Staggered-Adoption Difference-in-Differences - Part II

ECON 2400 - Applied Econometrics II

Juan C. Yamin

Brown University

Spring 2026

Outline

Outline

Recap: why TWFE can break

$$Y_{it} = \alpha_i + \beta_t + \tau^{TWFE} D_{it} + \varepsilon_{it}, \text{ with staggered adoption } D_{it} = \mathbb{1}\{t \geq E_i\}.$$

- With heterogeneous or dynamic effects, TWFE recovers an *implicit* weighted average:

$$\tau^{TWFE} = \sum_{(i,t)\in\Omega_1} w_{it}\,\tau_{it}, \qquad \sum_{(i,t)\in\Omega_1} w_{it} = 1,$$

  where some w_{it} can be negative.

- **Forbidden comparisons:** TWFE uses already-treated units as controls for later-treated units. If effects evolve with exposure length, treated-as-controls comparisons can flip signs and create negative weights

- **FWL mechanics:** residualize treatment on unit and time FE, $\tilde{D}_{it} = D_{it} - \hat{D}_{it}$. Then $\hat{\tau}^{TWFE} \propto \sum_{it} \tilde{D}_{it} Y_{it}$. If $\hat{D}_{it} > 1$ in a treated cell, then $\tilde{D}_{it} < 0$, so that treated cell enters with negative weight

Recap: why Event Studies can break

$$Y_{it} = \alpha_i + \beta_t + \sum_{h \neq -1} \tau_h \, \mathbb{1}\{t = E_i + h\} + \varepsilon_{it}.$$

- **Under-identification:** if there are no never-treated units, the fully-dynamic path is not point identified
- **Negative weights:** Same intuition as the TWFE case.
- **Contamination:** even when identified, $\hat{\tau}_h$ can put weight on treated cells at event times $\ell \neq h$, so it becomes a signed mix of effects at different horizons rather than a clean "effect at h".

Outline

From last time's τ_{it} to $ATT(g,t)$

- Let $G_i \in \{2, \ldots, T, \infty\}$ be unit i's adoption date
- Let $y_{it}(g)$ be the outcome at t when i becomes treated at g ($y_{it}(\infty)$ is the never-treated path).

Group-time effect (the main building block). For each cohort g and time $t \geq g$,

$$ATT(g,t) \equiv \mathbb{E}\big[y_{it}(g) - y_{it}(\infty) \mid G_i = g\big].$$

- Last time we worked with cell-level effects τ_{it} on treated cells.
- $ATT(g,t)$ is just the cohort-g average of those effects at calendar time t.
- Event-time effects (effects after h periods of exposure) are aggregations of $ATT(g, g+h)$ across cohorts.

Estimation menu: same goal, different routes

Goal: avoid forbidden comparisons. TWFE mixes in treated-as-controls comparisons. Modern estimators avoid forbidden comparisons by:

- *Manual Averaging*:
  - Callaway–Sant'Anna (2021). Estimate $\widehat{ATT}(g,t)$ using only never-treated and/or not-yet-treated units, then aggregate.
  - de Chaisemartin–D'Haultfœuille (2020, 2024). Average valid 2×2 DID comparisons (good for on/off treatments).
- *Imputation*: Gardner (2021), Borusyak–Jaravel–Spiess (2024). Predict untreated outcomes, impute counterfactuals, then difference; can be more precise under stronger structure.

Callaway–Sant'Anna: identification assumptions

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

No anticipation (no pre-treatment effects):

$$\text{for all } i \text{ and } t < g, \qquad Y_{it}(g) = Y_{it}(\infty).$$

Parallel trends for untreated potential outcomes (two options):

- Never-treated controls (fixed control group):

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) \mid G_i = g] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) \mid G_i = \infty], \quad \forall\, t \geq g.$$

- Not-yet-treated controls (time-varying control group):

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) \mid G_i = g] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) \mid G_i \neq g,\ D_{it} = 0], \quad \forall\, t \geq g.$$

One practical difference to keep in mind:

- Never-treated: control group is stable over time (cleaner interpretation, possibly smaller sample).
- Not-yet-treated: uses future-treated units as controls before they adopt (more sample, but the control group changes with t and the PT restriction is stronger).

Main result: $ATT(g,t)$ is a long-difference 2×2 DiD

Under no anticipation and PT with never-treated controls:

$$ATT^{nev}_{unc}(g,t) = \mathbb{E}[Y_t - Y_{g-1} \mid G_i = g] - \mathbb{E}[Y_t - Y_{g-1} \mid G_i = \infty]\,.$$

Under no anticipation and PT with not-yet-treated controls:

$$ATT^{ny}_{unc}(g,t) = \mathbb{E}[Y_t - Y_{g-1} \mid G_i = g] - \mathbb{E}[Y_t - Y_{g-1} \mid D_{it} = 0,\ G_i \neq g]\,.$$

Intuition:

- This is the usual two-period two-group DiD logic, except we take a long difference from $g-1$ to t.
- With not-yet-treated controls, the comparison group depends on t (who is still untreated by t).
- Same goal: avoid treated-as-controls comparisons, so the estimator is built from valid 2×2 comparisons for each (g,t).

Never-treated controls: computing $ATT^{nev}(g, g+h)$

Fix a cohort g and a horizon $h \geq 0$. Let pre $= g-1$ and post $= g+h$.

Compute the same long-difference DiD, but with a fixed control group:

$$ATT^{nev}_{unc}(g, g+h) = \underbrace{\mathbb{E}[Y_{g+h} - Y_{g-1} \mid G_i = g]}_{\text{treated cohort change}} - \underbrace{\mathbb{E}[Y_{g+h} - Y_{g-1} \mid G_i = \infty]}_{\text{never-treated change}}.$$

Concrete example in the figure (horizon $h = 1$):

$$ATT^{nev}_{unc}(2, 3) = (Y_{A,3} - Y_{A,1}) - (Y_{D,3} - Y_{D,1}).$$

(The control group is always the same: never-treated)

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$				
$t = 3$	$h = 1$			
$t = 4$		$h = 1$		
$t = 5$				
$t = 6$			$h = 1$	

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$				
$t = 3$	$h = 1$			
$t = 4$		$h = 1$		
$t = 5$				
$t = 6$			$h = 1$	

Not-yet-treated controls: computing $ATT^{ny}(g, g+h)$

Fix a cohort g and a horizon $h \geq 0$. Let pre $= g-1$ and post $= g+h$.

Compute a long-difference DiD:

$$ATT^{ny}_{unc}(g, g+h) = \underbrace{\mathbb{E}[Y_{g+h} - Y_{g-1} \mid G_i = g]}_{\text{treated cohort change}} - \underbrace{\mathbb{E}[Y_{g+h} - Y_{g-1} \mid D_{i,g+h} = 0,\ G_i \neq g]}_{\text{controls not yet treated by post}}.$$

Concrete example in the figure (horizon $h = 1$):

$$ATT^{ny}_{unc}(2,3) = (Y_{A,3} - Y_{A,1}) - \frac{1}{2}\Big[(Y_{C,3} - Y_{C,1}) + (Y_{D,3} - Y_{D,1})\Big].$$

(The control set uses everyone untreated at $t = 3$, so it can change with post $= g + h$.)

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$				
$t = 3$	$h = 1$			
$t = 4$		$h = 1$		
$t = 5$				
$t = 6$			$h = 1$	

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$				
$t = 3$	$h = 1$			
$t = 4$		$h = 1$		
$t = 5$				
$t = 6$			$h = 1$	

Adding covariates: outcome regression, IPW, and doubly robust (DR)

Up to now no covariates, $ATT(g,t)$ is identified by a long-difference 2×2 DiD:

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

With covariates X, the estimand is the same, but we use conditional parallel trends. Two ways to adjust for X:

- Outcome regression (OR): model the untreated change

$$m^{nev}_{g,t}(X) = \mathbb{E}[Y_t - Y_{g-1} \mid X,\ G = \infty]$$

- Inverse probability weighting (IPW): reweight controls to look like cohort g in terms of X via the propensity score

$$p_g(X) = \Pr(G = g \mid X,\ G \in \{g, \infty\})$$

Doubly robust (DR) estimator:

- Combines OR and IPW in a single formula to estimate the same $ATT(g,t)$.

What csdid/did actually estimates (DR version, conceptually)

Goal: estimate the same group-time effect

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

How the software does it (for each (g,t)):

1. Pick a control set: never-treated ($G = \infty$) or not-yet-treated at time t ($D_{it} = 0$).
2. Compute long differences: $\Delta Y_{it} = Y_{it} - Y_{i,g-1}$.
3. Fit two nuisance pieces (possibly with ML):
   - outcome regression: predict the untreated change $\hat{m}_{g,t}(X) \approx \mathbb{E}[\Delta Y \mid X, \text{control}]$,
   - propensity score: estimate $\hat{p}_g(X) \approx \Pr(G = g \mid X, \text{eligible sample})$.
4. Combine them using a DR score:

$$\widehat{ATT}(g,t) = \text{average treated } (\Delta Y - \hat{m}_{g,t}(X)) - \text{weighted average controls } (\Delta Y - \hat{m}_{g,t}(X)).$$

Why DR matters: consistent if either the propensity model or the outcome model is correct.

Aggregation: what you report after estimating all $ATT(g,t)$

After estimating $\widehat{ATT}(g,t)$ for all $g \leq t$, you report an aggregate

$$\widehat{\theta} = \sum_{g \leq t} w_{g,t}\, \widehat{ATT}(g,t), \qquad w_{g,t} \geq 0, \quad \sum_{g \leq t} w_{g,t} = 1.$$

Common choices (interpretation and tradeoffs):

- Overall ATT: a single summary number (average effect across all treated cohorts and post periods).
- Event-study (event time $e = t - g$): dynamics as a function of time since adoption.
- Cohort averages: average over t within each g to show heterogeneity across adoption cohorts.
- Calendar-time averages: average over treated cohorts at each t (effect among those treated by t).

DCDH (2020) in staggered adoption: target is the adoption effect ($h = 0$)

Staggered adoption (absorbing): $D_{it} = \mathbb{1}\{t \geq G_i\}$.

Chaisemartin and D'Haultfœuille (2020) targets the effect *at adoption* for each cohort g:

$$ATT(g, g) = \mathbb{E}[Y_g(g) - Y_g(\infty) \mid G_i = g].$$

Key implication:

- focus is on the $h = 0$ slice (instantaneous effect at $t = g$); the estimand is an *instantaneous* effect at adoption
- no dynamic path $ATT(g, g + h)$ for $h > 0$ in this design
- They allow for on/off treatment in general setting
- coincides with CS when there are no covariates + uses not-yet-treated units as the comparison group + focus on $h = 0$

Identification intuition: compare switchers to still-untreated in $(g-1, g)$

Switchers at time g are the newly treated: $G_i = g$ (equivalently $D_{i,g-1} = 0,\ D_{ig} = 1$).

Controls at time g are those still untreated: $D_{ig} = 0$ (equivalently $G_i > g$ or $G_i = \infty$).

Identification uses only the adoption window $(g-1, g)$:

$$ATT^{ny}_{unc}(g, g) = \mathbb{E}[Y_g - Y_{g-1} \mid G_i = g] - \mathbb{E}[Y_g - Y_{g-1} \mid D_{ig} = 0].$$

Why this avoids the TWFE problem:

- already-treated units never serve as controls;
- the comparison is a two-period DiD, so earlier pre-periods are not used.

Estimation: compute DID_t each period, then average across adoption times

For each $t \geq 2$, compute the DiD based on switchers at t:

$$DID_t = \mathbb{E}[Y_t - Y_{t-1} \mid G_i = t] - \mathbb{E}[Y_t - Y_{t-1} \mid D_{it} = 0].$$

Then average across t with weights proportional to the number of switchers:

$$DIDM = \sum_{t=2}^{T} \frac{N_t}{\sum_{s=2}^{T} N_s} DID_t, \qquad N_t = \#\{i : G_i = t\}.$$

Interpretation:

- $DIDM$ is an average of the adoption effects $ATT(g, g)$ across cohorts;
- each DID_t uses not-yet-treated units at time t as controls.

DCDH intuition in staggered adoption: only the adoption window $(g-1, g)$

Left panel: cohort $g = 2$ is unit A; not-yet-treated at $g = 2$ are B, C, D:

$$ATT^{ny}_{unc}(2,2) = (Y_{A,2} - Y_{A,1}) - \frac{1}{3}\Big[(Y_{B,2} - Y_{B,1}) + (Y_{C,2} - Y_{C,1}) + (Y_{D,2} - Y_{D,1})\Big].$$

Right panel: cohort $g = 3$ is unit B; not-yet-treated at $g = 3$ are C, D (since A is already treated):

$$ATT^{ny}_{unc}(3,3) = (Y_{B,3} - Y_{B,2}) - \frac{1}{2}\Big[(Y_{C,3} - Y_{C,2}) + (Y_{D,3} - Y_{D,2})\Big].$$

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$	$h = 0$			
$t = 3$		$h = 0$		
$t = 4$				
$t = 5$			$h = 0$	
$t = 6$				

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$	$h = 0$			
$t = 3$		$h = 0$		
$t = 4$				
$t = 5$			$h = 0$	
$t = 6$				

Imputation: estimate the untreated path, then subtract

Goal is still to learn cohort-time effects $ATT(g,t)$ (and then aggregate), but the strategy is different from CS/DCDH.

Imputation idea:

- Use untreated observations to learn the counterfactual outcome under no treatment, $Y_{it}(\infty)$.
- For treated observations, impute $\widehat{Y}_{it}(\infty)$ and define a unit-time effect

$$\widehat{\tau}_{it} = Y_{it} - \widehat{Y}_{it}(\infty).$$

- Then build whatever estimand you want (overall ATT, event-study, cohort averages) by averaging $\widehat{\tau}_{it}$ with chosen weights.

What you assume (baseline version):

- No anticipation: untreated (never-treated + not-yet-treated) observations satisfy $Y_{it} = Y_{it}(\infty)$.
- A model for untreated outcomes (workhorse):

$$\mathbb{E}[Y_{it}(\infty)] = \alpha_i + \beta_t \qquad \text{(unit + time effects)}.$$

BJS / Gardner algorithm (no covariates): two steps

Step 1 (fit untreated outcome model on untreated sample):

$$Y_{it} = \alpha_i + \beta_t + \varepsilon_{it} \qquad \text{using only } (i,t) \text{ with } D_{it} = 0.$$

This yields $\widehat{\alpha}_i, \widehat{\beta}_t$ and therefore $\widehat{Y}_{it}(\infty) = \widehat{\alpha}_i + \widehat{\beta}_t$.

Step 2 (impute effects on treated observations):

$$\widehat{\tau}_{it} = Y_{it} - \widehat{\alpha}_i - \widehat{\beta}_t \qquad \text{for } (i,t) \text{ with } D_{it} = 1.$$

From cell-level effects to cohort-time effects:

$$\widehat{ATT}(g,t) = \text{average of } \widehat{\tau}_{it} \text{ over treated units with } G_i = g \text{ at time } t.$$

Then aggregate $\{\widehat{ATT}(g,t)\}$ into:

- overall ATT (single number),
- event-study effects by event time $e = t - g$,
- cohort averages or calendar-time averages.

Why imputation can be attractive (and what it costs)

It can be statistically efficient under a strong (but transparent) model for untreated outcomes!

- If untreated outcomes truly satisfy

$$Y_{it}(\infty) = \alpha_i + \beta_t + \varepsilon_{it} \quad \text{with homoskedastic, serially uncorrelated } \varepsilon_{it},$$

  then estimating α_i, β_t by OLS on the untreated sample yields an efficient estimator for any weighted average of effects.

- Proof idea: the imputation estimator is equivalent to OLS in a saturated model

$$Y_{it} = \alpha_i + \beta_t + \text{a dummy for each treated cell } (i,t) \text{ with its own coefficient} + \varepsilon_{it}.,$$

  which assigns a separate parameter to each treated cell; Gauss–Markov implies OLS is efficient for any linear combination.

Tradeoffs:

- The identifying content is in the untreated-outcome model. If that model is wrong, estimates can be biased.
- Inference must account for the first-step estimation of $\widehat{\alpha}_i, \widehat{\beta}_t$.

Outline

Continuous Treatments

Many real-world treatments have a "dose":

- Pollution exposure (dissipates across space)
- Tax rates, minimum wages, subsidy levels
- Hours of training, amount of spending

Setup (Callaway, Goodman-Bacon & Sant'Anna, 2024)

- Two periods: $t \in \{1, 2\}$. No unit is treated in period 1.
- In period 2, units receive a continuous dose $D_i \in \mathcal{D} = \{0\} \cup \mathcal{D}_+$
- Potential outcomes: $Y_{i,t}(d)$ — outcome unit i would experience under dose d
- No anticipation: $Y_{i,1} = Y_{i,1}(0)$ for all units

Parallel Trends (natural extension of the binary case):

$$E[Y_2(0) - Y_1(0) \mid D = d] \;=\; E[Y_2(0) - Y_1(0) \mid D = 0] \quad \forall\, d \in \mathcal{D}_+$$

In words: the average untreated outcome evolution for each dose group d is the same as for the untreated group. *Same idea as binary DiD, but must hold for every dose level.*

Two Types of Causal Effects

With a continuous treatment, there are **two fundamentally different questions**:

Level effect (ATT)

"What is the total effect of receiving dose d compared to no treatment?"

→ Height of the dose-response curve

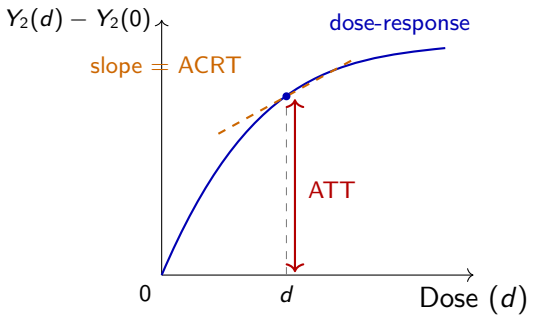
Relevant for: Should we implement this policy *at all*?

Causal response (ACRT)

"What is the effect of a small increase in dose at d?"

→ Slope of the dose-response curve

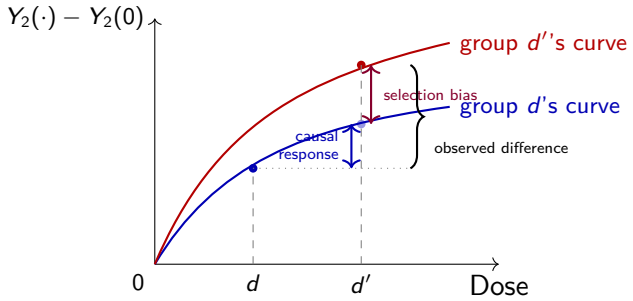
Relevant for: Should we *increase* the dose?



The Selection Bias Problem

ATT is straightforward: compare dose group d vs. untreated $\rightarrow$ standard DiD logic.

ACRT requires comparing *across dose groups*. But different groups may have different dose-response curves:



When we compare outcomes of group d' vs. group d, we get:

Observed difference = Causal response + Selection bias

Strong Parallel Trends

Strong Parallel Trends (SPT):

$$\mathbb{E}[Y_2(d) - Y_1(0)] = \mathbb{E}[Y_2(d) - Y_1(0) \mid D = d] \quad \forall d.$$

- Dose group d is representative: its average outcome under dose d matches the population average under dose d.
- Rules out selection into dose based on *gains* (which is what biases comparisons across dose groups).
- Not full homogeneity: treatment effects may vary, but not in a way that is systematically related to which dose is chosen $\Rightarrow$ Different dose groups don't have systematically different average dose-response curves. In expectation, the curve is the same no matter which dose group you look at.

Why TWFE Fails with Continuous Treatments

The standard approach: run $Y_{it} = \theta_t + \eta_i + \beta^{twfe} D_i \cdot \text{Post}_t + v_{it}$

This gives a **single number** β^{twfe} — but what does it estimate?
Problem: β^{twfe} admits multiple decompositions, and none is clean:

- As a weighted average of level effects $\rightarrow$ weights can be **negative** and integrate to zero
- As a weighted average of causal responses $\rightarrow$ weights are positive, but includes a selection bias term (unless SPT holds)
- Weights are driven by distance from mean dose — not by economic relevance

Even if strong PT holds:

- Weights depend on the *size of the untreated group*
- Dropping the untreated group changes β^{twfe} dramatically (78% in the CGS application!)

$\Rightarrow$ We need estimators that directly target well-defined causal parameters.

Estimation: Targeting What We Want

Callaway, Goodman-Bacon & Sant'Anna (2024)

Key idea: Define your target parameter first, then build an estimator for it.

1. Level effects — $ATT(d)$:

Compare dose group d to the untreated group: $\widehat{ATT}(d) = \bar{\Delta Y}\big|_{D=d} - \bar{\Delta Y}\big|_{D=0}$ or estimate $E[\Delta Y \mid D = d]$ flexibly (nonparametric).

For a **summary**, just binarize — compare any treated vs. untreated:

$$ATT^{glob} = E[\Delta Y \mid D > 0] - E[\Delta Y \mid D = 0]$$

2. Causal responses — $ACRT(d)$ (requires SPT): Take the derivative of the estimated $ATT(d)$ curve:

$$\widehat{ACRT}(d) = \frac{\partial\, E[\Delta Y \mid D = d]}{\partial\, d}$$

What If Treatment Is Truly Continuous?

de Chaisemartin, D'Haultfœuille & Vazquez-Bare (2024)

So far: We always had an untreated group ($D = 0$) to compare against.

Problem: Some treatments change for *everyone*. Think of temperatures, precipitation, prices — no county has *exactly* the same temperature two years in a row. There are **no stayers**.
Idea: Quasi-stayers Some units' treatment barely changed — a county whose temperature shifted by 0.01° is *almost* a stayer.

Large $\lvert\Delta D\rvert$	vs.	**Tiny** $\lvert\Delta D\rvert$
("treated")		("quasi-stayers" $\approx$ controls)

Compare outcome changes of units with large treatment shifts to those with near-zero shifts $\rightarrow$ recovers a **weighted average marginal effect**.
Trade-off: The "control group" shrinks as we demand $\lvert\Delta D\rvert$ closer to zero $\rightarrow$ less precise estimates. Parametric approaches can help.

Outline

Extension: Spatial Spillovers in DiD (Butts, 2024)

Standard DiD assumes treatment only affects treated units (SUTVA).
But many policies are assigned by geography and effects spill across borders:

- A new hospital serves residents from neighboring counties
- A factory boosts service-sector spending across the commuting zone

Potential outcomes now depend on two things: $Y_{it}(D_i,\ h_i(\mathbf{D}))$

- D_i: unit i's own treatment status
- $h_i(\mathbf{D})$: unit i's *exposure* to others' treatment (e.g. proximity to treated units)

$\Rightarrow$ some "control" units are exposed to spillovers, so their outcomes no longer reflect the counterfactual trend.

What Do We Want to Estimate?

Total effect — the national policymaker's question: "What was the overall effect of implementing the entire policy?"

$$\tau_{total} = E\big[Y_{i1}(1,\, h_i(\mathbf{D})) - Y_{i1}(0,\, 0) \mid D_i = 1\big]$$

Compares the world with treatment to a world with no treatment at all.
Spillover on controls — what happens to untreated but exposed units: "Did nearby untreated units benefit or lose out?"

$$\tau_{spill}(0) = E\big[Y_{i1}(0,\, h_i(\mathbf{D})) - Y_{i1}(0,\, 0) \mid D_i = 0\big]$$

Identification requires a key assumption — spillovers are *local*:

- Beyond some distance $\bar{d}$, units are unaffected: $h_i(\mathbf{D}) = 0$
- These "far-away" controls ($S_i(\bar{d}) = 0$) serve as the clean comparison group

$\rightarrow$ Total effect: compare treated to far-away controls
$\rightarrow$ Spillover effect: compare nearby controls to far-away controls

Estimation: Splitting the Control Group

The fix is simple — split controls into "potentially affected" and "clean":

$$Y_{it} = \tau\, D_i \cdot \mathbf{1}_{t=1} + \sum_{j=1}^{J} \gamma_j\, (1 - D_i)\, \text{Ring}_i^j \cdot \mathbf{1}_{t=1} + \mu_i + \lambda_t + \varepsilon_{it}$$

$\mu_i,\ \lambda_t$	Unit and time fixed effects (standard DiD)
τ	Total effect: treated vs. far-away controls
γ_j	Spillover on controls in ring j vs. far-away controls
Ring_i^j	Indicator for control unit i being in distance ring j (e.g. 0–25 mi, 25–50 mi, 50–100 mi from nearest treated)
Omitted category	Far-away controls with $S_i(\bar{d}) = 0$

Intuition: this is just standard DiD, but we acknowledge that not all controls are "clean." We let nearby controls have their own post-treatment shift (γ_j) and use only far-away units as the benchmark.

Event Study with Spillovers

With staggered adoption, we use an imputation approach
Stage 1: Estimate $\mu_i + \lambda_t$ using only observations that are both *untreated* and *unexposed*:

$$Y_{it} = \mu_i + \lambda_t + u_{it} \qquad \text{for } \{(i,t) : d_{it} = 0 \text{ and } s_{it} = 0\}$$

This is the key modification — we exclude nearby controls.
Stage 2: Residualize all observations: $\tilde{Y}_{it} = Y_{it} - \hat{\mu}_i - \hat{\lambda}_t$

Then regress on event-time dummies for treatment and spillover (this is just a convenient way to aggregate!):

$$\tilde{Y}_{it} = \sum_k \tau^k \, d_{it}^k + \sum_k \gamma^k \, s_{it}^k + \varepsilon_{it}$$

where d_{it}^k and s_{it}^k are indicators for being k years from treatment onset.
τ^k traces out the total effect over time (the treatment event study)
γ^k traces out how spillovers evolve over time

Application: Community Health Centers

Between 1965–1974, the federal government established Community Health Centers (CHCs) providing low-cost primary care to impoverished counties.

- Individuals in nearby counties could *travel* to CHCs to receive care
- If so, neighboring (untreated) counties also see mortality reductions
- Standard DiD would use these affected neighbors as controls → bias toward zero

Figure 2 — Total and Spillover Effects of Community Health Centers