)

Model $p(X)$

Fails if PS misspecified

Doubly robust

Model **both**

Consistent if either correct

Default estimator in `did` R package and in `diff_diff.CallawaySantAnna(doubly_robust=True)` in Python.

Each Estimator Makes Different Choices About Controls and Covariates

	Controls	Covariates	Inference
Callaway & Sant'Anna	NYT / NT / LT	Yes (DR)	Simultaneous CI
Sun & Abraham	NT / LT	No	Pointwise
Borusyak et al.	NYT / NT	Yes	Pointwise
Stacked regression	NYT / NT	Yes	Cluster-robust

NYT = not-yet-treated, NT = never-treated, LT = last-treated.

Alternative Estimators Recover the True ATT; TWFE Does Not

Baker, Larcker & Wang (2022, *JFE*) — Fig. 6 — CS, Stacked, SA concentrate on true ATT; TWFE is entirely wrong in Sim. 6

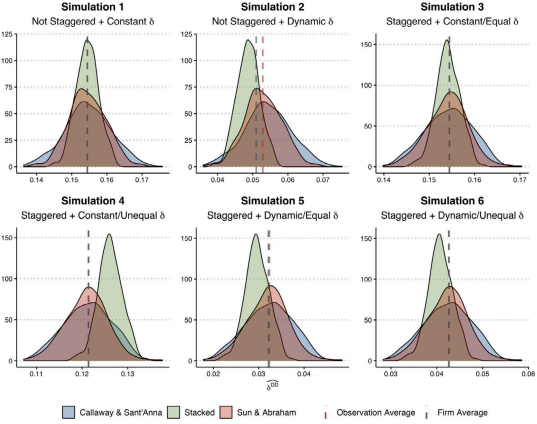


Fig. 6. Simulation: Distribution of Static Effect Estimates of Alternative Estimators.
Figure 6 plots the distribution of static treatment effect estimates for the three alternative estimators explained in Section 4. These distributions are generated based on applying the alternative estimators to each of the 500 simulated Compustat ROA panel datasets under Simulations 1–6. For each data generating process, we overlay the three distributions. The dashed vertical lines represent the observation-level or firm-level average ATT for the five-year period post treatment.

Robust Event Studies Trace the True Dynamic Path

Baker, Larcker & Wang (2022, *JFE*) — Fig. 7 — Pre-treatment estimates are zero, as they should be

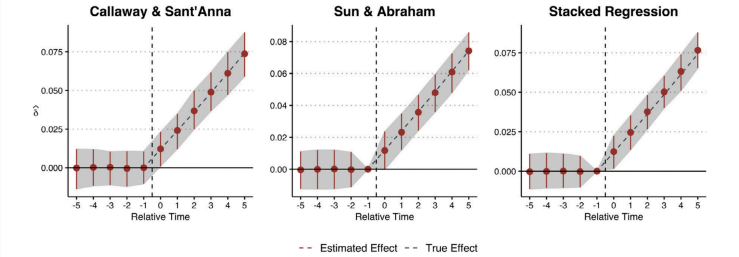


Fig. 7. Robust DiD Methods with Staggered Treatment Assignment and Dynamic Treatment Effects.
Figure 7 plots the distribution of treatment effect estimates by relative-time period for the three alternative estimators explained in Section 4. These distributions are generated based on applying the alternative estimators to each of the 500 simulated Compustat ROA panel datasets under Simulation 6 (Eq. (10)), for which TWFE DiD estimates are highly biased. For each relative-time period from −5 to 5, we plot the point estimate (the solid circle), the 95% confidence interval (the vertical lines intersecting the solid circles), and the observation-level average ATT for each relative-time period (the dashed blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Pre-trends and Sensitivity Analysis

Passing the Pre-trends Test Does Not Validate Parallel Trends

Roth (2022, *AER: Insights*)

Standard practice: test for pre-trends → if insignificant, proceed with DiD.

The problem: This is a **pre-test**, and pre-tests distort subsequent inference.

When the pre-trends test passes:

- Does **not** mean parallel trends holds
- Conditional on passing, treatment effects are biased towards significance
- Confidence intervals have poor coverage

Low power is the culprit:

- Pre-trends tests have low power against economically meaningful violations
- Violations too small to detect can still produce large post-treatment bias
- Median bias of 40%+ in simulations

Recommendation: Do **not** rely on the pre-trends test alone. Use sensitivity analysis.

Honest Confidence Intervals Bound the Violation Rather Than Test for Zero

Rambachan & Roth (2023, *ReStud*)

Key insight: Instead of testing whether pre-trends are zero, *bound how different* post-treatment trends could be from pre-treatment trends.

ΔRM($\bar{M}$): Relative magnitudes

Post-treatment violations $\leq \bar{M} \times$ max pre-treatment violation.

"Violations don't get worse after treatment."

ΔSD(M): Smoothness

Consecutive differences in violations bounded by M.

"Violations change smoothly over time."

Construct **honest CIs** valid under all violations in the chosen class.

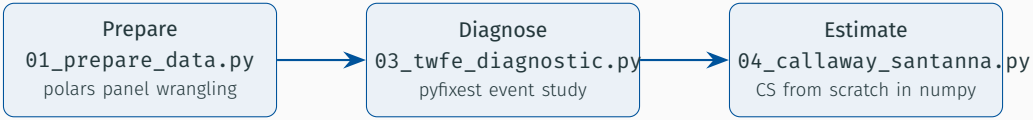
ACA Medicaid example: robust CI for $\widehat{\text{ATT}}(2014) = [-11.1, 5.1]$. Wide, but honest.

R: `HonestDiD`. Python: `diff_diff.HonestDiD`.

Python Implementation

Our Python Pipeline: `pyfixest`

The scripts below walk through each stage of the analysis. We use `uv` for environment management.



All code is in `week-7-code/`. Run `python main.py` to reproduce every figure and table.

For your dissertation, consider the `diff-diff` package (`uv pip install diff-diff`), which wraps CS, SA, BJS, and HonestDiD in a scikit-learn API.

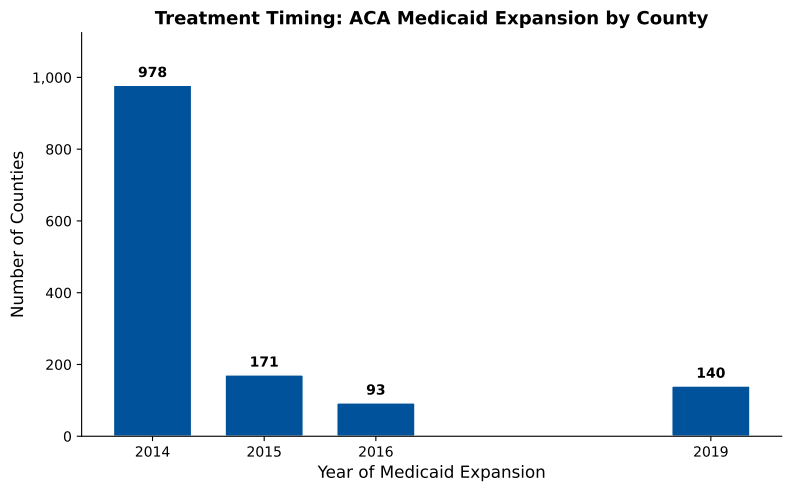
Loading and Preparing the Medicaid Panel

```python
import polars as pl
raw = pl.read_csv("county_mortality_data.csv",
                  null_values=["NA", ""])

# Drop pre-2014 adopters (DE, MA, NY, VT) and DC
df = raw.filter(
    ~pl.col("state").is_in(["DC","DE","MA","NY","VT"]))

# Treatment timing; 0 for never-treated
df = df.with_columns(
    pl.col("yaca").cast(pl.Int64, strict=False)
      .fill_null(0).alias("treat_year"))

# Population weight = county pop. in 2013
wt = df.filter(pl.col("year") == 2013).select(
    "county_code",
    pl.col("population_20_64").alias("set_wt"))
df = df.join(wt, on="county_code")
```

Key columns: `county_code` (unit), `year` (time), `crude_rate_20_64` (outcome), `treat_year` (expansion year; 0 = never), `set_wt` (2013 population weight).

Running Example: ACA Medicaid Expansion and County Mortality

Baker, Callaway, Cunningham, Goodman-Bacon & Sant'Anna (2025, *JEL*) — 2,702 counties × 11 years



Step 1: TWFE Diagnostic with `pyfixest`

```python
import pyfixest as pf

pdf = df.to_pandas()

# Static TWFE: population-weighted, clustered SEs
twfe = pf.feols(
    "crude_rate_20_64 ~ treated | county_code + year",
    data=pdf,
    weights="set_wt",          # 2013 population
    vcov={"CRV1": "county_code"})

# Event study with population weights
twfe_es = pf.feols(
    "crude_rate_20_64 ~ i(time_to_treat, Treat, ref=-1) "
    "| county_code + year",
    data=pdf,
    weights="set_wt",
    vcov={"CRV1": "county_code"})
```

Population weights mean each **adult** counts equally, not each county.

TWFE Static Result: −0.94 (Insignificant)

```
TWFE Static Estimate (population-weighted, clustered SEs)
Coefficient: -0.9352
SE:           2.1766
t-stat:      -0.43
```

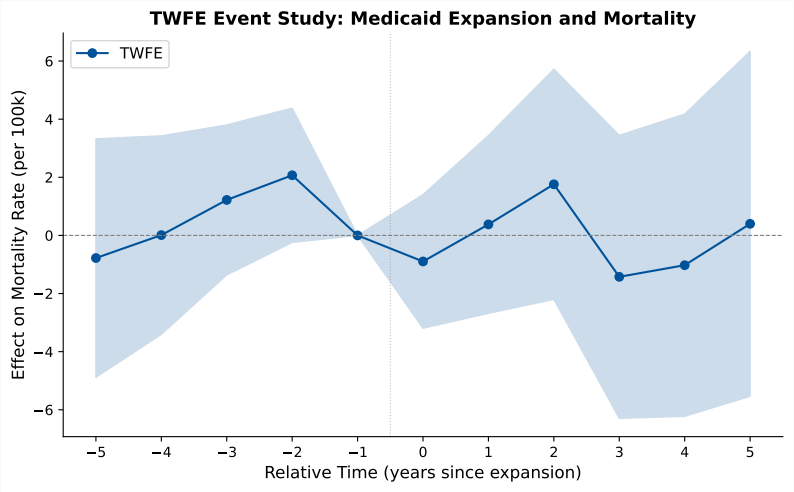
What this tells us:

- TWFE estimates a **negative** but **insignificant** effect of Medicaid expansion on mortality
- Is this the true ATT — or contaminated by staggered timing?

This is the TWFE estimate — a weighted average of all pairwise 2×2 DiDs, including contaminated ones. We need CS to get a clean answer.

TWFE Event Study on Medicaid Data: What Does It Show?

ACA Medicaid expansion — TWFE event study from `pyfixest`, relative to $e = -1$



Step 2: Callaway-Sant'Anna — Clean 2 × 2 DiDs

```
for g in cohorts:              # loop over treatment cohorts
    base_year = g - 1          # universal base period

    for t in years:
        e = t - g              # relative time

        # Not-yet-treated controls
        nyt_mask = (cohort_arr == 0) | (cohort_arr > t)

        # Population-weighted 2x2 DiD for this (g, t)
        att = (wavg(Y[treated, t], W[treated])
             - wavg(Y[treated, base], W[treated])) \
            - (wavg(Y[nyt, t], W[nyt])
             - wavg(Y[nyt, base], W[nyt]))

# Aggregate by relative time, weight by cohort size
# Bootstrap CIs: resample counties 1,000 times
```

Each cohort g gets a clean 2 × 2 DiD against not-yet-treated units. Population weights ensure each adult counts equally (matching Baker et al., 2025). CIs via cluster bootstrap.

CS Results: Pre-trends Near Zero, Overall ATT ≈ 0

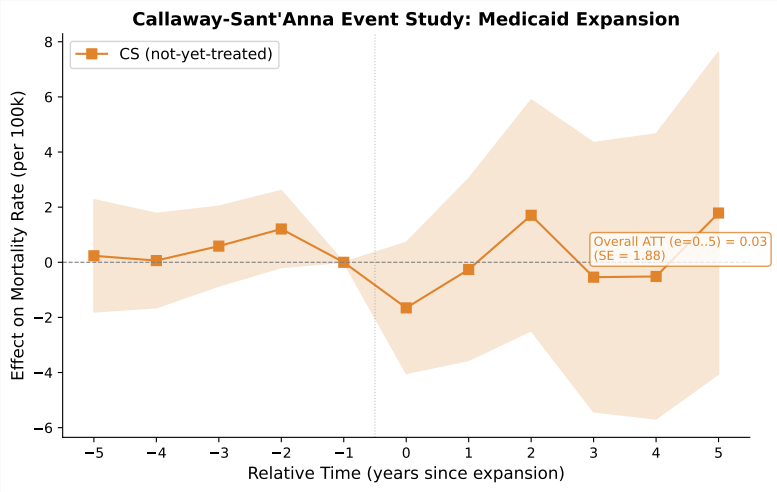
```
Callaway-Sant'Anna (not-yet-treated, pop. weighted)
Overall ATT (e in [0,5]): +0.03 (SE = 1.88)
95% CI: [-3.65, +3.70]

  e = -5: ATT = +0.23  [-1.80, +2.27]
  e = -4: ATT = +0.06  [-1.65, +1.77]
  e = -3: ATT = +0.59  [-0.86, +2.03]
  e = -2: ATT = +1.21  [-0.19, +2.61]
  e = -1: ATT = +0.00   (reference period)
  e = +0: ATT = -1.65  [-4.03, +0.72]
  e = +1: ATT = -0.26  [-3.56, +3.04]
  e = +2: ATT = +1.71  [-2.48, +5.89]
  e = +3: ATT = -0.54  [-5.43, +4.35]
  e = +4: ATT = -0.51  [-5.69, +4.66]
  e = +5: ATT = +1.79  [-4.05, +7.63]
```

CS aggregates clean 2×2 DiDs across **all four cohorts** (2014, 2015, 2016, 2019) over event times $e = 0$ to 5.

Pre-trends near zero; overall ATT ≈ 0 with wide CIs.

CS Event Study: Clean Estimates Without TWFE Contamination

Callaway-Sant'Anna with not-yet-treated controls — cluster bootstrap 95% CIs



Three Estimates, Three Different Questions

Baker et al. (2025) report multiple estimates from the **same data**. Each answers a different question:

Method	What it estimates	Estimate
Simple 2×2 DiD	2014 cohort only, immediate effect (2013 vs. 2014), pop. weighted	-2.6
TWFE static	Contaminated weighted average across all cohorts and all periods	-0.94
Callaway-Sant'Anna	Clean average across **all cohorts**, event times $e = 0$ to 5, pop. weighted	$+0.03$

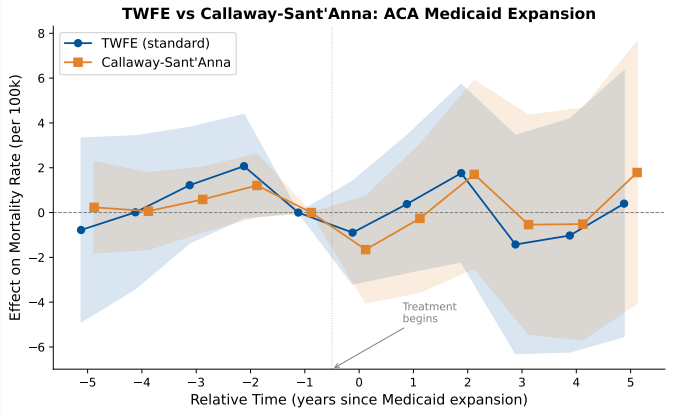
The 2×2 DiD (-2.6) looks at only the 2014 cohort's immediate effect. CS ($+0.03$) aggregates all four cohorts (2014–2019) over six post-treatment years — the positive effects of later cohorts cancel the negative 2014 effect. **Different estimands, not conflicting results.**

Step 3: Overlay TWFE and CS to See the Divergence

```python
import numpy as np
import matplotlib.pyplot as plt

twfe = np.load("outputs/twfe_event_study_data.npz")
cs = np.load("outputs/cs_event_study_data.npz")

fig, ax = plt.subplots(figsize=(9, 5.5))

# TWFE with CI band
ax.fill_between(twfe_t - 0.12, twfe_lo, twfe_hi,
                alpha=0.15, color="#00539B")
ax.plot(twfe_t - 0.12, twfe_e, "o-",
        label="TWFE (standard)")

# CS with CI band
ax.fill_between(cs_t + 0.12, cs_lo, cs_hi,
                alpha=0.15, color="#E0832B")
ax.plot(cs_t + 0.12, cs_e, "s-",
        label="Callaway-Sant'Anna")
```

TWFE and CS Agree Here — and That Is the Diagnostic

71% of treated counties expanded in 2014 — limited staggering means limited TWFE contamination



Agreement = good news. When TWFE and CS diverge, TWFE is contaminated (Baker et al., 2022 simulations). When they agree, TWFE is defensible — but you only know by running both.

For Your Dissertation: The `diff-diff` Package

For your own research, use a production package rather than writing estimators by hand.

`uv pip install diff-diff` · `diff-diff.readthedocs.io/en/stable/`

```python
from diff_diff import CallawaySantAnna, HonestDiD

cs = CallawaySantAnna(
    control_group='not_yet_treated',
    doubly_robust=True)
cs.fit(data, outcome='roa', group='cohort',
       time='year', unit_id='firm_id',
       covariates=['size', 'leverage', 'age'])
cs.print_summary()

# Sensitivity analysis: Rambachan & Roth (2023)
honest = HonestDiD(cs.results_)
honest.sensitivity_analysis(
    restriction='relative_magnitudes', m_bar=1.0)
honest.plot_sensitivity()
```

Also available: `SunAbraham`, `EfficientDiD`, `SyntheticDiD`.

Running Sensitivity Analysis Step by Step

```python
# Rambachan-Roth sensitivity: how robust is your result?
honest = HonestDiD(cs.results_)

# Test at M-bar = 0.5, 1.0, 1.5, 2.0
for m in [0.5, 1.0, 1.5, 2.0]:
    ci = honest.sensitivity_analysis(
        restriction='relative_magnitudes', m_bar=m)
    print(f"M-bar={m}: CI = [{ci.lower:.3f}, {ci.upper:.3f}]")

honest.plot_sensitivity()
plt.savefig('sensitivity.pdf', bbox_inches='tight')
```

Interpretation: $\bar{M} = 1$ means post-treatment violations are no worse than the largest pre-treatment violation. If the CI includes zero at $\bar{M} = 0.5$ but not at $\bar{M} = 0$, your result is fragile to even mild parallel-trends violations.

Empirical Lessons

Two JFE Findings Collapse Under Modern Methods

Baker, Larcker & Wang (2022) replications

Beck et al. (2010): Bank deregulation

TWFE: -0.022^{***} on log Gini
Goodman-Bacon: 86% of weight on problematic later-vs-earlier 2×2s

CS estimate: 0.001 (SE 0.007)
Stacked: 0.000 (SE 0.005)

The negative effect disappears entirely.

Fauver et al. (2017): Board reforms

TWFE: $+0.096^{***}$ on Tobin's Q
Goodman-Bacon: not feasible (unbalanced panel)

CS estimate: 0.062 (SE 0.135, insig.)
Stacked: 0.063 (SE 0.051, insig.)

Positive value effect not robust.

Both papers applied credible methodology *at the time*. The problem is the estimator, not the researchers.

“These New Methods Are Complicated and TWFE Was Good Enough for *JFE*”

The sceptic’s objections:

1. The new estimators are harder to implement and explain
2. TWFE worked for decades of published research
3. “My treatment timing isn’t that staggered”

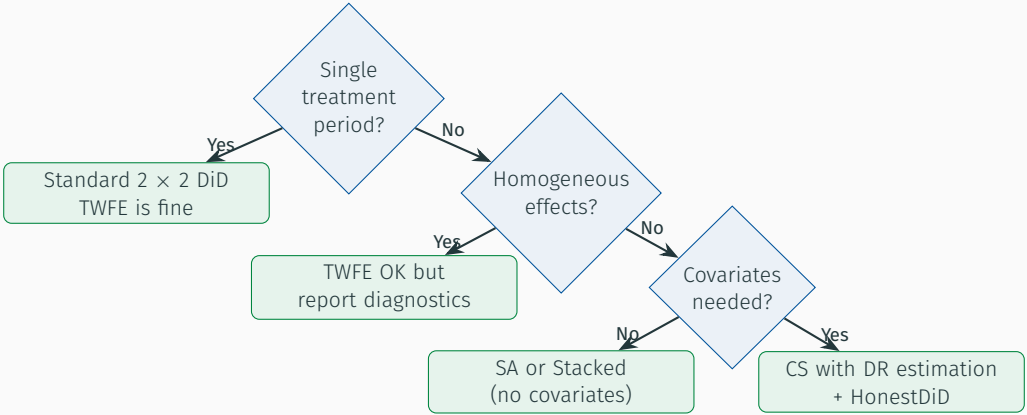
The response:

1. Software has caught up — `diff-diff` makes robust estimation no harder than TWFE
2. Two *JFE* results just collapsed under modern methods (you saw them on the last slide)
3. Even modest staggering with dynamic effects produces bias; the diagnostic takes 3 lines of code

The question is not whether you *can* use TWFE. It is whether Referee 2 will *let* you.

Practical Recommendations

A Decision Tree for Your Dissertation



Six Recommendations for Staggered DiD

Baker, Larcker & Wang (2022)

1. **Justify homogeneity** if using TWFE — or don't use it
2. **Diagnose bias:** plot treatment timing, run Goodman-Bacon decomposition
3. **Avoid binning** in event studies — fit the full set of relative-time indicators
4. **Apply $\geq$1 alternative estimator** (CS, SA, stacked, BJS)
5. **Justify your control group:** not-yet-treated, never-treated, or last-treated
6. **Report event-study plots** alongside any static estimate

For your dissertation: At minimum: (1) Goodman-Bacon diagnostic, (2) CS or SA event study, (3) HonestDiD sensitivity analysis. Report TWFE as a benchmark, not the main result. In Python: `diff-diff` handles all three steps.

Exercises

Exercise 1: Diagnose TWFE on Your Data

Try It Yourself

Using your dissertation panel dataset:

1. Load your data with `pl.read_csv()` or `pl.read_parquet()`
2. Run `TwoWayFixedEffects().fit()` and inspect the summary
3. Check the negative weight share — what fraction of weights are negative?
4. Run `check_parallel_trends()` and plot the results
5. If more than 10% of weights are negative, proceed to Exercise 2

If you do not yet have your own data, use the built-in example: `from diff_diff.datasets import load_mpdta`.

Exercise 2: Estimate and Visualise with Callaway-Sant'Anna

Try It Yourself

1. Run `CallawaySantAnna(control_group='not_yet_treated')` on your data
2. Plot the event study with `uniform_bands=True`
3. Re-run with `control_group='never_treated'` — do results change?
4. Export your event study plot as PDF with `fig.savefig('event_study.pdf')`

Questions to answer:

- Are the pre-treatment coefficients close to zero?
- How does the choice of control group affect the ATT?
- Do the simultaneous confidence bands differ much from the pointwise CIs?

Exercise 3: Sensitivity Analysis

Try It Yourself

1. Run `HonestDiD(cs.results_)` on your Callaway-Sant'Anna results
2. Test at $\bar{M} = 0.5, 1.0, 1.5, 2.0$ using `restriction='relative_magnitudes'`
3. At what value of $\bar{M}$ does the confidence interval first include zero?
4. Plot the sensitivity curve with `honest.plot_sensitivity()`

Interpret your findings:

- If the CI excludes zero at $\bar{M} = 1$, your result is robust to moderate violations
- If it includes zero at $\bar{M} = 0.5$, the result is fragile
- Report the breakdown value of $\bar{M}$ in your dissertation

Summary

Key Papers Covered Today

Paper	Key contribution
Diagnosis	
Goodman-Bacon (2021)	TWFE decomposition theorem
de Chaisemartin & D'H. (2020)	Negative weights diagnostic
Baker, Larcker & Wang (2022)	Practitioner guide, simulations, JFE replications
Solutions	
Callaway & Sant'Anna (2021)	Group-time ATT framework with DR estimation
Sun & Abraham (2021)	Interaction-weighted estimator, convex weights
Borusyak et al. (2024)	Imputation estimator, most efficient
Baker et al. (2025)	8-step forward-engineering framework
Sensitivity	
Roth (2022)	Pre-testing distorts inference
Rambachan & Roth (2023)	Honest sensitivity analysis

Software Ecosystem

Method	Python (`diff-diff`)	R	Stata
Callaway & Sant'Anna	`CallawaySantAnna`	`did`	`csdid`
Sun & Abraham	`SunAbraham`	`fixest::sunab()`	`eventstudyinteract`
Borusyak et al.	`EfficientDiD`	`didimputation`	`did_imputation`
de Chaisemartin & D'H.	`TwoWayFixedEffects`	`DIDmultiplegt`	`did_multiplegt`
Goodman-Bacon decomp.	—	`bacondecomp`	`bacondecomp`
Doubly robust DiD	`CS(dr=True)`	`DRDID`	—
Sensitivity analysis	`HonestDiD`	`HonestDiD`	—

Python: `uv pip install diff-diff` `diff-diff.readthedocs.io/en/stable/`

Guide: Baker et al. (2025) — worked examples in R, Stata, and Python.

DiD Checklist for Your Dissertation

Before you commit to a DiD strategy, work through these questions:

1. **Is your treatment staggered?** If yes, TWFE is suspect — proceed to step 2.
2. **Goodman-Bacon diagnostic:** What share of weights are negative? Does the decomposition show problematic timing comparisons?
3. **Which robust estimator?** CS for maximum flexibility; SA or stacked if you prefer regression-based workflows.
4. **Control group:** Are you using not-yet-treated (preferred), never-treated, or both as a robustness check?
5. **Pre-trends:** Do pre-treatment event-study coefficients look flat? Do not treat a passing pre-trends test as proof.
6. **Sensitivity:** At what $\bar{M}$ does the HonestDiD CI include zero? Report this breakdown value.

Week 7 — Key Takeaways

The diagnosis:

- TWFE uses already-treated units as controls, producing bias with dynamic effects
- Goodman-Bacon shows TWFE is a weighted average of all pairwise 2×2 DiDs
- Some weights can be negative — wrong sign, wrong inference

The cure:

- Define your target parameter first, then choose the estimator
- CS, SA, BJS, stacked — all estimate clean 2×2 building blocks
- Always run HonestDiD sensitivity analysis
- In Python: `diff-diff` handles the full pipeline

TWFE is a starting point, not a destination. Define your target → Diagnose TWFE → Estimate robustly → Sensitivity analysis.

55% of published DiD papers used staggered TWFE. Your dissertation will do better.