A split-treatment design

Jean-Baptiste Bonnier

Université de Franche-Comté, CRESE

November 2024

Abstract

I introduce the split-treatment design, a framework in which the response to a treatment is split in the reactions to two or more events. I show that estimators in standard regression-based methods have no sensible causal interpretation in this setting as they may be subject both to negative weights and contamination bias. I then propose a simple method, a first-difference regression with sample constraints - the FD-DiD -, that allows to identify and estimate sensible causal parameters of interest. This estimator is straightforward to compute and efficient under random walk errors and unrestricted heterogeneity across groups and events. It additionally applies more generally to settings with several nonlinearly-dependent treatments. I revisit the application of Bernstein et al. (2019) to identify the effects of the introduction of a clearinghouse by the NYSE on counterparty risk.

Keywords: Difference-in-differences, Local projections, Heterogeneous treatment effects, Multiple treatments, Contamination bias

JEL Classification: C21, C23

1 Introduction

Differences-in-differences (DiD) are widely used in social sciences for estimating causal effects. Typically, identification relies on the "no anticipation" assumption which states that the treatment has no causal effect prior to its implementation. However, economic and finance theory is usually based on some version of the rational expectation hypothesis, i.e., agents are forward-looking, rational, and use available information to make decisions. As a consequence, in some settings, when treatment is anticipated, it is likely that to-be-treated individuals will react preemptively and adjust their behaviour before treatment actually occurs. An archetypical example is the delay between the time a piece of legislation is passed into law and the time of its implementation. In such situations, it is not always clear as to whether agents adjust their behaviour at the time the law is passed, at its implementation, or at both.

Consider as an example the implementation of a central clearing counterparty (CCP) in a financial market. The aim of this change is to introduce an insurance system by which the losses incurred by the failure of one's counterparty are insured by the CCP. Theory says that counterparty risk should fall, and that, if counterparty risk is priced, asset prices should react at the time of implementation. Yet, given rational expectations, some agents may anticipate this future fall in counterparty risk and adjust their behaviour after the announcement date, not because counterparty risk actually falls, but because assets subject to this change now have a higher average expected return. Therefore, it is unclear whether one should consider that treatment occurs at the implementation or the announcement date. This reasoning easily expands to a general case with more than two events. Progressive disclosure of information prior to treatment constitutes as many potential reasons for anticipated effects. In this paper, I devise a DiD design that accounts for the possibility that the effect of a treatment is split in the reactions to two or more events. I call this setting a split-treatment design.[1] Contrary to the familiar event-study case (Sun and Abraham, 2021; de Chaisemartin and d'Haultfœuille, 2023b), the split-treatment design allows for heterogeneous latency time

[1] To my knowledge, I am the first to explicitly focus on such a treatment assignment. A similar design is discussed in Section C.2 of de Chaisemartin and d'Haultfœuille's (2023a) Web Appendix, but they focus on the case with two treatments and they, more importantly, make very different assumptions for identification.

between two responses across individuals. My aim is then to derive sufficient conditions for identification and to provide a method for estimation of meaningful causal parameters of interest, namely weighted average "event effects".

Because the split-treatment design can be seen either as a framework with a single treatment with irregular dynamics or as a framework with several treatments, identification faces important challenges. In particular, a recent and yet abundant literature has shown that standard regression-based estimators fail, in general, in identifying interpretable causal estimands when treatment effects are heterogeneous. A first difficulty, pervasive in both single and multiple treatment specifications, pertains to the well-known issue of negative weights (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Borusyak et al., 2024). Clean identification fails because standard methods rely on "forbidden comparisons" of newly treated observations with already treated ones that may experience heterogeneous effects across groups or relative time since treatment. Furthermore, specifications with multiple treatments are affected by another issue, a contamination bias that stems from the nonlinear dependence structure that often exists between different treatments (Hull, 2018; Goldsmith-Pinkham et al., 2022; de Chaisemartin and D'Haultfœuille, 2023a).

The split-treatment design is concerned by both issues. I show how a simple and practical regression-based method in the spirit of Dube et al.'s (2023) local-projection based DiD (LP-DiD) allows to bypass these hurdles and estimate average "event effects". This solution relies on a first-difference regression with sample constraints to select clean controls. These constraints are different from those imposed in Dube et al.'s (2023) framework and will generally be easy to meet in real-world applications. Still, identification comes at the cost of important restrictions on the dynamics of "event effects". Although limiting, the structure of the split-treatment design itself brings dynamics to a single treatment, and it makes sense in some finance and economics applications for information to be incorporated rapidly into the outcome. Besides, although it is not in the spirit of the original design, nothing prevents practitioners from defining additional events as lagged effects of past events. Finally, this method, called first-difference DiD (FD-DiD), is efficient under random walk errors and unrestricted heterogeneity across groups and events.

Although I develop the FD-DiD for the split-treatment design, it has a much larger appeal. Specifically, in settings with several treatments that have a non-linear relationship (as in the case of mutually-exclusive treatments, for instance), the FD-DiD allows estimation of average treatment effects on the treated (ATT) for each treatment without contamination. Since it is regression-based, it is straightforward to control for covariates and easy to implement. Therefore, it may be the main takeaway of this paper that Dube et al.'s (2023) method can be adapted to settings with multiple treatments to make any contamination bias disappear provided some constraints are made on treatment effect dynamics.

The remainder of the paper is organized as follows. Section 2 presents the split-treatment design. Section 3 highlights how contamination bias prevents simple local-projection regressions to be used for identification and introduces the FD-DiD. Section 4 discusses the related literature with a focus on staggered designs with binary treatments and parallel trends, thereby leaving out papers that have assumed randomized treatment timing. Section 5 is dedicated to efficiency and inference. Section 6 presents some additional results. In Section 7, I use the FD-DiD to revisit the application of Bernstein et al. (2019) to identify the effects of the introduction of a clearinghouse by the *New York Stock Exchange* (NYSE) on counterparty risk. Section 8 concludes.

2 Split-treatment design

2.1 Setup

I consider a framework with N units, $i \in \{1, \ldots, N\}$, T periods, $t \in \{1, \ldots, T\}$, and a treatment split into E binary events, $e \in \{1, \ldots, E\}$. Treatment is only completed after event E, although all events can have an effect on the outcome. Let $D^e_{i,t}$ denote the dummy that takes the value 1 once event e has occurred for unit i, and E^e_i denote the period at which event e occurred for unit i, i.e., $E^e_i = \min\{t : D^e_{i,t} = 1\}$. For a never-treated unit i, I note $E_i^e = \infty$, $\forall e \in \{1, \dots, E\}$. The information set $\mathbb{D} = (D_{i,t}^e)_{(i,t,e)}$ contains the timing of all events for all units. For every treated unit i, events always happen in the same increasing order, so that, for $(e, e') \in \{1, \dots, E\}^2$, if $e < e'$, then $E_i^e < E_i^{e'}$. The design is staggered in the sense that units can get through any event at a different time.

Units can be divided into G mutually exclusive groups according to the timing of all E events, $\mathcal{G} = \{1, \dots, G\}$. That is, two units belong to the same group if and only if they go through all E events at the same time. This condition leads to a simple extension of the sharp design assumption. Let $N_{g,t}$ denote the number of observations in group g at period t, then: $\forall (i, g, t, e) \in \{1, \dots, N_{g,t}\} \times \{1, \dots, G\} \times \{1, \dots, T\} \times \{1, \dots, E\}$, $D_{i(g),t}^e = D_{g,t}^e$ and $E_{i(g)}^e = E_g^e$. Events are absorbing states: $\forall (g, e)$, $\forall t \geq 2$, $D_{g,t}^e \geq D_{g,t-1}^e$.

The number of periods between two events can be heterogeneous, i.e., $\exists (i, i', e) \in \{1, \dots, N\}^2 \times \{1, \dots, E\} : E_i^e - E_i^{e-1} \neq E_{i'}^e - E_{i'}^{e-1}$. With homogeneous latency time between events, the split-treatment design is nested in the event-study case. The latter has already been the focus of considerable attention in the literature (Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2023b), so that the interest of the split-treatment design lays on the case with heterogeneous latency time between events.

$Y_{i,t}$ denotes an observed outcome of interest. $Y_{i,t}^0(\mathbf{0'_E})$ and $Y_{i,t}^E(E_i^1, \dots, E_i^E)$, with $\mathbf{0_E}$ a null vector of size E, denote potential outcomes of unit i at time t without and with treatment, respectively.[2] Notations of potential outcomes reflect the fact that, in the case of dynamic "event effects", potential outcomes will depend on the timing of all E events. Similarly, let $Y_{i,t}^e(E_i^1, \dots, E_i^e, \mathbf{0'_{E-e}})$ denote the potential outcome of unit i at time t after events 1 to e occurred in periods E_i^1 to E_i^e, respectively, and event $e+1$ did not occur yet. To simplify notation, I often drop the E-vector of event timing. Likewise, although it is a slight abuse of notation I sometimes note $Y_{i,t}^e(\mathbf{g})$ to refer to the potential outcome of $Y_{i,t}$ as it went through the first e events associated with the path of group g. Treatment assignment and potential outcomes are treated as random variables independent across groups. Expectations are taken with respect to these random variables. The number of individuals in each group is treated

[2] Although, given previous definitions, noting $Y_{i,t}^\infty(\mathbf{0'_E})$ would be more intuitive, I use $Y_{i,t}^0(\mathbf{0'_E})$ instead to simplify notations.

as non-random.

2.2 Parameters of interest

The main estimand of interest is a weighted average "event effect" on the treated h periods after event e occurred. First, define the individual event effect h periods after event e occurred as:

$$\tau_{i,h}^{e} = E[Y_{i,E_i^e+h}^{e}(E_i^1,\dots,E_i^e,\mathbf{0'_{E-e}}) - Y_{i,E_i^e+h}^{e-1}(E_i^1,\dots,E_i^{e-1},\mathbf{0'_{E-e+1}})|\mathbb{D}] \tag{1}$$

Although several causal parameters can be of interest to researchers, this most disaggregated brick offers flexibility to envision heterogeneous effects across events, groups, or horizons.[3]

The target estimand is then:

$$\tau_{h}^{e} = \sum_{i} \omega_i \tau_{i,h}^{e} \tag{2}$$

The target can be an equally-weighted average such that $w_i = 1/N_1$ for all i - with N_1 the number of units that go through event e -, but it is not necessary. In particular, weights can depend on the treatment design.

Group-specific average event effects will also prove useful. They are defined by:

$$\tau_{g,h}^{e} = \frac{1}{N_g}\sum_{i\in g} \tau_{i(g),h}^{e} \tag{3}$$

where N_g is the number of units in group g and $i(g)$ denotes an individual i in group g.

2.3 Identifying assumptions

Causal parameters of interest involve several potential outcomes that can never be observed simultaneously for a given individual. Identifying assumptions are therefore needed to impute the mean counterfactual untreated outcome for treated units.

Assumption 1. *(Parallel trends)*

[3] It follows that the individual treatment effect h periods after treatment for unit i is given by the sum of all event effects in period $E_i^E + h$.

For all $(t, t') \in \{1, \ldots, T\}^2$ *and* $(i, i') \in \{1, \ldots, N\}^2$*:*

$$E[Y_{i,t}^0 - Y_{i,t'}^0 | \mathbb{D}] = E[Y_{i',t}^0 - Y_{i',t'}^0 | \mathbb{D}] \tag{4}$$

Assumption 1 (A1) states that the expected mean change in the untreated outcome is the same for every unit in every group. This assumption ensures that, had treated units not been treated, they would have evolved in the same manner as control units. In line with the literature, it is helpful to be more specific and to define a simple data-generating process that respects the parallel trend assumption for the untreated potential outcome: $E[Y_{i,t}^0(\mathbf{0'_E})] = \alpha + \alpha_i + \delta_t$, where α_i and δ_t are individual and time non-stochastic effects, respectively.[4]

Assumption 2. *(No anticipation)*
For all $i \in \{1, \ldots, N\}$*,* $e \in \{0, 1, \ldots, E\}$*,* $k \in \{0, \ldots, E - e\}$*, and* $t < E_i^{e+1}$*:*

$$Y_{i,t}^e(E_i^1, \ldots, E_i^e, \mathbf{0_{E-e}}) = Y_{i,t}^{e+k}(E_i^1, \ldots, E_i^e, \ldots, E_i^{e+k}, \mathbf{0_{E-e-k}}) \tag{5}$$

Assumption 2 (A2) adapts the no anticipation assumption to the split-treatment design. It stipulates that, for two distinct event paths with the same first e events, the potential outcome of unit i along these two paths will be identical for periods prior to event $e+1$. Put more simply, an event does not have an influence on the outcome before it occurs.

Assumption 3. *(Strict exogeneity)*
For all $(i, t) \in \{1, \ldots, N\} \times \{1, \ldots, T\}$*:* $E[\epsilon_{i,t} | \mathbb{D}] = 0$

Assumption 3 (A3) is a standard assumption of strict exogeneity. It states that shocks are mean independent of the treatment design.

[4] It is clear from the following that a weaker parallel trends assumption would have been sufficient for identification. Since alternative restrictions depend on realized treatment timing, and given that parallel trends is an assumption on potential outcomes, I prefer to follow Borusyak et al. (2024) and rely on this stronger assumption that is easier to justify *ex ante*.

With these assumptions the expected value of $Y_{i,t+h}$ can be decomposed to obtain:[5]

$$
\begin{aligned}
E[Y_{i,t+h}|\mathbb{D}] = \alpha + \alpha_i + \delta_{t+h} &+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau_{i,h}^{e}\mathbb{1}_{\{E_g^e=t\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{\infty}\tau_{i,h+j}^{e}\mathbb{1}_{\{E_g^e=t-j\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{h}\tau_{i,h-j}^{e}\mathbb{1}_{\{E_g^e=t+j\}}\right)\right) \qquad (6)
\end{aligned}
$$

Eq. (6) highlights that $E[Y_{i,t+h}|\mathbb{D}]$ can be expressed as the sum of its potential outcome without treatment and all prior dynamic event effects up until period $t+h$. In addition, subtracting $E[Y_{i,t-1}|\mathbb{D}]$ from $E[Y_{i,t+h}|\mathbb{D}]$ gives:

$$
\begin{aligned}
E[\Delta_h Y_{i,t}|\mathbb{D}] = \delta_t^h &+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau_{i,h}^{e}\mathbb{1}_{\{E_g^e=t\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{\infty}(\tau_{i,h+j}^{e}-\tau_{i,j-1}^{e})\mathbb{1}_{\{E_g^e=t-j\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{h}\tau_{i,h-j}^{e}\mathbb{1}_{\{E_g^e=t+j\}}\right)\right) \qquad (7)
\end{aligned}
$$

with $\delta_t^h = \delta_{t+h} - \delta_{t-1}$.

The intricacy of Eqs. (6) and (7) reveals that it will be difficult to identify event effects without additional assumptions. Dynamic effects of events prior to e make it impossible to use untreated units, while using other treated units requires additional constraints. Identification is possible, however, if event effects are static. I show in Section 6.2 that it is possible to weaken this assumption, but I choose to focus on static effects in the baseline specification for ease of exposition. Although restrictive, it is a reasonable assumption in many finance settings - especially since events themselves introduce flexible dynamics to a single treatment. Specifically, even with static effects, there can still be heterogeneous effects both across groups and in the latency time between events. It is an advantage of the split-treatment design -

[5] Detailed calculations can be found in Appendix A.

compared to event-study regressions - to cater for group heterogeneity in the effects relative to the time of treatment - i.e., a given event e may not be as far apart from the "complete treatment" for two different groups.

Assumption 4. *(Static effects)* $\forall (h, h') \in \mathbb{N}^2$*:*

$$\begin{aligned} E[Y^{e}_{i,E_i^e+h}(E_i^1, \dots, E_i^e, \mathbf{0'_{E-e}}) - Y^{e-1}_{i,E_i^e+h}(E_i^1, \dots, E_i^{e-1}, \mathbf{0'_{E-e+1}})|\mathbb{D}] = \\ E[Y^{e}_{i,E_i^e+h'}(E_i^1, \dots, E_i^e, \mathbf{0'_{E-e}}) - Y^{e-1}_{i,E_i^e+h'}(E_i^1, \dots, E_i^{e-1}, \mathbf{0'_{E-e+1}})|\mathbb{D}] \end{aligned} \tag{8}$$

that is: $\forall (h, h') \in \mathbb{N}^2$, $\tau^e_{i,h} = \tau^e_{i,h'}$.

Assumption 4 (A4) simply implies that the effect of an event e for individual i is the same for all periods $t \geq E_i^e$.[6] Therefore, under A4, $\tau^e_{i,h} = \tau^e_{i,0}$, $\forall h \in \mathbb{N}$, so that it is sufficient to focus on $h = 0$. The horizon subscript can then be discarded, and Eqs. (6) and (7) reduce to:

$$E[Y_{i,t}|\mathbb{D}] = \alpha + \alpha_i + \delta_t + \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e=1}^{E} \left(\sum_{j=0}^{\infty} \tau_i^e \mathbb{1}_{\{E_g^e = t-j\}} \right) \right) \tag{9}$$

and:

$$E[\Delta Y_{i,t}|\mathbb{D}] = \delta_t^0 + \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e=1}^{E} \left(\tau_i^e \mathbb{1}_{\{E_g^e = t\}} \right) \right) \tag{10}$$

2.4 Illustration

At this point, an illustration may help digest the notations and assumptions. Consider an example with 3 groups, 4 periods and 2 events. Group A is never treated, group B goes through event 1 in period 2 and event 2 in period 3, while group C goes through event 1

[6] An alternative to static effects would be to rely on assumptions suggested in Section C.2 of de Chaisemartin and D'Haultfœuille's (2023a) Web Appendix. In the two-treatment case, to identify the effect of the second treatment, in addition to parallel trends, no anticipation, and strong exogeneity, they need to rely on assumptions: (i) that the effect of the first treatment evolves over time in the same way in every group, and (ii) that there exists at least two groups that received the first treatment at the same date and the second treatment at different dates. These assumptions are equally very restrictive, and the choice of one or the other approach will depend on the type of applications at hand.

in period 3 and event 2 in period 4. Following the motivation of the paper, event 1 can be the announcement of a policy expected to have an impact on financial markets, while event 2 may correspond to the time of its implementation. Table 1 summarizes the treatment assignment. For simplicity, assume without loss of generality that there is a single individual in each group.

Table 1: Example - treatment assignment

	$D^1_{i,t}$			$D^2_{i,t}$		
	A	B	C	A	B	C
$t=1$	0	0	0	0	0	0
$t=2$	0	1	0	0	0	0
$t=3$	0	1	1	0	1	0
$t=4$	0	1	1	0	1	1

As a way to visualize the implication of static effects, Tables 2 and 3 summarize the expected value of the outcome conditional on the treatment assignment under A1-A3 and under A1-A4, respectively.

Table 2: Example - expected outcome under A1-A3

		$E[Y_{i,t}\vert\mathbb{D}]$	
	$g=\text{A}$	$g=\text{B}$	$g=\text{C}$
$t=1$	$\gamma_A+\delta_1$	$\gamma_B+\delta_1$	$\gamma_C+\delta_1$
$t=2$	$\gamma_A+\delta_2$	$\gamma_B+\delta_2+\tau^1_{B,0}$	$\gamma_C+\delta_2$
$t=3$	$\gamma_A+\delta_3$	$\gamma_B+\delta_3+\tau^1_{B,1}+\tau^2_{B,0}$	$\gamma_C+\delta_3+\tau^1_{C,0}$
$t=4$	$\gamma_A+\delta_4$	$\gamma_B+\delta_4+\tau^1_{B,2}+\tau^2_{B,1}$	$\gamma_C+\delta_4+\tau^1_{C,1}+\tau^2_{C,0}$
		$E[\Delta Y_{i,t}\vert\mathbb{D}]$	
	$g=\text{A}$	$g=\text{B}$	$g=\text{C}$
$t=2$	δ^0_2	$\delta^0_2+\tau^1_{B,0}$	δ^0_2
$t=3$	δ^0_3	$\delta^0_3+\tau^1_{B,1}-\tau^1_{B,0}+\tau^2_{B,0}$	$\delta^0_3+\tau^1_{C,0}$
$t=4$	δ^0_4	$\delta^0_4+\tau^1_{B,2}-\tau^1_{B,1}+\tau^2_{B,1}-\tau^2_{B,0}$	$\delta^0_4+\tau^1_{C,1}-\tau^1_{C,0}+\tau^2_{C,0}$

Table 2 makes it clear that identification will prove difficult without additional assumptions on the heterogeneity of event effects. By contrast, with static effects, Table 3 is sparser. A lot less event effects appear in the panel representing the expected change in the outcome.

Table 3: Example - expected outcome under A1-A4

	$E[Y_{i,t}\|\mathbb{D}]$		
	$g =$ A	$g =$ B	$g =$ C
$t = 1$	$\gamma_A + \delta_1$	$\gamma_B + \delta_1$	$\gamma_C + \delta_1$
$t = 2$	$\gamma_A + \delta_2$	$\gamma_B + \delta_2 + \tau_B^1$	$\gamma_C + \delta_2$
$t = 3$	$\gamma_A + \delta_3$	$\gamma_B + \delta_3 + \tau_B^1 + \tau_B^2$	$\gamma_C + \delta_3 + \tau_C^1$
$t = 4$	$\gamma_A + \delta_4$	$\gamma_B + \delta_4 + \tau_B^1 + \tau_B^2$	$\gamma_C + \delta_4 + \tau_C^1 + \tau_C^2$
	$E[\Delta Y_{i,t}\|\mathbb{D}]$		
	$g =$ A	$g =$ B	$g =$ C
$t = 2$	δ_2^0	$\delta_2^0 + \tau_B^1$	δ_2^0
$t = 3$	δ_3^0	$\delta_3^0 + \tau_B^2$	$\delta_3^0 + \tau_C^1$
$t = 4$	δ_4^0	δ_4^0	$\delta_4^0 + \tau_C^2$

I will use this simple example to illustrate the next section's results. In the following, I first show with the example of local projection (LP) regressions that standard regression-based methods do not identify a proper estimand in the split-treatment design. I then demonstrate, in the spirit of Dube et al.'s (2023) local projection based difference-in-differences (LP-DiD), that a simple first-difference regression combined with the assumption of static event effects allows to identify a convex combination of all group-specific event effects for a given event e.

3 A first-difference DiD estimator

3.1 Local projections and the split-treatment design

Although less popular than TWFE regressions, local projections (LP) can also be used in some settings to estimate treatment effects under appropriate assumptions of parallel trends and no anticipation.[7] In the two-group and multiple-period case, for instance, the coefficient associated with the regression, $\Delta_h Y_{i,t} = \delta^h_t + \beta^{h,LP}\Delta D_{i,t} + \epsilon^h_{i,t}$, where $\Delta_h Y_{i,t} = Y_{i,t+h} - Y_{i,t-1}$, provides an unbiased estimate of the ATT.[8]

Consider the LP regression for a given event e:

$$\Delta_h Y_{i,t} = \delta^h_t + \beta^{h,LP,e}\Delta D^e_{i,t} + \epsilon^h_{i,t} \tag{11}$$

One would want for $E[\hat{\beta}^{h,LP,e}]$ to capture the effects of event e for all treated groups. Under assumptions A1 to A3, it is clear that this regression will not identify a sensible estimand.[9] For the sake of simplicity, let's assume for now that $\delta^h_t = \delta^h$, $\forall(t,h)$. One has:

$$E[\hat{\beta}^{h,LP,e}|\mathbb{D}] = E[\Delta_h Y_{i,t}|\mathbb{D}, \Delta D^e_{i,t} = 1] - E[\Delta_h Y_{i,t}|\mathbb{D}, \Delta D^e_{i,t} = 0]$$

with:

$$\begin{aligned}
E[\Delta_h Y_{i,t}|\mathbb{D}, \Delta D^e_{i,t} = 0] = \delta^h &+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e'=1,e'\neq e}^{E}\left(\tau^{e'}_{i,h}\mathbb{1}_{\{E^{e'}_g=t\}}\right)\mathbb{1}_{\{E^e_g\neq t\}}\right)\\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e'=1}^{E}\left(\sum_{j=1}^{\infty}(\tau^{e'}_{i,h+j} - \tau^{e'}_{i,j-1})\mathbb{1}_{\{E^{e'}_g=t-j\}}\right)\mathbb{1}_{\{E^e_g\neq t\}}\right)\\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e'=1}^{E}\left(\sum_{j=1}^{h}\tau^{e'}_{i,h-j}\mathbb{1}_{\{E^{e'}_g=t+j\}}\right)\mathbb{1}_{\{E^e_g\neq t\}}\right)
\end{aligned}$$

[7] Since the first-difference difference-in-differences (FD-DiD) estimator I develop is directly related to LP regressions, I focus on their shortcomings to highlight the relevance of my approach. Complementary results on the shortcomings of TWFE under A1 to A4 are postponed in Section 6.1.

[8] See Dube et al. (2023) for more details on DiD settings where LP regressions can be useful.

[9] To highlight the usefulness of A4, I do not assume static effects in this development.

and:

$$
\begin{aligned}
E[\Delta_h Y_{i,t} | \mathbb{D}, \Delta D^e_{i,t} = 1] = \delta^h &+ \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \tau^e_{i,h} \mathbb{1}_{\{E^e_g = t\}} \right) \\
&+ \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=1}^{e-1} \left(\sum_{j=1}^{\infty} (\tau^{e'}_{i,h+j} - \tau^{e'}_{i,j-1}) \mathbb{1}_{\{E^{e'}_g = t-j\}} \right) \mathbb{1}_{\{E^e_g = t\}} \right) \\
&+ \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=e+1}^{E} \left(\sum_{j=1}^{h} \tau^{e'}_{i,h-j} \mathbb{1}_{\{E^{e'}_g = t+j\}} \right) \mathbb{1}_{\{E^e_g = t\}} \right)
\end{aligned}
$$

where, since the expected value is taken with respect to $\Delta D^e_{i,t} = 1$ in the second equation, $\Delta D^{e'}_{i,t'} = 0$ for $e' > e$ and $t' \leq t$ (events after e must happen after t), and $\Delta D^{e'}_{i,t'} = 0$ for $e' < e$ and $t' \geq t$ (events before e must happen before t). Hence:

$$
E[\hat{\beta}^{h,LP,e} | \mathbb{D}] = E[\Delta_h Y_{i,t} | \mathbb{D}, \Delta D^e_{i,t} = 1] - E[\Delta_h Y_{i,t} | \mathbb{D}, \Delta D^e_{i,t} = 0] \tag{12}
$$

$$
= \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \tau^e_{i,h} \mathbb{1}_{\{E^e_g = t\}} \right) \tag{12.1}
$$

$$
- \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=1, e' \neq e}^{E} \left(\tau^{e'}_{i,h} \mathbb{1}_{\{E^{e'}_g = t\}} \right) \mathbb{1}_{\{E^e_g \neq t\}} \right) \tag{12.2}
$$

$$
+ \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=1}^{e-1} \left(\sum_{j=1}^{\infty} (\tau^{e'}_{i,h+j} - \tau^{e'}_{i,j-1}) \mathbb{1}_{\{E^{e'}_g = t-j\}} \right) \mathbb{1}_{\{E^e_g = t\}} \right) \tag{12.3}
$$

$$
- \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=1}^{E} \left(\sum_{j=1}^{\infty} (\tau^{e'}_{i,h+j} - \tau^{e'}_{i,j-1}) \mathbb{1}_{\{E^{e'}_g = t-j\}} \right) \mathbb{1}_{\{E^e_g \neq t\}} \right) \tag{12.4}
$$

$$
+ \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=e+1}^{E} \left(\sum_{j=1}^{h} \tau^{e'}_{i,h-j} \mathbb{1}_{\{E^{e'}_g = t+j\}} \right) \mathbb{1}_{\{E^e_g = t\}} \right) \tag{12.5}
$$

$$
- \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=1}^{E} \left(\sum_{j=1}^{h} \tau^{e'}_{i,h-j} \mathbb{1}_{\{E^{e'}_g = t+j\}} \right) \mathbb{1}_{\{E^e_g \neq t\}} \right) \tag{12.6}
$$

After application of the law of iterated expectations, the population regression coefficient $E[\hat{\beta}^{h,LP,e}]$ will identify a weighted average of the individual effects of event e (12.1), plus five bias terms that highlight different sorts of "forbidden comparisons" (Borusyak et al., 2024). A plus sign in front of a bias term in Eq. (12) denotes a bias that stems from units "treated" for event e themselves, while a minus sign denotes a bias that comes from comparing it with inappropriate control units. The first bias (12.2) comes from comparisons with (i, t) cells in which individuals gets through events other than e, i.e., observations such that $\mathbb{1}_{\{E_g^e=t\}} = 0$ but $\mathbb{1}_{\{E_g^{e'}=t\}} = 1$ for some $e' \neq e$. The second bias (12.3) comes from potential dynamic effects of events prior to e for cells considered as "treated" for event e, i.e., observations such that $\mathbb{1}_{\{E_g^e=t\}} = 1$ and $\mathbb{1}_{\{E_g^{e'}=t-j\}} = 1$ for some j such that $1 \leq j < \infty$ and $e' \in \{1, \ldots, e-1\}$. The third bias (12.4) comes from comparison with (i, t) cells that are subject to dynamic effects of events that unit i went through in periods prior to t, i.e., observations such that $\mathbb{1}_{\{E_g^e=t\}} = 0$ and $\mathbb{1}_{\{E_g^{e'}=t-j\}} = 1$ for some j such that $1 \leq j < \infty$ and $e' \in \{1, \ldots, E\}$. (12.3) and (12.4) exist as long as there is some e' such that $\tau_{i,h+j}^{e'} \neq \tau_{i,j-1}^{e'}$, i.e., as long as event effects are dynamic.[10] The fourth bias (12.5) comes from the potential presence of units that go through one or several events subsequent to e between $t+1$ and $t+h$ among cells "treated" for event e, i.e., observations such that $\mathbb{1}_{\{E_g^e=t\}} = 1$ and $\mathbb{1}_{\{E_g^{e'}=t+j\}} = 1$ for some j such that $1 \leq j \leq h$ and e' such that $e' \in \{e+1, \ldots, E\}$. The last bias (12.6) stems from comparisons with (i, t) cells that go through one or several events between $t+1$ and $t+h$, i.e., observations such that $\mathbb{1}_{\{E_g^e=t\}} = 0$ and $\mathbb{1}_{\{E_g^{e'}=t+j\}} = 1$ for some j such that $1 \leq j \leq h$ and e' such that $e' \in \{1, \ldots, E\}$.

Although Eq. (12) has been derived for simple LP regressions, it clearly highlights that the clean control conditions used by Dube et al. (2023) will not be sufficient to identify a proper estimand in the case with several treatments and dynamic effects. I now assume event effects are static (A4) and explicitly characterize the weights associated with each group-specific event effect in regression (11) for $h = 0$.[11] Eq. (11) becomes a simple first-difference (FD)

[10] Also note that (12.3) and (12.4) do not partially cancel each other out without additional assumptions. Indeed, groups treated and not treated for event e at period t are mutually exclusive. It implies that: (i) for a given period t, the values of indicators $\mathbb{1}_{\{E_g^{e'}=t-j\}}$, $e' \in \{1, \ldots, e-1\}$, $1 \leq j < \infty$, need not be the same for units for which event e occurs at period t and for units for which it doesn't, and, (ii) even if there are treated and control groups that follow the same path of events prior to t, event effects can be heterogeneous across individuals and groups. A similar reasoning holds for (12.5) and (12.6).

[11] Depending on the context, researchers may be reluctant to assume static event effects, and would rather make an assumption of homogeneous effects across units:

Assumption 4'. *(Homogeneous effects) An event e have the same expected effect at a horizon h, $h \geq 0$, for* regression:

$$\Delta Y_{i,t} = \delta_t^0 + \beta^{FD,e} \Delta D_{i,t}^e + \epsilon_{i,t} \tag{13}$$

Proposition 1. *Suppose A1-A4 hold, then:*

$$\begin{aligned} E[\hat{\beta}^{FD,e}] &= E\left[\sum_{e'=1}^{E} \sum_{g,t:\Delta D^{e'}_{g,t}=1} \frac{N_g \widetilde{\Delta D}^{e}_{g,t}}{\sum_{g,t:\Delta D^{e}_{g,t}=1} N_g \widetilde{\Delta D}^{e}_{g,t}} \tau_g^{e'}\right] \\ &= E\left[\sum_{g,t:\Delta D^{e}_{g,t}=1} \frac{N_g (1-\Delta D^{e}_{.,t})}{\sum_{g,t:\Delta D^{e}_{g,t}=1} N_g (1-\Delta D^{e}_{.,t})} \tau_g^{e}\right. \\ &\quad \left. - \sum_{e'=1,e'\neq e}^{E} \sum_{g,t:\Delta D^{e'}_{g,t}=1} \frac{N_g \Delta D^{e}_{.,t}}{\sum_{g,t:\Delta D^{e}_{g,t}=1} N_g (1-\Delta D^{e}_{.,t})} \tau_g^{e'}\right] \end{aligned} \tag{14}$$

Even under the assumption of static effects, Proposition 1 shows that the coefficient in the FD regression with one event doesn't identify a proper estimand.[12] The population regression coefficient identifies a weighted sum of all group-specific event effects with weights that sum to 1 only for event e.[13] Note also that weights are all positive and bounded between 0 and 1 for event e, while they will all be negative for events $e' \neq e$.

While including indicators for all events in the regression solves the omitted variable

all units, i.e., $\forall i$*:*

$$\begin{aligned} &E[Y^{e}_{i,E^e_i+h}(E^1_i, \ldots, E^e_i, \mathbf{0'_{E-e}}) - Y^{e-1}_{i,E^e_i+h}(E^1_i, \ldots, E^{e-1}_i, \mathbf{0'_{E-e+1}})|\mathbb{D}] \\ &\qquad = E[Y^{e}_{i',E^e_{i'}+h}(E^1_{i'}, \ldots, E^e_{i'}, \mathbf{0'_{E-e}}) - Y^{e-1}_{i',E^e_{i'}+h}(E^1_{i'}, \ldots, E^{e-1}_{i'}, \mathbf{0'_{E-e+1}})|\mathbb{D}] \end{aligned}$$

In other words: $\tau^e_{i,h} = \tau^e_{i',h}$.

Eq. (12) may give the impression that the population regression coefficient in a FD-DiD regression similar to the one described below - but with adapted sample restrictions - could identify a proper estimand for the average event effect under such an assumption. This is misguided. Even if one restricts the sample to control units that go through events prior to e at the same time as treated units, and to a horizon h such that no other event occurs, both for treated and control units, between t and $t+h$, dynamic effects of past events of treated and control units will not cancel each other out as their weights will be different. In this situation, if such a control group does exist, it is still possible to compute a sum of simple difference-in-differences to identify the average event effect on the treated under A1-A3 and A4'. Yet, it is clear that Assumption 4' and associated sample conditions are very restrictive. Similar considerations are developed in Section C.2 of de Chaisemartin and D'Haultfœuille's (2023a) Web Appendix.

[12] The proofs of Proposition 1 and 2 are reported in Section B.1 of the Appendix.

[13] As shown in the Appendix, $E[\hat{\beta}^{FD,e}]$ actually identifies a weighted sum of individual event effects. Yet, since all individual effects in a same group have the same weight, it can be restated as a weighted sum of group-specific average event effects.

problem, it does not prevent contamination. Consider the regression:

$$\Delta Y_{i,t} = \delta^0_t + \sum_{e'=1}^{E} \beta^{FDm,e'} \Delta D^{e'}_{i,t} + \epsilon_{i,t} \tag{15}$$

Proposition 2. *Suppose A1-A4 hold, then:*

$$E[\hat{\beta}^{FDm,e}] = E\left[\sum_{e'=1}^{E} \sum_{g,t:\Delta D^{e'}_{g,t}=1} \frac{N_g \widetilde{\Delta D}^{e}_{g,t}}{\sum_{g,t:\Delta D^{e}_{g,t}=1} N_g \widetilde{\Delta D}^{e}_{g,t}} \tau^{e'}_{g}\right] \tag{16}$$

where $\widetilde{\Delta D}^{e}_{g,t}$ *are residuals obtained from the auxiliary regression of* $\Delta D^{e}_{g,t}$ *on time fixed effects and* $\{\Delta D^{e'}_{i,t}\}_{e',e'\neq e}$*.*

Despite their same expressions, the weights in Eqs. (14) and (16) are different. $E[\hat{\beta}^{FDm,e}]$ still identifies a weighted sum of all group-specific average event effects with weights that sum to 1 only for event e. Additionally, in this case, weights associated with events e' other than e sum to 0. Contamination from other events then only arises when event effects are heterogeneous across groups. It implies that, by contrast to TWFE regressions, when all event indicators are used as regressors, LP regressions have the advantage over TWFE that contamination from other events may be dampened if their effects are not too heterogeneous across groups (because of contamination weights summing to 0 for events $e' \neq e$).

Contamination bias stems from the nonlinear dependence structure between events (Goldsmith-Pinkham et al., 2022).[14] If events were linearly dependent, the residuals $\Delta\widetilde{D}^{e}_{i,t}$ in the auxiliary regression would be independent of $\Delta D^{e'}_{i,t}$, $e' \in \{1, \ldots, E\}$, $e' \neq e$, the contamination bias

[14] As highlighted by Goldsmith-Pinkham et al. (2022), contamination bias is an issue distinct from omitted variable bias. To see this, consider event effects are homogeneous such that $\tau^e_i = \tau^e$, $\forall(i,e)$, and denote $\Delta\widetilde{D}^e$ the vector of residuals in the auxiliary regression of ΔD^e on time fixed-effects and indicators for other events. One gets:

$$E[\hat{\beta}^{FDm,e}] = E\left[(\Delta\widetilde{D}^{e'}\Delta\widetilde{D}^{e})^{-1}\Delta\widetilde{D}^{e'}\Delta Y\right] = E\left[(\Delta\widetilde{D}^{e'}\Delta\widetilde{D}^{e})^{-1}\Delta\widetilde{D}^{e'}(\delta^0 + \sum_{f=1}^{E}\Delta D^f \tau^f + \epsilon)\right]$$

$$= E\left[(\Delta\widetilde{D}^{e'}\Delta\widetilde{D}^{e})^{-1}\Delta\widetilde{D}^{e'}\Delta\widetilde{D}^{e}\right]\tau^e + E\left[\sum_{f=1}^{E}(\Delta\widetilde{D}^{e'}\Delta\widetilde{D}^{e})^{-1}\Delta\widetilde{D}^{e'}\Delta D^f\right]\tau^f = \tau^e$$

would disappear, and the estimand would correspond to a weighted sum of group-specific event effects for event e. Instead, the dependence between events comes from the fact that the occurrence of event e is both conditioned by prior events having already happened and necessary for subsequent events to happen.[15] Hence: $\Delta\widetilde{D}^e_{i,t} \neq \Delta D^e_{i,t} - E[\Delta D^e_{i,t}|\mathbb{D}]$.

To see this more clearly, note that $E[\hat{\beta}^{FDm,e}]$ can be obtained as a two-step residualization. Let $\Delta\ddot{D}^{e'}_{i,t}$, $e' \in \{1,\dots,E\}$, be the demeaned residuals in the regression of $\Delta D^{e'}_{i,t}$ on time fixed effects. $\Delta\widetilde{D}^e_{i,t}$ can be obtained as the residuals in the regression of $\Delta\ddot{D}^e_{i,t}$ on $\Delta\ddot{D}^{e'}_{i,t}$, $e' \in \{1,\dots,E\}$, $e' \neq e$. Since for all e one generally has that: $E[\Delta D^e_{i,t}] \neq E[\Delta D^e_{i,t'}]$ for $t \neq t'$, the relationship between residuals $\Delta\ddot{D}^e_{i,t}$ and $\Delta\ddot{D}^{e'}_{i,t}$ varies by period, and the regression between the two averages across this relationship. As a result, the line of best fit does not isolate the within period variation in $\Delta D^e_{i,t}$, and the remaining variation in $\Delta\widetilde{D}^e_{i,t}$ will tend to predict variation in $\Delta D^{e'}_{i,t}$ within periods, making the contamination weights non-zero.

Another way to understand the respective issue of these two estimators is that, while $\hat{\beta}^{FD,e}$ leverages comparisons of treated units with units in the same period that did not go through event e but that may have been through other events, $\hat{\beta}^{FDm,e}$ leverages comparisons with all cells in periods when some units got treated for any event and infer erroneous relationships between events so that the effect of other events is not averaged away.[16]

As an example, for the illustration introduced in Section 2.4, for event 2, the LP with only

[15] The dates of occurrence of different events are not only mutually exclusive, but also ordered.

[16] To see this, note that $\hat{\beta}^{FD,e}$ can be decomposed as (see Section C of the Appendix for details):

$$\hat{\beta}^{FD,e} = \sum_t (1 - N^e_t/N) \sum_{i=1}^{N^e_t} \frac{\Delta Y_{i,t} - (1/N^0_t)\sum_{i=1}^{N^0_t} \Delta Y_{i,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g(1 - \Delta D^e_{\cdot,t})}$$

where N^e_t and N^0_t denote the number of units that go through event e in period t and the number of units that don't, respectively. This expression shows that $\hat{\beta}^{FD,e}$ leverages comparisons of the outcome of units treated for e with the outcome of units untreated for event e in the same period. In doing so, it leverages forbidden comparisons with units treated for another event in the same time period. The decomposition of $\hat{\beta}^{FDm,e}$ is not as telling:

$$\hat{\beta}^{FDm,e} = \sum_t \frac{\sum_{i:\Delta D^e_{i,t}=1} \widetilde{\Delta D}^e_{i,t} \Delta Y_{i,t} + \sum_{e'=1,e'\neq e} \sum_{i:\Delta D^{e'}_{i,t}=1} \widetilde{\Delta D}^e_{i,t} \Delta Y_{i,t}}{\sum_{i,t} (\widetilde{\Delta D}^e_{i,t})^2}$$

where $\widetilde{\Delta D}^e_{i,t}$ denote residuals in auxiliary regression of $\Delta D^e_{i,t}$ on time indicators and $\Delta D^{e'}_{i,t}$, $e' \in \{1,\dots,E\}$, $e' \neq e$.

the first event indicator gives: $E[\hat{\beta}^{FD,2}|\mathbb{D}] = 1/2\tau_B^2 + 1/2\tau_C^2 - 1/4\tau_C^1$. Because the regression doesn't take into account event 1, it leverages the outcome of group C in period 3 to compare it to the outcome of group B that is newly treated for event 2. By comparison, the LP with both event indicators gives: $E[\hat{\beta}^{FDm,2}|\mathbb{D}] = 7/15\tau_B^2 + 8/15\tau_C^2 + 2/15\tau_B^1 - 2/15\tau_C^1$. The within-period covariance between $\Delta\ddot{D}_{i,t}^1$ and $\Delta\ddot{D}_{i,t}^2$ is -1/9 in period 3 and 0 in other periods. After averaging between this relationship, remaining variations in $\Delta\widetilde{D}_{i,t}^1$ predict variations in $\Delta D_{i,t}^2$. This simple example illustrates how the announcement of a policy - if it is expected to have an effect on the outcome and if it is not accounted for - can contaminate the estimation of the treatment. The contamination effect can be large if the effect of event 1 is comparatively larger than the effect of event 2 and if the effects of the second event are heterogeneous across groups. Also note that, depending on the group that has the larger effect, $E[\hat{\beta}^{FDm,2}]$ can be both positively or negatively biased.

A possible solution would be to interact time indicators with event indicators so that the regression be saturated and thus capture the nonlinear dependence between events. This solution can be problematic, however, when there are a large number of events and treated groups relative to the size of the sample.

3.2 A first-difference DiD estimator

Dube et al. (2023) show that Jorda's (2005) local projection (LP) approach can be adapted to a DiD setting in order to solve the issue of negative weights that arise in standard (both static and dynamic) TWFE regressions in presence of heterogeneity across groups and relative time since treatment. Their approach, a local projection based difference-in-differences (LP-DiD), simply consists in estimating a LP regression, $\Delta_h Y_{i,t} = \delta_t^h + \beta^{h,LP-DiD}\Delta D_{i,t} + \epsilon_{i,t}^h$, on the restricted sample of newly treated observations ($\Delta D_{i,t} = 1$) and not-yet-treated ones ($\Delta D_{i,t-j} = 0$ for $-h \leq j < \infty$). These restrictions make "unclean comparisons" disappear. This approach can easily be adapted to the split-treatment design to make any contamination bias disappear provided that one makes appropriate restrictions.

From prior developments, consider the first difference based difference-in-differences (FD-

DiD) regression:

$$\Delta Y_{i,t} = \delta_t^0 + \beta^{FD-DiD,e} \Delta D_{i,t}^e + \epsilon_{i,t} \tag{17}$$

restricted to a sample of (i) observations newly "treated" for event e, i.e., (i,t) cells such that $\mathbb{1}_{\{E_i^e = t\}} = 1$, and (ii) corresponding control observations for which no event occurs at period t, i.e., (i,t) cells such that $\mathbb{1}_{\{E_i^{e'} = t\}} = 0$ for all $e' \in \{1, \dots, E\}$. I show below that the population analogue of $\hat{\beta}^{FD-DiD,e}$ identifies a convex combination of all group-specific effects for event e and I characterize explicitly the weights assigned to each group-specific average event effect τ_g^e.

First, I recompose the sets of groups $\mathcal{G} = \{1, \dots, G\}$ into a smaller set of groups $\mathcal{G}' = \{1, \dots, G'\}$ according to the timing of event e, so that each group in $\mathcal{G}'$ is composed of units that go through event e at the same date. I use $(\mathcal{G})$ or $(\mathcal{G}')$ to make it clear whether a group belongs to $\mathcal{G}$ or $\mathcal{G}'$. I can then define the clean control sample (CCS) for an event e for a particular group $g(\mathcal{G}')$, denoted $CCS_{g(\mathcal{G}')}^e$, as the set of observations at time $t = E_g^e$ that satisfy the restrictions associated with Eq. (17). With these definitions, one gets the following result.[17]

Theorem 1. *Suppose A1-A4 hold, then:*

$$E[\hat{\beta}^{FD-DiD,e}] = E\left[\sum_{g'=1}^{G} \omega_{g'}^{FD-DiD,e} \tau_{g'(\mathcal{G})}^e\right] \tag{18}$$

with:

$$\begin{aligned} \omega_{g'}^{FD-DiD,e} &= \frac{N_{g'}(1 - N_{g(\mathcal{G}')}/N_{CCS_{g(\mathcal{G}')}^e})}{\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1 - N_{g(\mathcal{G}')}/N_{CCS_{g(\mathcal{G}')}^e})} \\ &= \frac{N_{CCS_{g(\mathcal{G}')}^e} n_{g'}^e n_g^e n_{c,g}^e}{\sum_{g=1}^{G'} N_{CCS_{g(\mathcal{G}')}^e} n_g^e n_{c,g}^e} \end{aligned} \tag{19}$$

where $N_{CCS_{g(\mathcal{G}')}^e}$ is the number of units in $CCS_{g(\mathcal{G}')}^e$. $n_g^e = N_{g(\mathcal{G}')}/N_{CCS_{g(\mathcal{G}')}^e}$ and $n_{c,g}^e = N_{g(\mathcal{G}')}^0/N_{CCS_{g(\mathcal{G}')}^e} = 1 - N_{g(\mathcal{G}')}/N_{CCS_{g(\mathcal{G}')}^e}$ are the shares of treated units and control units in $CCS_{g(\mathcal{G}')}^e$, respectively, while $n_{g'}^e = N_{g'}/N_{g(\mathcal{G}')}$ is the share of group $g' \in \mathcal{G}$ in treated units of

[17] The proof of this result is reported in Appendix B.2.

recomposed group $g \in \mathcal{G}'$.

The product $n^e_g n^e_{c,g}$ is maximal for $n^e_g = 0.5$, while it goes to 0 as n^e_g gets closer to 0 or 1. It means that more weights is given to a group when it has a balanced share of treated units and controls, while it goes to 0, when treated or control units completely dominate the sample. Weights are all positive, proportional to the variance of the treatment dummy, $\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}}$, on subsample $CCS^e_{g(\mathcal{G}')}$ and to its size, $N_{CCS^e_{g(\mathcal{G}')}}$. Moreover, weights sum to 1 so that $E[\hat{\beta}^{FD-DiD,e}]$ identifies a convex combination of all group-specific effects for event e.[18] Coming back a final time to the example introduced in Section 2.4, the coefficient associated with the indicator for event 2 in a FD-DiD gives: $E[\hat{\beta}^{FD-DiD,2}|\mathbb{D}] = 3/7\tau^2_B + 4/7\tau^2_C$, which indeed corresponds to a convex sum of the appropriate event effects.

The FD-DiD estimator associated with Eq. (17) identifies a variance-weighted average event effect on the treated (AET).[19] In general, researchers may be more interested in an equally-weighted average effect instead. It can be obtained with a re-weighted FD-DiD regression. Indeed, Eq. (18) and (19) imply that the estimation of an FD-DiD regression through weighted least squares, assigning to an observation belonging to $CCS^e_{g(\mathcal{G}')}$ a weight equal to $\sqrt{1/(\omega^{FD-DiD,e}_{g'}/N_{g'})}$, where $g'(\mathcal{G}) \subset g(\mathcal{G}')$, identifies the equally-weighted AET for event e.[20]

[18] In the simplified case where an event e always occurs at a different date for two different groups, one gets the simplified result that:

$$E[\hat{\beta}^{FD-DiD,e}] = E\left[\sum_{g=1}^{G} \omega^{FD-DiD,e}_g \tau^e_g\right]$$

with:

$$\begin{aligned}\omega^{FD-DiD,e}_g &= \frac{N_g(1 - N_g/N_{CCS^e_g})}{\sum_{g=1}^{G} N_g(1 - N_g/N_{CCS^e_g})}\\ &= \frac{N_{CCS^e_g} n^e_g n^e_{c,g}}{\sum_{g=1}^{G} N_{CCS^e_g} n^e_g n^e_{c,g}}\end{aligned}$$

where $n^e_g = N_g/N_{CCS^e_g}$ and $n^e_{c,g} = 1 - N_g/N_{CCS^e_g}$ are the shares of treated units and of control units in CCS^e_g, respectively. Weights are positive and sum to 1.

[19] Eqs. (18) and (19) actually reveal that the FD-DiD estimator is a variance-weighted average of de Chaisemartin and D'Haultfœuille's (2023a) DID_{M} estimator for several treatments when there are only switchers that gets from untreated to treated for event e - because events are absorbing states - and with a larger set of control units - because of the static effect assumption.

[20] $\omega^{FD-DiD,e}_{g'}/N_{g'}$ can be computed from the data using Eq. (19) or with an auxiliary regression.

An alternative way to obtain an equally-weighted AET relies on an imputation approach.[21] First, use clean control units to estimate a counterfactual outcome change for each treated unit. To do so, regress $\Delta Y_{i,t}$ on time indicators using only clean control observations of the restricted sample, and use the estimated coefficients to get a predicted value of each treated unit in the absence of treatment $\widehat{\Delta Y_{i,t}}$. Then, compute the equally-weighted AET as follows: $N_1^{-1} \sum_{i,t:\Delta D^e_{i,t}=1} (\Delta Y_{i,t} - \widehat{\Delta Y_{i,t}})$, where N_1 is the number of (i,t) cells in which event e occurs.

To conclude this section, note that the FD-DiD has a larger appeal than the split-treatment design as it applies to settings with several nonlinear dependent treatments - as in the case of mutually exclusive treatments, for instance. One only needs to replace "events" by "treatments" in prior developments. It is the main takeaway of the paper. Under restricted treatment effect dynamics, the approach of Dube et al. (2023) can be adapted using different clean control conditions to settings with several treatments - including the split-treatment design - to properly estimate average treatment effects without contamination. For a given treatment, it only requires to focus on a restricted sample of units newly treated and control units that do not go through any event in that same period.

4 Connection with the literature

Two-way fixed effects (TWFE) regressions, i.e., regressions of an outcome variable on group and time fixed effects and a treatment, were up until recently routinely used to estimate treatment effects. de Chaisemartin and D'Haultfœuille (2023b) found that out of the 100 most-cited papers published in the *American Economic Review* from 2015 to 2019, 26 estimate a TWFE regression. Motivated by the fact that it is indeed true in the canonical case with two groups and two periods, TWFE estimators were thought to be equivalent to DiD estimators. As shown by several recent seminal papers, it is now clear that this assumption was misguided.

[21] I show in Section 5.1 how the FD-DiD estimator can be obtained by imputation (Borusyak et al., 2024; Harmon, 2023).

Under the assumptions of parallel trends and no anticipation, TWFE regressions with one treatment identify a weighted sum of effects for treated (g, t) cells with weights that may be negative, sum to one, and are not proportional to the share of that cell in the treated population (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Borusyak et al., 2024). In a staggered design, when treatment is binary, the TWFE estimator can be decomposed in a weighted sum of 2×2 DiD comparing the outcome evolution of two groups from a pre-period to a post-period with positive weights that differ from 0 if and only if one group switches treatment while the other does not (Goodman-Bacon, 2021). Some of these DiD compare a group that starts receiving treatment to another that remains untreated at both dates, while others compare a group that starts receiving treatment to one already treated at both dates. Negative weights come from this second type of DiD, which involves "forbidden comparisons" in presence of heterogeneous effects across groups or relative time. In some cases, it can even lead to all group-time average treatment effects being of the same sign, while the estimated treatment coefficient is of the opposite sign.

Dynamic TWFE regressions - or event-study regressions - are also a popular tool to estimate the effects of a binary treatment. They consist in regressing the outcome variable on group and time fixed effects, and on relative-time indicators equal to 1 if group g started receiving the treatment h periods ago. They are often used when treatment is thought to have dynamic effects. Although coefficients in this dynamic specification yield a sensible causal estimand of treatment effect when there is heterogeneity across relative time, they do not when one adds heterogeneity across groups. Sun and Abraham (2021) show in a staggered design with a binary treatment that the coefficient associated with indicator h equals the sum of two terms. The first one is a weighted sum across groups of the cumulative effect of $h + 1$ treatment periods, with weights summing to 1 but that may be negative. This term resembles the decomposition of the coefficient in the static TWFE. The second term is a weighted sum across relative periods $h' \neq h$ and across groups of the cumulative effect of $h' + 1$ treatment periods in group g, with weights summing to 0. This result implies that the estimate of the cumulative effect of $h + 1$ treatment periods may be contaminated by the cumulative effects of all other relative treatment periods.[22]

Contamination from dynamic effects in event-study regressions can be seen as a byproduct of a more general issue that arises in TWFE regressions with several treatments. In an early paper, Hull (2018) studied TWFE regressions in a framework that involves a choice between different treatment options. Each indicator in the regression then corresponds to a possible value of this multinomial treatment. His results already highlighted a contamination phenomenon from effects of other treatment states. In a more general framework, de Chaisemartin and D'Haultfœuille (2023a) show that, in TWFE regressions with several treatments, the coefficient on a given treatment identifies a weighted sum of that treatment effects across (g, t) cells with weights that sum to 1 but may be negative, plus weighted sums of the effects of the other treatments, with weights summing to 0 under certain conditions.[23] Moreover, including only one treatment in the regression does not prevent contamination - although contamination weights will be different.[24]

In a last related paper, Goldsmith-Pinkham et al. (2022) explain how contamination bias can arise in regressions with mutually exclusive treatments and a set of controls such that the treatments can be assumed to be independent of the potential outcomes conditional on those controls. Although the core of their paper focuses on "design-based" identifying assumptions, their analysis extend to "model-based" frameworks as well. A key aspect of the explanation lies in the fact that, to properly identify the effect of a treatment, the residuals in the auxiliary regression of this treatment indicator on covariates and other treatment indicators must be mean independent of these regressors - and not simply uncorrelated with them, which they are by construction. This condition is not satisfied when the dependence between treatments is nonlinear, as is the case with mutually exclusive treatments or, in

[22] Indicators for $h < 0$ have also been used in event-study regressions to test the validity of the parallel trends assumption. Sun and Abraham's (2021) results notably imply that this approach for testing parallel trends is misguided.

[23] Even when weights sum to 0, there can be contamination in the case of heterogeneous effects across groups.

[24] I derive similar results for the split-treatment design in Section 6.1. de Chaisemartin and D'Haultfœuille (2023a) go further and derive the maximal bias of the average effect of the first treatment both when it is the only treatment included in the regression and when all treatments are included. It shows that controlling for the other treatments does not necessarily lead to less biased estimators than not controlling for them.

the split-treatment design, with events whose occurrence is conditioned by the assignment of other events.

Because the split-treatment design can either be regarded as (i) a framework with a single treatment - the first event - with irregular dynamic effects - subsequent events -, or, alternatively, as (ii) a framework with several treatments, it is directly affected by the issues mentioned above. Several estimators have been proposed to address these problems. I present them briefly and then emphasize their differences with the estimator I propose.

de Chaisemartin and D'Haultfœuille (2023b) review recent estimators in the single-treatment case. A common intuition behind their development has been to carefully select valid controls in order to bypass the "forbidden comparisons" problem. de Chaisemartin and D'Haultfœuille's (2020) pioneering estimator allows to identify a proper causal effect by a weighted average of DiD. Sun and Abraham (2021) develop a regression-based estimator that allows dynamic treatment effects but requires homogeneity of those effects across units of a same cohort - i.e., units treated at the same time must experience the same path of treatment effects. Callaway and Sant'Anna (2021) propose estimators with similar properties but allow parallel trends to hold only after conditioning on covariates. Borusyak et al. (2024) propose a flexible imputation estimator that allows to efficiently estimate any combination of individual treatment effects.

In a recent working paper, Dube et al. (2023) exploit the fact that the local projection approach, originally developed in macroeconomics to estimate average treatment responses that are heterogeneous and dynamic, can be linked to DiD regressions. They propose the local projection based difference-in-differences (LP-DiD) approach that combines local projections with a clean control condition to estimate unbiased dynamic effects. Since, in a staggered design, bias comes from "forbidden comparisons" of newly treated units with already treated units that may be experiencing dynamic and heterogeneous effects, their clean control condition restricts the sample to ensure that only untreated units are used as controls for newly-treated ones.

The literature on DiD with several treatments is still narrow. de Chaisemartin and

D'Haultfœuille (2023a) propose an estimator that relies on common-trends assumptions and that is robust to heterogeneous effects and contamination bias. It generalizes de Chaisemartin and D'Haultfœuille's (2020) estimator to several treatments. To isolate the effect of the first treatment, their estimator compares the $t-1$-to-t evolution of switching groups, whose first treatment switches from $t-1$ to t while their other treatments do not change, to the outcome evolution of control groups (i) whose treatments all remain the same between $t-1$ to t, and (ii) that had the same treatments as the switching groups in period $t-1$. These conditions ensure that their estimator is robust to heterogeneous effects of all treatments.

In the split-treatment design, the latency between events is unrestricted across groups. It means that the effect h periods after the first event may correspond to the second event for some units but not for others. As a consequence, the target estimand defined in recent papers that allow for some form of dynamic effects become inappropriate (Sun and Abraham, 2021; Dube et al., 2023). Specifically, in my framework, a group-specific event effect will not overlap with the group-specific treatment effect h periods after the initial treatment date. In addition, compared to Sun and Abraham (2021), my estimator allows unrestricted heterogeneity of event effects across units, when theirs requires homogeneous effects for units of a same cohort.

Compared to de Chaisemartin and D'Haultfœuille's (2023a) framework with several treatments, I do not allow for units to switch back to their original state, i.e., events are absorbing. In return, my estimator requires far less stringent conditions for using units as controls. In real-world applications of the split-treatment design, the condition that allows to use only units with the same treatments as the switching groups in period $t-1$ seems very restrictive. With these restrictions, identification in the application proposed in Section 7 would not have been possible, for instance. In my baseline specification, all units that do not go through any event in t can be used as controls.

The approach I develop for identification is closely related to Dube et al.'s (2023) LP-DiD. It is a simple regression-based method with sample constraints that guarantee clean comparisons. As such, it contributes to further our comprehension of the link between local projections and difference-in-differences. The split-treatment design highlights the challenges that pertain to identification in settings with multiple treatments and unrestricted effect heterogeneity. Although identification requires important restrictions on event effect dynamics - restrictions that will be too limiting in many settings -, keep in mind that the "event" structure itself brings dynamics to a single treatment, and that, as explained in the introduction, these restrictions can be justified in some finance and macroeconomic settings. Specifically, in my baseline specification, I assume event effects are static, and I only weaken this assumption in Section 6.2. Since, with static effects, it is not necessary to analyse effects at horizon $h > 0$, I call the method first-difference DiD (FD-DiD) by analogy with Dube et al. (2023). An important takeaway is that, under some conditions, Dube et al.'s (2023) method can be adapted to settings with multiple treatments.

Finally, this work can also be directly related to the flexible framework of Borusyak et al. (2024) and its link with other recent DiD estimators as highlighted by Harmon (2023). In Harmon's (2023) terminology, the FD-DiD is a subgroup DiD estimator. As such, I show in the next section that it has a simple imputation form, and that it is efficient under random walk errors and unrestricted treatment effect heterogeneity.

5 Properties of the FD-DiD

In this section, I use the additional notations Ω_N, Ω_T, and Ω_1 to denote the sets of different units, of different time periods, and of units treated for event e in the restricted sample, respectively.

5.1 Efficiency

I consider a setting similar to Borusyak et al. (2024) and Harmon (2023) with unrestricted treatment effect heterogeneity to show that the FD-DiD estimator is efficient under random walk errors.

Assumption 5. *(Random walk errors)*

Let $\epsilon_{i,t} = \Delta u_{i,t}$. $\forall (i,t) \in \{1,\dots,N\} \times \{2,\dots,T\}$, $E[\epsilon_{i,t}] = 0$, $Var(\epsilon_{i,t}) = \sigma^2$, and for all $(i,t) \neq (i',t')$, $cov(\epsilon_{i,t}, \epsilon_{i',t'}) = 0$.

Assumption 5 (A5) states that $u_{i,t}$ follows a random walk without drift. It implies that $\epsilon_{i,t}$ are spherical errors, i.e., random variables with mean zero, homoskedastic, and uncorrelated over time and units.

Theorem 2. *Suppose A1-A5 hold, then $\hat{\beta}^{FD-DiD,e}$ is the efficient estimator of $\tau^e = \sum_{i\in\Omega_1} \omega_i \tau_i^e$ with $\tau_i^e = E[Y^e_{i,E_i^e} - Y^{e-1}_{i,E_i^e}|\mathbb{D}]$ and $\omega_i = (1-N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})/(\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1-N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}))$.*

Theorem 2 states that, under random walk errors, $\hat{\beta}^{FD-DiD,e}$ is the best linear unbiased estimator of a convex sum of event-e effects across all treated observations. Although random walk errors is only a benchmark, note that it is particularly relevant in finance applications with asset prices as the outcome - given that their behaviour is closely related to a random walk. Standard cluster-robust inference is likely to be preferred in applications where A5 is not satisfied.

5.2 Asymptotic properties

I have shown that the FD-DiD has an imputation representation and is efficient under random walk errors. I now derive asymptotic properties.[25] Convergence is studied along a sequence of unbalanced panels indexed by sample size $\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}$. This approach has the appeal that it applies to asymptotic sequences where both the number of units and the number of time periods may grow, although the assumptions are least restrictive when the number of time periods remains constant or grows slowly. While efficiency required spherical errors for $\epsilon_{i,t}$, I now make the standard assumption that errors are clustered by units.

Assumption 5'. *(Clustered errors)*

Error terms $\epsilon_{i,t}$ are independent across units i and have bounded variance $Var(\epsilon_{i,t}|\mathbb{D}) \leq \bar{\sigma}^2$ for all $(i,t) \in \Omega_N \times \Omega_T$ uniformly.

[25] Proofs of Theorems 3 and 4 can be found in Appendix E.

I assume that $N_g/N_{CCS^e_{g(\mathcal{G}')}}$ is bounded away from 0 and 1. It implies that, as the sample grows, the number of treated observations in group g does not grow disproportionately faster than the number of untreated observations in that clean control sample, and conversely.

Theorem 3. *Denote* $\omega_g = (1-N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})/(\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1-N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}))$*. Assume that A1-A4 and A5' hold and that,* $\forall g \in \mathcal{G}'$, $G'N_g/\left(\sum_{g=1}^{G'} N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}}\right)^2 \rightarrow 0$*. Then:* $\hat{\beta}^{FD-DiD,e} - \tau^e \xrightarrow{L_2} 0$.

The condition $G'N_g/\left(\sum_{g=1}^{G'} N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}}\right)^2 \rightarrow 0$ ensures that the weights are not too concentrated. It covers a large range of situations. In the simplified case where $N_g = N_g^0 = \bar{N}$, $\forall g \in \mathcal{G}'$, for instance, it is sufficient that G' or $\bar{N}$ goes to infinity to ensure consistency.

Theorem 4. *Under A1-A4 and A5', if there exists* $\kappa > 0$ *such that* $E[|\epsilon_{i,t}|^{2+\kappa}|\mathbb{D}]$ *is uniformly bounded, that* $\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}}/\left(\sum_{g=1}^{G'} N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}}\right) \rightarrow 0$*, and that:* $\sigma_e^2 \sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} > 0$ *with* $\sigma_e^2 = Var(\hat{\beta}^{FD-DiD,e})$*, then:* $\sigma_e^{-1}(\hat{\beta}^{FD-DiD,e} - \tau^e) \xrightarrow{d} \mathcal{N}(0,1)$.

This result establishes conditions under which the difference between estimator and estimand is asymptotically normal. $\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}}/\left(\sum_{g=1}^{G'} N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}}\right) \rightarrow 0$ ensures that the weights are not too concentrated, while $\sigma_e^2 \sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} > 0$ ensures that the variance does not vanish too quickly. It is also sufficient that G' or $\bar{N}$ goes to infinity to ensure asymptotic normality in the simplified case where $N_g = N_g^0 = \bar{N}$, $\forall g \in \mathcal{G}'$.

In some applications, it may be preferable to assume errors are clustered at a different level. In Section 7, for instance, treated units are stocks traded at the NYSE while control units are those exact same stocks traded at the CSE. It is therefore likely that residuals of observations in a same clean control sample will still exhibit correlation. Moreover, all stocks are used in one period and one period only, so that there is no need to account for serial correlation at the unit level. In this case, it seems more appropriate to rely on asymptotic properties that assume errors are clustered at the clean control sample level. It is straightforward to derive consistency and asymptotic normality weight conditions by adapting the proofs of Theorems 3 and 4. Specifically, consistency requires that $\forall g \in \mathcal{G}'$, $(N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}})^2/(\sum_{g=1}^{G'} N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}})^2 \rightarrow 0$, while asymptotic normality requires that

$\forall g \in \mathcal{G}'$, $\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}} \left(N_g N_g^0 / N_{CCS^e_{g(\mathcal{G}')}}\right) / \left(\sum_{g=1}^{G'} N_g N_g^0 / N_{CCS^e_{g(\mathcal{G}')}}\right) \to 0.$[26]

5.3 Inference

In Borusyak et al.'s (2024) imputation approach, individual treatment effects are estimated by fitting the observed outcome perfectly. As a consequence, residuals are null for all treated observations. By contrast, and although it can be obtained by imputation, the regression-based FD-DiD doesn't face such an issue. Indeed, the coefficient averages event effects across all treated units and does not require to estimate them individually. Inference can therefore rely on standard cluster-robust inference methods.[27]

6 Additional results

6.1 TWFE

TWFE regressions were routinely used to identify treatment effects. As explained in Section 2, a recent literature made it clear that this approach was misguided. Assume A1 to A4 hold. There is a clear parallel to be made between the analysis of TWFE regressions for the split-treatment design and de Chaisemartin and D'Haultfœuille's (2023a) analysis of TWFE regressions with several treatments. Each event can be considered as a distinct treatment. Consider first a TWFE regression that includes indicators for all events:

$$Y_{i,t} = \alpha + \alpha_g + \delta_t + \sum_{e'=1}^{E} \beta^{TWFEm,e'} D_{i,t}^{e'} + \epsilon_{i,t}$$

One finds:

$$E[\hat{\beta}^{TWFEm,e}] = E\left[\sum_{e'=1}^{E} \sum_{g,t:D_{g,t}^{e'}=1} \frac{N_g \widetilde{D}_{g,t}^{e}}{\sum_{g,t:D_{g,t}^{e}=1} N_g \widetilde{D}_{g,t}^{e}} \tau_g^{e'}\right] \tag{20}$$

[26] Proofs of these results are reported in Appendix F.

[27] See MacKinnon et al. (2023) for a recent literature review.

where $\widetilde{D}^e_{g,t}$ is the residual in the regression of $D^e_{i,t}$ on a constant, time and group fixed effects, and all other event indicators. $E[\hat{\beta}^{TWFEm,e}]$ identifies the sum of E terms. Each term is a weighted sum of group-specific AET for a different event. Weights on event e can be negative and sum to 1, while weights on other events do not. Additionally, weights on events $e' \neq e$ do not sum to 0 as events are not mutually exclusive (see de Chaisemartin and D'Haultfœuille's (2023a) Theorem 2).

Eq. (20) is similar to Theorem 2 in de Chaisemartin and D'Haultfœuille (2023a). Likewise, analogous to their Theorem 3, the expression of the coefficient associated with event e in a TWFE with only one event is given by:[28]

$$\begin{aligned}
E[\hat{\beta}^{TWFE,e}] &= E\left[\sum_{e'=1}^{E} \sum_{g,t:D^{e'}_{g,t}=1} \frac{N_g \widetilde{D}^e_{g,t}}{\sum_{g,t:D^e_{g,t}=1} N_g \widetilde{D}^e_{g,t}} \tau^{e'}_g \right] \qquad (21)\\
&= E\left[\sum_{e'=e}^{E} \sum_{g,t:D^{e'}_{g,t}=1} \frac{N_g(1 - D^e_{g,.} - D^e_{.,t} + D^e_{.,.})}{\sum_{g,t:D^e_{g,t}=1} N_g(1 - D^e_{g,.} - D^e_{.,t} + D^e_{.,.})} \tau^{e'}_g \right.\\
&+ \left(\sum_{e'=1}^{e-1} \sum_{g,t:D^e_{g,t}=1} \frac{N_g(1 - D^e_{g,.} - D^e_{.,t} + D^e_{.,.})}{\sum_{g,t:D^e_{g,t}=1} N_g(1 - D^e_{g,.} - D^e_{.,t} + D^e_{.,.})} \right.\\
&\left.\left. + \sum_{e'=1}^{e-1} \sum_{g,t:D^{e'}_{g,t}=1, D^e_{g,t}=0} \frac{N_g(-D^e_{g,.} - D^e_{.,t} + D^e_{.,.})}{\sum_{g,t:D^e_{g,t}=1} N_g(1 - D^e_{g,.} - D^e_{.,t} + D^e_{.,.})} \right) \tau^{e'}_g \right]
\end{aligned}$$

where, this time, $\widetilde{D}^e_{g,t}$ is the residual in the regression of $D^e_{i,t}$ on a constant, and time and group fixed effects only. Eq. (21) is the same as Eq. (20) but with different weights. Weights associated with event e can still be negative and still sum to 1, while those associated with other events do not. Therefore, TWFE regressions for the split-treatment design are affected both by negative weights (as highlighted, for instance, by de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2023b; Goldsmith-Pinkham et al., 2022; Borusyak et al., 2024) and contamination bias (Goldsmith-Pinkham et al., 2022).

[28] A notable difference is that de Chaisemartin and D'Haultfœuille's (2023a) Theorem 3 focuses on the two-treatment case instead of E treatments as in Eq. (21).

Also note that, compared to de Chaisemartin and D'Haultfœuille's (2023a) analysis, the split-treatment design imposes a particular dependence structure: it is not possible to keep the value of an event at 0 if the value of a subsequent event is at 1, or, conversely, to move the value of an event from 0 to 1 if the values of all preceding events are not 1. As shown in the development of Eq. (21), it has implications for the weights. For events prior to e, the form of the weights across (g, t) cells in which event e already occurred will differ from those cells in which event e did not occur yet. Contamination from cells where event e did not happen yet will enter the equation with a lower weight, that will more likely be negative.[29]

de Chaisemartin and D'Haultfœuille (2023a) also propose an estimator to solve this issue. It requires to select valid controls (i) that have the same treatments in $t-1$ as the switching group, and (ii) that keep the same treatments between $t-1$ and t. The first condition is quite restrictive for real-world applications of the split-treatment design. The FD-DiD solution I propose below only requires for the second condition to be satisfied. It also has the practical advantage of being estimable by regression. Its obvious drawback is that it is developed for a more restrictive setting than de Chaisemartin and D'Haultfœuille's (2023a).

6.2 Not-quite static effects

It is straightforward to see from the analysis in Section 4 that the split-treatment design can accommodate some form of dynamic effects: so long as both treated and counterfactual units included in the FD-DiD at time E_g^e don't experience additional effects from prior events between $E_g^e - 1$ and E_g^e, the FD-DiD will identify the proper estimand.

It directly implies that, in cases when it actually makes sense for an event to have dynamic effects for a few periods, it is possible to use a local-projection based difference-in-differences

[29] I do not go further in analyzing contamination bias in TWFE regressions as it is not particularly relevant to the main argument of the paper - so long as there is indeed contamination bias in TWFE regressions - and given that they have already attracted a lot of attention in the literature (de Chaisemartin and D'Haultfœuille, 2023a, for several treatments; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; Borusyak et al., 2024, among others, for a single treatment).

(LP-DiD) to identify them:

$$\Delta_h Y_{i,t} = \delta_t^h + \beta^{LP-DiD,e,h} \Delta D_{i,t}^e + \epsilon_{i,t}$$

where $\Delta_h Y_{i,t} = Y_{i,t+h} - Y_{i,t}$, and where the sample is restricted to (i) observations newly "treated" for event e and for which no other event occurs before $t+h$, i.e., cells such that: $\mathbb{1}_{\{E_i^e = t\}} = 1$ and $D_{i,t+h}^{e+1} = 0$, and (ii) corresponding control observations for which no event occurs between t and $t+h$, i.e., cells such that: $\forall (e', j) \in \{1, \ldots, E\} \times \{0, \ldots, h\}$, $\mathbb{1}_{\{E_i^{e'} = t+j\}} = 0$. It additionally requires the additional restriction that (iii) the effects of events that affected these units - both treated and controls - prior to t do not change between $t-1$ and $t+h$.

Under these conditions, $\hat{\beta}^{LP-DiD,e,h}$ identifies a convex combination of all group-specific effects for event e at horizon h. It only remains efficient under random walk errors (i) if no untreated observation enters or leaves the sample between t and $t+h$, (ii) if none of the units that could serve as control is left out because it goes through an event between t and $t+h$, and (iii) if none of the units that could serve as control is left out because it is expected to experience changes in the effect of prior events between t and $t+h$ (see Harmon, 2023).[30]

6.3 Covariates

To account for heterogeneous trends between units, consider now an extension of the DGP such that parallel trends only hold conditional on covariates.

Assumption 1'. *(Conditional parallel trends)*
For all $(t, t') \in \{1, \ldots, T\}^2$ and $(i, i') \in \{1, \ldots, N\}^2$:

$$E[Y_{i,t}^0 - Y_{i,t'}^0 | \mathbb{D}, x_{i,t}^1, \ldots, x_{i,t}^Q] = E[Y_{i',t}^0 - Y_{i',t'}^0 | \mathbb{D}, x_{i,t}^1, \ldots, x_{i,t}^Q] \quad (22)$$

Also assume a linear functional form of potential outcomes such that, on the restricted sample

[30] The later condition implies that the LP-DiD would not take advantage of the sample in the most efficient manner as variations in the outcome of such observations could be used in a stepwise difference-in-differences (SWDD) estimator (Harmon, 2023) in periods when the effects of past events are not expected to change.

associated with Eq. (17), the following specification is valid:

$$\Delta Y_{i,t} = \delta_t^0 + \beta^{FD-DiD,e} \Delta D_{i,t}^e + \sum_{q=1}^{Q} \gamma_q x_{i,t}^q + \epsilon_{i,t} \tag{23}$$

In this more general setting, one has:

$$E[\hat{\beta}^{FD-DiD,e}] = E\left[\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \frac{\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}} \sum_{g' \in g(\mathcal{G}')} \mathbb{1}_{\{i \in g'(\mathcal{G})\}} \tau_i^e}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} (\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2}\right]$$

where $\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}}$ now refers to the residuals in the regression of $\Delta D^e_{i,t}$ on time indicators and covariates on the restricted FD-DiD sample. $\hat{\beta}^{FD-DiD,e}$ still identifies a combination of all group-specific effects for event e. However, covariates alter the weighting scheme so that it becomes difficult to characterize weights analytically. Still, they can be computed in empirical applications through an auxiliary regression. Moreover, an equally-weighted average effect can be estimated using an imputation approach (see Section 5.1). It is also possible to control for pre-treatment characteristics semi-parametrically with methods in the spirit of Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021).

7 Application

7.1 Context

I use the method developed in this paper to revisit the application of Bernstein et al. (2019) published in the *Journal of Political Economy.*[31] Bernstein et al. (2019) use an almost-perfect historical experiment to study the effect of the multilateral netting function of a clearinghouse on counterparty risk. Before the introduction of multilateral netting, NYSE equities settled on a bilateral basis, which implies that brokers needed to write and receive checks/securities for every transaction. At the time, settlement needed to be made by the next day at 2:15 p.m. and brokers rarely had enough assets on hand to pay every single

[31] I would like to thank the authors for kindly providing data and information regarding their paper.

transaction. Additionally, customers would also buy securities on margin so that brokers often had to borrow the necessary additional funds. Banks were then forced to extend significant uncollateralized credit and day loans to brokers, and effectively provided them short-term leverage to finance their daily positions. Brokers also needed to finance positions via overnight collateralized borrowing at the call loan rate.

The volatility of the call loan rate led to an important number of broker defaults. One then understands that this bilateral system involved a large degree of exposure with both direct and contagion counterparty risks. With multilateral netting, transactions are netted by the clearinghouse so that the transfers of checks/securities vastly fall. Consequently, the number of transactions that can be defaulted on drops, and there is a reduction in both direct and contagion counterparty risk.

The historical experiment relies on the fact that, during the late nineteenth and early twentieth centuries, two exchanges, the New York Stock Exchange (NYSE) and the Consolidated Stock Exchange (CSE), coexisted. The CSE traded securities listed on the NYSE and was located just across the street. While the small exchange netted stock transactions through a clearinghouse starting in 1886, the NYSE did not until May 1892. Bernstein et al. (2019) use identical securities on the two exchanges to identify the effect of the introduction of multilateral netting at the NYSE using CSE-traded securities as controls. Since the two exchanges were so close geographically, arbitragers could indeed prevent price discrepancies not due to market liquidity or counterparty risk premia. They find that the introduction of netting on the NYSE increased the value of stocks by 24 basis points relative to the CSE.

7.2 Methodology

Periods of panics and threats from banks to suspend overcertification to NYSE brokers led to the creation of the NYSE clearinghouse in May 1892. The clearinghouse then engaged in multilateral netting across all NYSE members for a gradually-growing list of stocks. The NYSE started to introduce multilateral netting for four stocks on May 17 1892, and progressively extended this system to more stocks throughout the end of the nineteenth and beginning of the twentieth centuries as members became more familiar with it. The Committee of the clearinghouse of the NYSE met when they decided to clear additional stocks.

Bernstein et al. (2019) rely on monthly data to identify the effect of the introduction of the clearing house on counterparty risk. They focus on dual-listed stocks that are part of the original Dow Jones index between September 1886 and October 1896, and of its successors, the Dow Jones Railroad Index and the Industrial index, following its split. In finance theory, if markets are sufficiently liquid, new information should be integrated rapidly into prices. Therefore, the fall in counterparty risk should be priced in NYSE-traded stocks shortly after multilateral clearing is introduced. I extend Bernstein et al.'s (2019) analysis using daily prices hand-collected from archives of the *New York Times*.

Starting from Bernstein et al.'s (2019) analysis, the price of a stock can be decomposed as follows:

$$P_{i,t,E} = P_{i,t}^{Fun} - P_{i,t,E}^{MktLq} - P_{i,t,E}^{CP} + \epsilon_{i,t,E}$$

Where $P_{i,t,E}$ is the price on exchange E for stock i at time t, $P_{i,t}^{Fun}$ is the stock's fundamental value, $P_{i,t,E}^{MktLq}$ is the discount caused by the market illiquidity premia, $P_{i,t,E}^{CP}$ is the discount caused by the counterparty risk premium, and $\epsilon_{i,t,E}$ is market microstructure noise with mean zero. Taking expected value of the first difference gives:

$$E[\Delta P_{i,t,E}] = E[\Delta P_{i,t}^{Fun}] - E[\Delta P_{i,t,E}^{MktLq}] - E[\Delta P_{i,t,E}^{CP}]$$

The focus is on the estimation of $E[\Delta P_{i,t,E}^{CP}]$. Because the introduction of the clearinghouse may not be exogenous but related to market turmoil, one needs to account for changes in fundamental value.

Contrary to Bernstein et al. (2019), I do not restrict my analysis to stocks in the Dow Jones index and its successors. Instead, I use the minutes of the Committee of the clearinghouse of the NYSE to collect the dates at which all dual-listed stocks start being cleared through the clearinghouse.[32] I choose to do so to increase the size of the dataset that would

[32] Some stocks may be added to the clearing list before being dropped and added again. I only consider first addition to the list. From their dataset, it seems to have been the choice of Bernstein et al. (2019) as

have otherwise been too short, since there are not necessarily transactions in both exchanges for all stocks around the dates at which they start being centrally cleared at the NYSE. I then construct two-day price changes as the difference between the closing price of the day following the first clearing day and the closing price of the day preceding it.[33] I choose this two-day periods to account for the fact that market liquidity was not as important as today, and that information may take a little bit longer to be integrated into prices. Because this historical experiment provides a natural counterfactual for every stock, I only include stock prices at the CSE in the control group of each stock being newly centrally cleared at the NYSE. Occasionally, there are a few dates with more than two individuals as a few stocks may be added to the list simultaneously. The dataset is eventually composed of 158 data points for 79 individual stocks traded on both exchanges.

How does this application relate to the split-treatment design? Consider the fact that, following the approval by the Committee of the clearinghouse that a stock would join clearing, market participants anticipate a rise in the price of this stock - because of a lower expected counterparty-risk discount -, and that they decide to make money off of it. Informed traders could buy this stock and sell it a few days later after it actually joined clearing. Even though it is a risky bet, in the sense that counterparty risk did not actually fall just yet, and that its holder would then be exposed to changes in fundamentals for a few days, these stocks may still have had high expected returns. There are typically between two days and two weeks between the approval by the Committee and the implementation of the new system. Therefore, some of the reduction in counterparty risk may already be priced the day before a stock joins clearing. In this situation, analyzing the effect of multilateral netting on the change of stock prices just around the dates they start being centrally cleared may lead to an underestimation of the fall in counterparty risk. Prices can then react to two events: (i) the approval by the Committee that a stock will be cleared through the clearinghouse, and (ii) the actual implementation of this new system. Hence, I also construct two-day price

well.

[33] If there is no transaction either on the preceding day or the following day for one of the exchange, I exclude this stock from the dataset. If, however, the preceding or following day is a Sunday or a holiday, I take the closest closing price instead.

changes around dates of meetings of the Committee of the clearinghouse when it is decided that some new stocks will be centrally cleared. This dataset is composed of 154 data points for 77 individual stocks traded on both exchanges.

I estimate the following regressions:

$$\Delta P_{i,t,E} = \delta_t + \beta^{FD-DiD,e} \Delta D^e_{i,t} + \epsilon_{i,t,E}$$

with $e \in \{ann, imp\}$, where *ann* and *imp* denote announcement and implementation, respectively. δ_t denote time fixed effects. In most time periods, when there are only two individuals - the price of the stock on the NYSE being newly centrally cleared and its control on the CSE -, stock fixed effects would be confounded with time fixed effects. Although taking the first difference of a variable should make individual fixed effects disappear, I still use cluster-robust variance estimates at the stock level to account for possible changes in fundamental value. Changes in the fundamental value are time varying, but since, in this setting, I only include prices of individual stocks at one date, clustering at the stock level should be sufficient. $\Delta D^e_{i,t}$ is a dummy that takes the value 1 for stocks listed on the NYSE - treated individuals - and 0 for stocks listed on the CSE - controls.[34]

7.3 Results

Table 4 shows the results. Columns 1 and 3 show results for the announcement with, respectively, price change and return as the dependent variable, while columns 2 and 4 show results for the implementation. Results suggest that there was no price reaction following the formal approval by the Committee of the clearinghouse that a stock will be centrally cleared. There is, however, a significant rise in NYSE prices following the implementation of multilateral netting. Using the return specification - although it is just below standard significance thresholds -, the introduction of multilateral clearing on the NYSE reduces the

[34] I also controlled for the average volume, volume change, and volume percentage change over the three-day period to account for difference in market liquidity between the two exchanges as well as for possible change in market liquidity premia. It is not significant and does not change the results. Similarly, including day fixed effects does not affect the results. Results are available upon request.

average counterparty risk premium by 28 bp. It is reassuring that this result is of the same order as Berstein et al.'s (2019) first specification (column 1 in Table 2 of their paper).[35] It suggests that, even at the time, prices reacted to new information rapidly.

Table 4: Application - results

	(1)	(2)	(3)	(4)
$\hat{\beta}^{FD-DiD,ann}$	0.018 (0.103)		0.190 (0.875)	
$\hat{\beta}^{FD-DiD,imp}$		0.214** (0.084)		0.285 (0.174)

Notes: Figures in parentheses are standard errors. *, **, and *** denote rejections of the null hypothesis, H0: $\hat{\beta}^{FD-DiD} = 0$, at the 10%, 5%, and 1% significance level, respectively.

8 Conclusion

In this paper, I introduce the split-treatment design, a framework in which the response to some treatment is split in the reactions to two or more events. I show that estimators in standard regression-based methods have no sensible causal interpretation in this setting as they may be subject both to negative weights and contamination bias. I then propose a simple method, a first-difference regression with sample constraints - the FD-DiD -, that allows to identify and estimate sensible causal parameters of interest. This estimator is straightforward to compute and efficient under random walk errors and unrestricted heterogeneity across groups and events.

Although I develop the FD-DiD with the split-treatment design in mind, it has a larger appeal. Specifically, in settings with several treatments that have a nonlinear relationship - as in the case of mutually exclusive treatments -, the FD-DiD allows estimation of average treatment effects on the treated (ATT) for each treatment without contamination. It is a regression-based estimator, which implies that it is straightforward to control for covariates and easy to implement. Moreover, although the estimator is derived for static effects for

[35] Berstein et al. (2019) found a reduction in the counterparty risk premium of 24 bp. The slightly larger estimation obtained in this paper is most likely due to the different specification of the dependent variable or to the different sample.

ease of exposition, this assumption can weakened to allow for some dynamics. Therefore, it may be the main takeaway of this paper that Dube et al.'s (2023) method can be adapted to settings with multiple treatments provided it is sensible to constraint treatment effect dynamics.

References

Bernstein, A., E. Hughson, and M. Weidenmier (2019). Counterparty Risk and the Establishment of the New York Stock Exchange Clearinghouse. *Journal of Political Economy 127*(2), 689–729.

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies*.

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996.

de Chaisemartin, C. and X. D'Haultfœuille (2023). Two-way fixed effects and differences-in-differences estimators with several treatments. *Journal of Econometrics 236*(2), 105480.

de Chaisemartin, C. and X. D'Haultfœuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econometrics Journal 26*(3), C1–C30.

Dube, A., D. Girardi, O. Jordà, and A. M. Taylor (2023). A Local Projections Approach to Difference-in-Differences Event Studies. NBER Working Paper No. w31184.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2022). Contamination Bias in Linear Regressions. Working Papers 2022-15, Princeton University. Economics Department.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Harmon, N. A. (2023). Difference-in-Differences and Efficient Estimation of Treatment Effects. Working Paper.

Hull, P. (2018). Estimating Treatment Effects in Mover Designs. arXiv:1804.06721 [econ].

Jordà, O. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review 95*(1), 161–182.

MacKinnon, J. G., M. O. Nielsen, and M. D. Webb (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*(2), 272–299.

Sant'Anna, P. H. C. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*(1), 101–122.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Appendix

A Identifying assumptions

According to my setup and identifying assumptions, it is possible to decompose the expected value of $Y_{i,t+h}$ and $\Delta_h Y_{i,t}$ as follows.

$$
\begin{aligned}
E[Y_{i,t+h}|\mathbb{D}] &= E[Y^0_{i,t+h}|\mathbb{D}] + \sum_{g=1}^{G}(E[Y_{i,t+h}(\mathbf{g}) - Y^0_{i,t+h}|\mathbb{D}]\mathbb{1}_{\{i\in g\}}) \\
&= E[Y^0_{i,t+h}|\mathbb{D}] + \sum_{g=1}^{G}\left(\sum_{e=1}^{E}\left(E[Y^e_{i,t+h}(\mathbf{g}) - Y^0_{i,t+h}|\mathbb{D}]\mathbb{1}_{\{E^e_g\leq t+h<E^{e+1}_g\}}\right)\mathbb{1}_{\{i\in g\}}\right) \\
&= E[Y^0_{i,t+h}|\mathbb{D}] + \sum_{g=1}^{G}\left(\sum_{e=1}^{E}\left(E[Y^e_{i,t+h}(\mathbf{g}) - Y^0_{i,t+h}|\mathbb{D}](D^e_{g,t+h} - D^{e+1}_{g,t+h})\right)\mathbb{1}_{\{i\in g\}}\right) \\
&= E[Y^0_{i,t+h}|\mathbb{D}] + \sum_{g=1}^{G}\left(\sum_{e=1}^{E}\left(E\left[\sum_{r=1}^{e}(Y^r_{i,t+h}(\mathbf{g}) - Y^{r-1}_{i,t+h})|\mathbb{D}\right](D^e_{g,t+h} - D^{e+1}_{g,t+h})\right)\mathbb{1}_{\{i\in g\}}\right) \\
&= E[Y^0_{i,t+h}|\mathbb{D}] + \sum_{g=1}^{G}\left(\sum_{e=1}^{E}\left(\sum_{r=1}^{e}E[Y^r_{i,t+h}(\mathbf{g}) - Y^{r-1}_{i,t+h}|\mathbb{D}](D^e_{g,t+h} - D^{e+1}_{g,t+h})\right)\mathbb{1}_{\{i\in g\}}\right) \\
&= E[Y^0_{i,t+h}|\mathbb{D}] + \sum_{g=1}^{G}\left(\sum_{e=1}^{E}\left(\sum_{r=1}^{e}\tau^r_{i,t+h-E^r_g}(D^e_{g,t+h} - D^{e+1}_{g,t+h})\right)\mathbb{1}_{\{i\in g\}}\right) \\
&= \alpha + \alpha_i + \delta_{t+h} + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left((D^e_{g,t+h} - D^{e+1}_{g,t+h})\sum_{r=1}^{e}\tau^r_{i,t+h-E^r_g}\right)\right) \\
&= \alpha + \alpha_i + \delta_{t+h} + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau^e_{i,t+h-E^e_g}\sum_{r=e}^{E}(D^r_{g,t+h} - D^{r+1}_{g,t+h})\right)\right)
\end{aligned}
$$

Note that:

$$
\begin{aligned}
&\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau^e_{i,t+h-E^e_g}\sum_{r=e}^{E}(D^r_{g,t+h} - D^{r+1}_{g,t+h})\right) \\
&\qquad = \mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau^e_{i,t+h-E^e_g}D^e_{g,t+h}\right)
\end{aligned}
$$

Hence:

$$\begin{aligned}
E[Y_{i,t+h}|\mathbb{D}] &= \alpha + \alpha_i + \delta_{t+h} + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau^e_{i,t+h-E^e_g}D^e_{g,t+h}\right)\right) \\
&= \alpha + \alpha_i + \delta_{t+h} + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=-h}^{\infty}\tau^e_{i,h+j}\mathbb{1}_{\{E^e_g=t-j\}}D^e_{g,t+h}\right)\right) \\
&= \alpha + \alpha_i + \delta_{t+h} + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=-h}^{\infty}\tau^e_{i,h+j}\mathbb{1}_{\{E^e_g=t-j\}}\right)\right) \\
&= \alpha + \alpha_i + \delta_{t+h} + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau^e_{i,h}\mathbb{1}_{\{E^e_g=t\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{\infty}\tau^e_{i,h+j}\mathbb{1}_{\{E^e_g=t-j\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{h}\tau^e_{i,h-j}\mathbb{1}_{\{E^e_g=t+j\}}\right)\right)
\end{aligned}$$

Where the third equality stems from the fact that $\mathbb{1}_{\{E^e_g=t-j\}} = 1$ always implies that $D^e_{g,t+h} = 1$ for $-h \leq j < \infty$. For $h = -1$, the equation above results in:

$$E[Y_{i,t-1}|\mathbb{D}] = \alpha + \alpha_i + \delta_{t-1} + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{\infty}\tau^e_{i,j-1}\mathbb{1}_{\{E^e_g=t-j\}}\right)\right)$$

And subtracting $E[Y_{i,t-1}|\mathbb{D}]$ from $E[Y_{i,t+h}|\mathbb{D}]$ gives:

$$\begin{aligned}
E[\Delta_h Y_{i,t}|\mathbb{D}] &= \delta^h_t + \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\tau^e_{i,h}\mathbb{1}_{\{E^e_g=t\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{\infty}(\tau^e_{i,h+j} - \tau^e_{i,j-1})\mathbb{1}_{\{E^e_g=t-j\}}\right)\right) \\
&+ \sum_{g=1}^{G}\left(\mathbb{1}_{\{i\in g\}}\sum_{e=1}^{E}\left(\sum_{j=1}^{h}\tau^e_{i,h-j}\mathbb{1}_{\{E^e_g=t+j\}}\right)\right)
\end{aligned}$$

with $\delta^h_t = \delta_{t+h} - \delta_{t-1}$.

B A first-difference DiD estimator

B.1 First-difference estimator

I develop here the expression of the population regression coefficient associated with regression (13). First, according to the Frisch-Waugh-Lovell theorem:

$$E[\hat{\beta}^{FD,e}|\mathbb{D}] = \frac{\sum_{i,t} \widetilde{\Delta D}^{e}_{i,t} E[\Delta Y_{i,t}|\mathbb{D}]}{\sum_{i,t}(\widetilde{\Delta D}^{e}_{i,t})^2}$$

where $\widetilde{\Delta D}^{e}_{i,t} = \Delta D^e_{i,t} - \Delta D^e_{.,t}$ are residuals from the auxiliary regression of $\Delta D^e_{i,t}$ on time fixed effects. Using Eq. (10) and the fact that $\sum_i \widetilde{\Delta D}^{e}_{i,t} = 0$, one gets:

$$\begin{aligned}
E[\hat{\beta}^{FD,e}|\mathbb{D}] &= \frac{\sum_{i,t} \widetilde{\Delta D}^{e}_{i,t} \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=1}^{E} \left(\tau_i^{e'} \mathbb{1}_{\{E_g^{e'}=t\}}\right)\right)}{\sum_{i,t}(\widetilde{\Delta D}^{e}_{i,t})^2} \\
&= \frac{\sum_{i,t} \widetilde{\Delta D}^{e}_{i,t} \sum_{g=1}^{G} \left(\mathbb{1}_{\{i \in g\}} \sum_{e'=1}^{E} \left(\tau_i^{e'} \Delta D^{e'}_{g,t}\right)\right)}{\sum_{i,t} \widetilde{\Delta D}^{e}_{i,t} \Delta D^e_{i,t}} \\
&= \frac{\sum_{e'=1}^{E} \sum_{g=1}^{G} \sum_{i,t: i \in g} \widetilde{\Delta D}^{e}_{i,t} \tau_i^{e'} \Delta D^{e'}_{g,t}}{\sum_{i,t:\Delta D^e_{i,t}=1} \widetilde{\Delta D}^{e}_{i,t}} \\
&= \frac{\sum_{e'=1}^{E} \sum_{g=1}^{G} \sum_{i,t: i \in g, \Delta D^{e'}_{g,t}=1} \widetilde{\Delta D}^{e}_{g,t} \tau_i^{e'}}{\sum_{i,t:\Delta D^e_{g,t}=1} N_g \widetilde{\Delta D}^{e}_{g,t}} \\
&= \frac{\sum_{e'=1}^{E} \sum_{g,t:\Delta D^{e'}_{g,t}=1} \widetilde{\Delta D}^{e}_{g,t} N_g \tau_g^{e'}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g \widetilde{\Delta D}^{e}_{g,t}} \\
&= \sum_{e'=1}^{E} \sum_{g,t:\Delta D^{e'}_{g,t}=1} \frac{N_g \widetilde{\Delta D}^{e}_{g,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g \widetilde{\Delta D}^{e}_{g,t}} \tau_g^{e'} \\
&= \sum_{g,t:\Delta D^{e}_{g,t}=1} \frac{N_g \widetilde{\Delta D}^{e}_{g,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g \widetilde{\Delta D}^{e}_{g,t}} \tau_g^{e} + \sum_{e'=1, e' \neq e}^{E} \sum_{g,t:\Delta D^{e'}_{g,t}=1} \frac{N_g \widetilde{\Delta D}^{e}_{g,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g \widetilde{\Delta D}^{e}_{g,t}} \tau_g^{e'} \\
&= \sum_{g,t:\Delta D^{e}_{g,t}=1} \frac{N_g (1 - \Delta D^e_{.,t})}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g (1 - \Delta D^e_{.,t})} \tau_g^{e} - \sum_{e'=1, e' \neq e}^{E} \sum_{g,t:\Delta D^{e'}_{g,t}=1} \frac{N_g \Delta D^e_{.,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g (1 - \Delta D^e_{.,t})} \tau_g^{e'}
\end{aligned}$$

where the fifth equality comes from $\tau_g^{e'} = \frac{1}{N_g} \sum_{i=1}^{N_g} \tau_i^{e'}$. Eq. (14) follows from the law of iterated expectations applied to the sixth equality. Eq. (16) is obtained with the same steps.

In addition, for proposition 2, for $e' \neq e$, $\sum_{g,t:\Delta D^{e'}_{g,t}=1} N_g \widetilde{\Delta D}^e_{g,t} = \sum_{g,t} N_g \widetilde{\Delta D}^e_{g,t} \Delta D^{e'}_{g,t} = 0$ (by definition of the linear projection and orthogonality).

B.2 First-difference based difference-in-differences (FD-DiD)

Proof of Theorem 1. I recompose the sets of groups $\mathcal{G} = \{1, \ldots, G\}$ into a smaller set of groups $\mathcal{G}' = \{1, \ldots, G'\}$ according to the timing of event e, so that each group in $\mathcal{G}'$ is composed of units that go through event e at the same date. In the following, I use $(\mathcal{G})$ or $(\mathcal{G}')$ to make it clear whether a group belongs to $\mathcal{G}$ or $\mathcal{G}'$. I can then define the clean control sample (CCS) for an event e for a particular group $g(\mathcal{G}')$, denoted $CCS^e_{g(\mathcal{G}')}$, as the set of observations at time $t = E^e_g$ that satisfy the restrictions associated with Eq. (17).

One then has an unbalanced panel dataset defined by the clean control condition that can be ordered as a stacked dataset in which observations are grouped into consecutive and non-overlapping $CCS^e_{g(\mathcal{G}')}$. For variable $\Delta D^e_{i,t}$, for instance, for a given event e, this vector would first be composed of the subvector of ΔD^e, at time $t = E^e_1$, whose observations satisfy CCS^e_1, i.e., observations of units that either go through event e at time $t = E^e_1$ or are clean controls (i.e., units that do not go through any event at time $t = E^e_1$), then of a second subvector of ΔD^e, at time $t = E^e_2$, whose observations satisfy CCS^e_2, and so on until the subvector whose observations satisfy $CCS^e_{G'}$.

Moreover, for any observation $\{i, t\} \in CCS^e_{g(\mathcal{G}')}$, $\Delta D^e_{i,t} = \Delta D^e_{i,E^e_{g(\mathcal{G}')}} = D^e_{i,E^e_{g(\mathcal{G}')}}$. This equality follows from the fact that $D^e_{i,t-1} = D^e_{i,E^e_{g(\mathcal{G}')}-1} = 0$ by virtue of the clean control condition.

One can create a set of G' binary indicators that identify the CCS that an observation belongs to. For each g, the corresponding indicator is equal to 1 if $\{i, t\} \in CCS^e_{g(\mathcal{G}')}$, and 0 otherwise. By definition of treatment groups and CCS, these indicators are fully collinear with time indicators.

Let $\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}}$ be the residuals from a regression of ΔD^e on time indicators in the sample satisfying restrictions associated with Eq. (17). In matrix form, the regression is:

$$\Delta D^e = \lambda\delta + \widetilde{\Delta D}^e$$

with $\lambda = (\lambda^e_1, \dots, \lambda^e_{G'})$, where $\lambda^e_{g(\mathcal{G}')}$ is a $(\sum_{g'=1}^{G'} N_{CCS^e_{g'(\mathcal{G}')}} \times 1)$-vector that takes the value 1 when observation i belongs to $CCS^e_{g(\mathcal{G}')}$ and 0 otherwise, and $N_{CCS^e_{g'(\mathcal{G}')}}$ is the number of observations in $CCS^e_{g'(\mathcal{G}')}$; $\delta = (\delta_1, \dots, \delta_{G'})'$ is a $(G' \times 1)$ vector of coefficients; $\Delta D^e = (\Delta D^e_{1,E^e_1}, \dots, \Delta D^e_{N_{CCS^e_1},E^e_1}, \Delta D^e_{1,E^e_2}, \dots, \Delta D^e_{N_{CCS^e_{G'}},E^e_{G'}})'$ is a $(\sum_{g'=1}^{G'} N_{CCS^e_{g'(\mathcal{G}')}} \times 1)$ vector; and $\widetilde{\Delta D}^e = (\widetilde{\Delta D}^e_{1,E^e_1}, \dots, \widetilde{\Delta D}^e_{N_{CCS^e_1},E^e_1}, \widetilde{\Delta D}^e_{1,E^e_2}, \dots, \widetilde{\Delta D}^e_{N_{CCS^e_{G'}},E^e_{G'}})'$ is a $(\sum_{g'=1}^{G'} N_{CCS^e_{g'(\mathcal{G}')}} \times 1)$ vector of residuals. λ are CCS indicators that conveniently replace time indicators. Using OLS to estimate δ, one gets $\hat{\delta} = (N_1/N_{CCS^e_1}, \dots, N_{G'}/N_{CCS^e_{G'}})$, where $N_{g(\mathcal{G}')}$ is the number of observations in group g, $g \in \mathcal{G}'$. One then obtains:

$$\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}} = \Delta D^e_{i,E^e_{g(\mathcal{G}')}} - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}} = D^e_{i,E^e_{g(\mathcal{G}')}} - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}$$

and, for all observations i that belongs to group g:

$$\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}} = \Delta D^e_{g(\mathcal{G}'),E^e_{g(\mathcal{G}')}} - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}} = 1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}$$

Using FWL:

$$E[\hat{\beta}^{FD-DiD,e}|\mathbb{D}] = \frac{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}\left[\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}} E[\Delta Y_{i,E^e_{g(\mathcal{G}')}}|\mathbb{D}]\right]}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2}$$

One can develop $E[\Delta Y_{i,E^e_{g(\mathcal{G}')}}|\mathbb{D}]$:[36]

$$
\begin{aligned}
E[\Delta Y_{i,E^e_{g(\mathcal{G}')}}|\mathbb{D}] &= \delta_{E^e_{g(\mathcal{G}')}} - \delta_{E^e_{g(\mathcal{G}')}-1} + \sum_{g'=1}^{G} \left(\mathbb{1}_{\{i \in g'(\mathcal{G})\}} \sum_{e'=1}^{E} \left(\tau_i^{e'} \mathbb{1}_{\{E^{e'}_{g'(\mathcal{G})} = E^e_{g(\mathcal{G}')}\}} \right) \right) \\
&= \delta^0_{E^e_{g(\mathcal{G}')}} + \sum_{g' \in g(\mathcal{G}')} \left(\mathbb{1}_{\{i \in g'(\mathcal{G})\}} \tau_i^e \right)
\end{aligned}
$$

where $g(\mathcal{G}')$ is the group in $\mathcal{G}'$ that includes all groups of $\mathcal{G}$ that have event e occur at time $E^e_{g(\mathcal{G}')}$. The second inequality comes from the fact that only event e for treated group $g(\mathcal{G}')$ can occur at time $E^e_{g(\mathcal{G}')}$. The intercept can then get out of the sum on i and since, for a

[36] With the FD-DiD, the sample is restricted to satisfy a clean control condition. Let $\mathbf{s} = (s_{i,t})_{(i,t) \in \{1,\ldots,N\} \times \{1,\ldots,T\}}$ be a vector of selection indicators with $s_{i,t}$ equal to 1 if cell $\{i,t\}$ is used in the FD-DiD and 0 otherwise. The required strict exogeneity assumption now is: $E[\epsilon_{i,t}|\mathbb{D}, \mathbf{s}] = 0$, $\forall (i,t)$. Since the selection process is made based on the double condition that, at period t, $\Delta D^{e'}_{i,t} \neq 0$ for some units, and $\Delta D^e_{i,t} = 0$, $e = 1, \ldots, E$, for some other units, $\mathbf{s}$ is a deterministic function of $\mathbb{D}$. It implies that $E[\epsilon_{i,t}|\mathbb{D}, \mathbf{s}] = E[\epsilon_{i,t}|\mathbb{D}] = 0$. Hence, the strict exogeneity assumption necessary for the selected sample immediately follows from A3.

given $g(\mathcal{G}')$, $\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}} = 0$, the expression of $E[\hat{\beta}^{FD-DiD,e}]$ becomes:

$$
\begin{aligned}
E[\hat{\beta}^{FD-DiD,e}|\mathbb{D}] &= \frac{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}\left[\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}}\sum_{g'\in g(\mathcal{G}')}\mathbb{1}_{\{i\in g'(\mathcal{G})\}}\tau_i^e\right]}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}\frac{\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}}\sum_{g'\in g(\mathcal{G}')}\mathbb{1}_{\{i\in g'(\mathcal{G})\}}\tau_i^e}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g=1}^{G'}\sum_{g'\in g(\mathcal{G}')}\sum_{i=1}^{N^{g'(\mathcal{G})}_{CCS^e_{g(\mathcal{G}')}}}\frac{\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}}\sum_{g'\in g(\mathcal{G}')}\mathbb{1}_{\{i\in g'(\mathcal{G})\}}\tau_i^e}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g=1}^{G'}\widetilde{\Delta D}^e_{g(\mathcal{G}'),E^e_{g(\mathcal{G}')}}\sum_{g'\in g(\mathcal{G}')}\sum_{i=1}^{N^{g'(\mathcal{G})}_{CCS^e_{g(\mathcal{G}')}}}\frac{\sum_{g'\in g(\mathcal{G}')}\mathbb{1}_{\{i\in g'(\mathcal{G})\}}\tau_i^e}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g=1}^{G'}\widetilde{\Delta D}^e_{g(\mathcal{G}'),E^e_{g(\mathcal{G}')}}\sum_{g'\in g(\mathcal{G}')}\frac{N^{g'(\mathcal{G})}_{CCS^e_{g(\mathcal{G}')}}\tau^e_{g'(\mathcal{G})}}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g=1}^{G'}\widetilde{\Delta D}^e_{g(\mathcal{G}'),E^e_{g(\mathcal{G}')}}\sum_{g'\in g(\mathcal{G}')}\frac{N_{g'}\tau^e_{g'(\mathcal{G})}}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g=1}^{G'}\sum_{g'\in g(\mathcal{G}')}\frac{\widetilde{\Delta D}^e_{g(\mathcal{G}'),E^e_{g(\mathcal{G}')}}N_{g'}\tau^e_{g'(\mathcal{G})}}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g'=1}^{G}\frac{\widetilde{\Delta D}^e_{g'(\mathcal{G}),E^e_{g'(\mathcal{G})}}N_{g'}\tau^e_{g'(\mathcal{G})}}{\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}}(\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \sum_{g'=1}^{G}\omega^{FD-DiD,e}_{g'}\tau^e_{g'(\mathcal{G})}
\end{aligned}
$$

where $N^{g'(\mathcal{G})}_{CCS^e_{g(\mathcal{G}')}}$ is the number of units in subgroup $g'(\mathcal{G})$ of $g(\mathcal{G}')$ such that $g'(\mathcal{G})$ is a treated group, and $\tau^e_{g'(\mathcal{G})} = (1/N^{g'(\mathcal{G})}_{CCS^e_{g(\mathcal{G}')}})\sum_{i=1}^{N^{g'(\mathcal{G})}_{CCS^e_{g(\mathcal{G}')}}}\tau_i^e$. In the eighth equality, $\widetilde{\Delta D}^e_{g'(\mathcal{G}),E^e_{g'(\mathcal{G})}}$ will be the same for two groups $g'(\mathcal{G})$ for which event e occurs at the same period. Eq. (18) follows from the application of the law of iterated expectations to this result.

Moreover, one has:

$$
\begin{aligned}
\omega_{g'}^{FD-DiD,e} &= \frac{N_{g'} \widetilde{\Delta D}^{e}_{g'(\mathcal{G}),E^e_{g'(\mathcal{G})}}}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} (\widetilde{\Delta D}^{e}_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \frac{N_{g'}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \widetilde{\Delta D}^{e}_{i,E^e_{g(\mathcal{G}')}} (\Delta D^e_{i,E^e_{g(\mathcal{G}')}} - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})} \\
&= \frac{N_{g'}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \widetilde{\Delta D}^{e}_{i,E^e_{g(\mathcal{G}')}} \Delta D^e_{i,E^e_{g(\mathcal{G}')}} - \sum_{g=1}^{G'} N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \widetilde{\Delta D}^{e}_{i,E^e_{g(\mathcal{G}')}}} \\
&= \frac{N_{g'}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \widetilde{\Delta D}^{e}_{i,E^e_{g(\mathcal{G}')}} \Delta D^e_{i,E^e_{g(\mathcal{G}')}}} \\
&= \frac{N_{g'}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})}{\sum_{g=1}^{G'} \sum_{i=1:D^e_{i,E^e_{g(\mathcal{G}')}}=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \widetilde{\Delta D}^{e}_{i,E^e_{g(\mathcal{G}')}}} \\
&= \frac{N_{g'}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})}{\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})} \\
&= \frac{N_{CCS^e_{g(\mathcal{G}')}} n^e_{g'} n^e_g n^e_{c,g}}{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} n^e_g n^e_{c,g}}
\end{aligned}
$$

where, in the numerator, $1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}$ is the residual in the auxiliary regression of $\Delta D^e_{i,E^e_{g(\mathcal{G}')}}$ on λ associated with group $g' \in \mathcal{G}$, which is the same for groups $g' \in \mathcal{G}$ that have event e occur in the same period. $n^e_g = N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}$ and $n^e_{c,g} = 1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}$ are the shares of treated units and of control units in $CCS^e_{g(\mathcal{G}')}$, respectively, while $n^e_{g'} = N_{g'}/N_{g(\mathcal{G}')}$ is the share of group $g' \in \mathcal{G}$ in treated units of recomposed group $g \in \mathcal{G}'$. Weights are all positive, proportional to the variance of the treatment dummy, $\widetilde{\Delta D}^{e}_{i,E^e_{g(\mathcal{G}')}}$, on subsample

$CCS^e_{g(\mathcal{G}')}$ and to its size, $N_{CCS^e_{g(\mathcal{G}')}}$. Moreover, weights sum to 1:

$$\begin{aligned}
\sum_{g'=1}^{G} \omega_{g'}^{FD-DiD,e} &= \sum_{g'=1}^{G} \frac{N_{CCS^e_{g(\mathcal{G}')}} n^e_{g'} n^e_g n^e_{c,g}}{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} n^e_g n^e_{c,g}} \\
&= \sum_{g=1}^{G'} \sum_{g' \in \mathcal{G}} \frac{N_{CCS^e_{g(\mathcal{G}')}} n^e_{g'} n^e_g n^e_{c,g}}{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} n^e_g n^e_{c,g}} \\
&= \sum_{g=1}^{G'} \frac{N_{CCS^e_{g(\mathcal{G}')}} n^e_g n^e_{c,g} \sum_{g' \in \mathcal{G}} n^e_{g'}}{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} n^e_g n^e_{c,g}} \\
&= \sum_{g=1}^{G'} \frac{N_{CCS^e_{g(\mathcal{G}')}} n^e_g n^e_{c,g}}{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} n^e_g n^e_{c,g}} \\
&= 1
\end{aligned}$$

so that $E[\beta^{FD-DiD,e}]$ identifies a convex combination of all group-specific effects for event e.

Consider the special case where an event e always occurs at a different date for two different groups. It implies that each observation satisfying restrictions associated with Eq. (17) enters into one and only one CCS. This case leads to the simplified result:

$$\begin{aligned}
E[\hat{\beta}^{FD-DiD,e}|\mathbb{D}] &= \sum_{g=1}^{G} \frac{N_g \widetilde{\Delta D}^e_{g,E^e_g}}{\sum_{g=1}^{G} \sum_{i=1}^{N_{CCS^e_g}} (\widetilde{\Delta D}^e_{i,E^e_g})^2} \tau^e_g \\
&= \sum_{g=1}^{G} \omega_g^{FD-DiD,e} \tau^e_g
\end{aligned}$$

with:

$$\begin{aligned}
\omega_g^{FD-DiD,e} &= \frac{N_g(1 - N_g/N_{CCS^e_g})}{\sum_{g=1}^{G} N_g(1 - N_g/N_{CCS^e_g})} \\
&= \frac{N_{CCS^e_g} n^e_g n^e_{c,g}}{\sum_{g=1}^{G} N_{CCS^e_g} n^e_g n^e_{c,g}}
\end{aligned}$$

Weights are positive and sum to 1. Footnote 17 follows from application of the law of iterated expectations to the result above.

C Decompositions

The FD estimator with one event, $\hat{\beta}^{FD,e}$, can be decomposed to show that it leverages comparisons of treated units with units in the same period that may have been through other events:

$$\begin{aligned}
\hat{\beta}^{FD,e} &= \frac{\sum_{i,t} \widetilde{\Delta D}^{e}_{i,t} \Delta Y_{i,t}}{\sum_{i,t} (\widetilde{\Delta D}^{e}_{i,t})^2} \\
&= \frac{\sum_{i,t:\Delta D^e_{i,t}=1} (1-\Delta D^e_{.,t}) \Delta Y_{i,t} - \sum_{i,t:\Delta D^e_{i,t}=0} \Delta D^e_{.,t} \Delta Y_{i,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g (1-\Delta D^e_{.,t})} \\
&= \sum_t \frac{(1-N^e_t/N) \sum_{i:\Delta D^e_{i,t}=1} \Delta Y_{i,t} - (N^e_t/N) \sum_{i:\Delta D^e_{i,t}=0} \Delta Y_{i,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g (1-\Delta D^e_{.,t})} \\
&= \sum_t \frac{(1-N^e_t/N) \sum_{i=1}^{N^e_t} \Delta Y_{i,t} - (N^e_t/N) \sum_{i=1}^{N^0_t} \Delta Y_{i,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g (1-\Delta D^e_{.,t})} \\
&= \sum_t (1-N^e_t/N) \sum_{i=1}^{N^e_t} \frac{\Delta Y_{i,t} - (1/N^0_t) \sum_{i=1}^{N^0_t} \Delta Y_{i,t}}{\sum_{g,t:\Delta D^e_{g,t}=1} N_g (1-\Delta D^e_{.,t})}
\end{aligned}$$

where N^e_t and N^0_t denote the number of units that go through event e in period t and the number of units that don't, respectively.

The FD-DiD estimator can be developed to show that it leverages comparisons of the outcome of every unit treated for event e with every control unit untreated at the same period:

$$
\begin{aligned}
\hat{\beta}^{FD-DiD,e} &= \frac{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} \widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}} \Delta Y_{i,E^e_{g(\mathcal{G}')}}}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} (\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \frac{\sum_{g=1}^{G'} \left((1 - \Delta D^e_{.,E^e_{g(\mathcal{G}')}}) \sum_{i=1}^{N_{g(\mathcal{G}')}} \Delta Y_{i,E^e_{g(\mathcal{G}')}} - \Delta D^e_{.,E^e_{g(\mathcal{G}')}} \sum_{i=1}^{N^0_{g(\mathcal{G}')}} \Delta Y_{i,E^e_{g(\mathcal{G}')}} \right)}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} (\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \frac{\sum_{g=1}^{G'} \left((1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}) \sum_{i=1}^{N_{g(\mathcal{G}')}} \Delta Y_{i,E^e_{g(\mathcal{G}')}} - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}} \sum_{i=1}^{N^0_{g(\mathcal{G}')}} \Delta Y_{i,E^e_{g(\mathcal{G}')}} \right)}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} (\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \frac{\sum_{g=1}^{G'} (1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}) \sum_{i=1}^{N_{g(\mathcal{G}')}} \left(\Delta Y_{i,E^e_{g(\mathcal{G}')}} - (1/N^0_{g(\mathcal{G}')}) \sum_{j=1}^{N^0_{g(\mathcal{G}')}} \Delta Y_{j,E^e_{g(\mathcal{G}')}} \right)}{\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} (\widetilde{\Delta D}^e_{i,E^e_{g(\mathcal{G}')}})^2} \\
&= \frac{\sum_{g=1}^{G'} (1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}) \sum_{i=1}^{N_{g(\mathcal{G}')}} \left(\Delta Y_{i,E^e_{g(\mathcal{G}')}} - (1/N^0_{g(\mathcal{G}')}) \sum_{j=1}^{N^0_{g(\mathcal{G}')}} \Delta Y_{j,E^e_{g(\mathcal{G}')}} \right)}{\sum_{g=1}^{G'} N_{g(\mathcal{G}')} (1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})}
\end{aligned}
$$

D Efficiency

Proof of Theorem 2. Under unrestricted treatment effect heterogeneity, a well-defined model is: $Y_{i,t} = \alpha + \alpha_i + \delta_t + \sum_{e=1}^{E} D^e_{i,t} \tau^e_i + u_{i,t}$. Taking the first difference gives:

$$\Delta Y_{i,t} = \delta^0_t + \sum_{e=1}^{E} \Delta D^e_{i,t} \tau^e_i + \epsilon_{i,t} \tag{24}$$

where $\epsilon_{i,t} = \Delta u_t$. Recall that under A4 the estimand of interest is: $\tau^e = \sum_i \omega_i \tau^e_i$, with $\tau^e_i = E[Y^e_{i,E^e_i} - Y^{e-1}_{i,E^e_i} | \mathbb{D}]$. The proof is divided in four steps. First, an efficient estimator for $\tau = (\tau^e_i)_{i,e}$ is identified using the law of total variance. Second, an efficient estimator is then straightforwardly derived for τ^e. Third, I show that this estimator has an imputation form. Finally, I demonstrate that this imputation estimator is identical to the FD-DiD.

Step 1. Let $\hat{\tau}^*$ denote the OLS estimator of the vector composed of all $\{\hat{\tau}^{e*}_i\}_{i,e}$ in the regression of $\Delta Y_{i,t}$ on time fixed effects and individual treatment indicators (Eq. (24)). By the law of total variance: $\text{Var}(\hat{\tau}^*) = \text{Var}(E[\hat{\tau}^*|\mathbb{D}]) + E[\text{Var}(\hat{\tau}^*|\mathbb{D})]$. $\hat{\tau}^*$ is conditionally unbiased for any treatment assignment $\mathbb{D}$ so that $\text{Var}(\hat{\tau}^*) = E[\text{Var}(\hat{\tau}^*|\mathbb{D})]$. Let $\hat{\tau}$ be any other linear unbiased estimator of τ. Conditional on information $\mathbb{D}$, the Gauss-Markov theorem implies that $\text{Var}(\hat{\tau}^*|\mathbb{D})$ will be lower than $\text{Var}(\hat{\tau}|\mathbb{D})$. Averaging over $\mathbb{D}$, $\hat{\tau}^*$ is best linear unbiased for τ.

Step 2. Define $\hat{\tau}^{e*} = \sum_i \omega_i \hat{\tau}_i^{e*}$. Here, I extend the efficiency of $\hat{\tau}^*$ for τ to $\hat{\tau}^{e*}$ for τ^e. For every linear estimator $\hat{\tau}^e$ unbiased for τ^e for all τ, there is a linear unbiased estimator $\hat{\tau}$ of τ for which $\hat{\tau}^e = \omega'\hat{\tau}$, where ω is the vector of relevant weights ω_i for τ_i^e. Conditional on information set $\mathbb{D}$, $\hat{\tau}^*$ is best linear unbiased for τ with variance $\Sigma_{\hat{\tau}^*}$ that is minimal among the variances of linear unbiased estimators of τ. Hence, $\text{Var}(\omega'\hat{\tau}^*|\mathbb{D}) - \text{Var}(\omega'\hat{\tau}|\mathbb{D}) = \omega'(\Sigma_{\hat{\tau}^*} - \Sigma_{\hat{\tau}})\omega \leq 0$ establishes efficiency conditional on the treatment design. Using the law of total variance again: $\text{Var}(\hat{\tau}^{e*}) = \text{Var}(E[\hat{\tau}^{e*}|\mathbb{D}]) + E[\text{Var}(\hat{\tau}^{e*}|\mathbb{D})]$ where the first term is null and the second averages estimator variances that are minimal among linear unbiased estimators conditional on $\mathbb{D}$.

Step 3. In matrix form, Eq. (24) is equivalent to $\mathbf{\Delta Y} = \lambda\delta + \mathbf{\Delta D}\tau + \epsilon$. λ is a matrix of T time indicators. δ is a vector of T coefficients. $\mathbf{\Delta D}$ is a $(NT \times N_1)$ matrix, where N_1 is the number of (i, t) cells in which an event occurs, i.e., a cell such that $\Delta D_{i,t}^e = 1$, for some $e \in \{1, \ldots, E\}$. τ is a vector of coefficients of size N_1. Using FWL, the OLS estimator of δ can be obtained from the regression of the residuals of $\mathbf{\Delta Y}$ on the residuals of λ with respect to $\mathbf{\Delta D}$:

$$\begin{aligned}\hat{\delta}^* &= (\lambda'(\mathbb{I}_{NT} - \mathbf{\Delta D}(\mathbf{\Delta D'\Delta D})^{-1}\mathbf{\Delta D'})\lambda)^{-1}\lambda'(\mathbb{I}_{NT} - \mathbf{\Delta D}(\mathbf{\Delta D'\Delta D})^{-1}\mathbf{\Delta D'})\mathbf{\Delta Y} \\ &= (\lambda_\mathbf{0}'\lambda_\mathbf{0})^{-1}\lambda_\mathbf{0}'\mathbf{\Delta Y_0}\end{aligned}$$

where $\lambda_\mathbf{0}$ and $\mathbf{\Delta Y_0}$ are the $(N_0 \times T)$ matrix of time indicators and $(N_0 \times 1)$ vector of outcomes for the restricted sample of observations such that no event occurred in cells (i, t), i.e., (i, t) cells such that $\Delta D_{i,t}^e = 0$, for all $e \in \{1, \ldots, E\}$.

The OLS estimator $\hat{\tau}^*$ of τ in $\mathbf{\Delta Y} = \lambda\delta + \mathbf{\Delta D}\tau + \epsilon$ is the same as the OLS estimator in the regression of $\mathbf{\Delta Y} - \lambda\hat{\delta}^*$ on $\mathbf{\Delta D}$. Indeed, $(\hat{\delta}^*, \hat{\tau}^*)$ minimizes the sum of squares $||\mathbf{\Delta Y} -$

$\lambda\hat{\delta} - \mathbf{\Delta D}\hat{\tau}||^2$ over choices $(\hat{\delta}, \hat{\tau})$ so that $\hat{\tau}^*$ minimizes $||\mathbf{\Delta Y} - \lambda\hat{\delta}^* - \mathbf{\Delta D}\hat{\tau}||^2$ over $\hat{\tau}$ given $\hat{\delta}^*$. One gets:

$$\begin{aligned}\hat{\tau}^* &= (\mathbf{\Delta D}'\mathbf{\Delta D})^{-1}\mathbf{\Delta D}'(\mathbf{\Delta Y} - \lambda\hat{\delta}^*) \\ &= \mathbf{\Delta D}'\mathbf{\Delta Y} - \mathbf{\Delta D}'\lambda\hat{\delta}^* \\ &= \mathbf{\Delta Y_1} - \lambda_{\mathbf{1}}\hat{\delta}^*\end{aligned}$$

where $\lambda_{\mathbf{1}}$ and $\mathbf{\Delta Y_1}$ are the $(N_1 \times T)$ matrix of time indicators and $(N_1 \times 1)$ vector of outcomes, respectively, for the restricted sample of observations such that an event occurred in cells (i, t), i.e., (i, t) cells such that $\Delta D^e_{i,t} = 1$, for some $e \in \{1, \ldots, E\}$.

It implies that the efficient estimator of τ^e can be obtained by imputation with the following procedure:

1. Estimate $\hat{\delta}^*_t$ in the regression $\Delta Y_{i,t} = \delta^0_t + \epsilon_{i,t}$ on the set of (i, t) cells such that no event occurs.

2. Estimate the counterfactual outcomes for treated units as $\widehat{\Delta Y^0_{i,t}} = \hat{\delta}^*_t$.

3. Estimate individual event effects as $\hat{\tau}^{e*}_i = \Delta Y_{i,t} - \widehat{\Delta Y^0_{i,t}}$.

4. Compute $\hat{\tau}^{e*} = \sum_i \omega_i \hat{\tau}^{e*}_i$.

Step 4. It remains to show that the FD-DiD estimator is the same as this imputation estimator. First, note that the FD-DiD estimator has the following form:[37]

$$\hat{\beta}^{FD-DiD,e} = \frac{\sum_{g=1}^{G'}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})\sum_{i=1}^{N_{g(\mathcal{G}')}}\left(\Delta Y_{i,E^e_{g(\mathcal{G}')}} - (1/N^0_{g(\mathcal{G}')})\sum_{j=1}^{N^0_{g(\mathcal{G}')}}\Delta Y_{j,E^e_{g(\mathcal{G}')}}\right)}{\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})}$$

This result highlights that $\hat{\beta}^{FD-DiD,e}$ leverages comparisons of the outcome evolution of every unit treated for event e with every control unit untreated at the same period. $\hat{\beta}^{FD-DiD,e}$ is a subgroup difference-in-differences (SGDD) estimator in the terminology of Harmon (2023),

[37] This result is shown in Section C of the Appendix.

i.e., the last untreated period is used as the baseline. It has the form: $\hat{\beta}^{FD-DiD,e} = \sum_i \omega_i \hat{\tau}_i^e$, where the sum is taken over all individuals treated for event e, where weights $\omega_i = (1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})/(\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}))$ are identical for all individuals in the same group and are considered conditional on the treatment design, and $\hat{\tau}_i^e = \Delta Y_{i,E^e_{g(\mathcal{G}')}} - (1/N^0_{g(\mathcal{G}')}) \sum_{j=1}^{N^0_{g(\mathcal{G}')}} \Delta Y_{j,E^e_{g(\mathcal{G}')}}$.

$\hat{\delta}_t^*$ obtained in step 3 is simply the average of observations $\Delta Y_{i,t}$ at time t across untreated units i, that is $(1/N^0_{g(\mathcal{G}')}) \sum_{j=1}^{N^0_{g(\mathcal{G}')}} \Delta Y_{j,t}$. Plugging this estimate into $\hat{\tau}_i^{e*}$ gives:

$$\hat{\tau}_i^{e*} = \Delta Y_{i,E^e_{g(\mathcal{G}')}} - (1/N^0_{g(\mathcal{G}')}) \sum_{j=1}^{N^0_{g(\mathcal{G}')}} \Delta Y_{j,E^e_{g(\mathcal{G}')}}$$

Finally, with the vector of individual treatment estimators $(\hat{\tau}_i^{e*})_{i,e}$, the estimator of the target estimand for a given e is $\hat{\tau}^{e*} = \sum_i \omega_i \hat{\tau}_i^{e*}$, with $\omega_i = (1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})/(\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}))$, i.e., the FD-DiD estimator: $\hat{\tau}^{e*} = \hat{\beta}^{FD-DiD,e}$.

E Asymptotic properties

Proof of Theorem 3. First, from the decomposition of $\hat{\beta}^{FD-DiD,e}$ in Section C of the Appendix, it is straightforward to show that the FD-DiD has the following representation:

$$\hat{\beta}^{FD-DiD,e} = \sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} v_{i,E^e_i} Y_{i,E^e_i}$$

Where the sum is taken over the restricted set of observations that satisfy Eq. (15), and with:

$$v_{i,E^e_i} = \begin{cases} \omega_{i(g)} \quad \text{if } i \in g(\mathcal{G}') \\ (-N_{g(\mathcal{G}')}/N^0_{g(\mathcal{G}')})\omega_{i(g)} \quad \text{if } i \in CCS^e_{g(\mathcal{G}')} \text{ and } i \notin g(\mathcal{G}') \end{cases}$$

To prove consistency of $\hat{\beta}^{FD-DiD,e}$, it is sufficient that $E[\hat{\beta}^{FD-DiD,e}] = \tau_e$ - which has been proven in Section 4 - and that $\text{Var}(\hat{\beta}^{FD-DiD,e}) \to 0$. One has:

$$\begin{aligned}
\text{Var}(\hat{\beta}^{FD-DiD,e}|\mathbb{D}) &= \text{Var}\left(\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} v_{i,E^e_i} Y_{i,E^e_i} |\mathbb{D}\right) \\
&= \text{Var}\left(\sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} v_{i,E^e_i} \epsilon_{i,E^e_i} |\mathbb{D}\right) \\
&= \text{Var}\left(\sum_{i\in\Omega_N} \sum_{g=1}^{G'} v_{i,E^e_{i(g)}} \epsilon_{i,E^e_{i(g)}} |\mathbb{D}\right) \\
&= \sum_{i\in\Omega_N} \text{Var}\left(\sum_{g=1}^{G'} v_{i,E^e_{i(g)}} \epsilon_{i,E^e_{i(g)}} |\mathbb{D}\right) \\
&\leq \sum_{i\in\Omega_N} \left(\sum_{g=1}^{G'} |v_{i,E^e_g}|\right)^2 \bar{\sigma}^2
\end{aligned}$$

Where I use A5' in the fourth equality and the inequality. Moreover:

$$\begin{aligned}
\sum_{i\in\Omega_N} \left(\sum_{g=1}^{G'} |v_{i,E^e_g}|\right)^2 &\leq 2\sum_{i\in\Omega_N} \left(\sum_{g=1:\Delta D^e_{i,E^e_g}=0}^{G'} |v_{i,E^e_g}|\right)^2 + 2\sum_{i\in\Omega_N} \left(\sum_{g=1:\Delta D^e_{i,E^e_g}=1}^{G'} |v_{i,E^e_g}|\right)^2 \\
&\leq 2G' \sum_{i\in\Omega_N} \sum_{g=1:\Delta D^e_{i,E^e_g}=0}^{G'} v^2_{i,E^e_g} + 2\sum_{i\in\Omega_N} \left(\sum_{g=1:\Delta D^e_{i,E^e_g}=1}^{G'} |w_{i(g)}|\right)^2 \\
&\leq 2G' \sum_{g=1}^{G'} \sum_{i=1}^{N^0_g} v^2_{i,E^e_g} + 2\sum_{g=1}^{G'} \sum_{i=1}^{N_g} w^2_i \\
&= 2G' \sum_{g=1}^{G'} (N_g/N^0_g) N_g w^2_g + 2\sum_{g=1}^{G'} N_g w^2_g
\end{aligned}$$

Where I use Jensen's inequality in the first two steps: $\forall b \geq 1$, $S \in \mathbb{N}$, $\forall(a_1, \ldots, a_s)$, $|\sum_{s=1}^S a_s|^b \leq S^{b-1} \sum_{s=1}^S |a_s|^b$. It is straightforward to show that, with $N_g/N_{CCS^e_{g(\mathcal{G}')}}$ bounded between 0 and 1 and with $G' \in \mathbb{N}^*$, $G' \sum_{g=1}^{G'} (N_g/N^0_g) N_g w^2_g \to 0 \Rightarrow \sum_{g=1}^{G'} N_g w^2_g \to 0$. Therefore, it is sufficient to focus on convergence of the first sum.

Note that $N_g/N^0_g = (N_g/N_{CCS^e_{g(\mathcal{G}')}})(N_{CCS^e_{g(\mathcal{G}')}}/N^0_g)$ with $N_g/N_{CCS^e_{g(\mathcal{G}')}}$ bounded between 0 and 1 and with $N_{CCS^e_{g(\mathcal{G}')}}/N^0_g < \infty$. Hence, $2G' \sum_{g=1}^{G'} (N_g/N^0_g) N_g w^2_g$ goes to 0 if $G' N_g w^2_g$ goes to 0 for all g. Moreover:

$$\begin{aligned} G' N_g w_g^2 &= G' N_g \left(\frac{N_g^0 / N_{CCS^e_{g(\mathcal{G}')}}}{\sum_{g=1}^{G'} N_g (N_g^0 / N_{CCS^e_{g(\mathcal{G}')}})} \right)^2 \\ &\leq \frac{G' N_g}{\left(\sum_{g=1}^{G'} N_g N_g^0 / N_{CCS^e_{g(\mathcal{G}')}} \right)^2} \end{aligned}$$

Therefore, using the assumption of Proposition 1: $\text{Var}(\hat{\beta}^{FD-DiD,e}|\mathbb{D}) \to 0$. Moreover, by the law of total variance:

$$\text{Var}(\hat{\beta}^{FD-DiD,e}) = \text{Var}(E[\hat{\beta}^{FD-DiD,e}|\mathbb{D}]) + E[\text{Var}(\hat{\beta}^{FD-DiD,e}|\mathbb{D})] = E[\text{Var}(\hat{\beta}^{FD-DiD,e}|\mathbb{D})] \to 0$$

Hence, $\hat{\beta}^{FD-DiD,e}$ converges in quadratic mean. The proof of Theorem 3 is completed.

Proof of Theorem 4. One has:

$$\hat{\beta}^{FD-DiD,e} - \tau^e = \sum_{g=1}^{G'} \sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} v_{i,E_i^e} \epsilon_{i,E_i^e} = \sum_{i \in \Omega_N} \xi_i$$

with $\xi_i = v_i' \epsilon_i = \sum_{t:(i,t)\in\Omega} v_{i,t}\epsilon_{i,t}$, $E[\xi_i|\mathbb{D}] = 0$, $\text{Var}(\xi_i|\mathbb{D}) = v_i'\Sigma_i v_i = (\sum_{t:(i,t)\in\Omega} v_{i,t}\epsilon_{i,t})^2$. Write $p = 2+\kappa$ and let q be the solution to $1/p+1/q = 1$ (so $1 < q < 2 < p$). Using Hölder's inequality to establish: $\sum_{t:(i,t)\in\Omega} |v_{i,t}|^{1/q} \left(|v_{i,t}|^{1/p}|\epsilon_{i,t}|\right) \leq \left(\sum_{t:(i,t)\in\Omega} |v_{i,t}|^{q/q}\right)^{1/q} \left(\sum_{t:(i,t)\in\Omega} |v_{i,t}|^{p/p}|\epsilon_{i,t}|^p\right)^{1/p}$,

$\forall i$, and using $E[|\epsilon_{i,t}|^p|\mathbb{D}] \leq C$ and $p/q+1=p$, one has:

$$\begin{aligned}
E[|\xi_i|^{2+\kappa}|\mathbb{D}] &= E\left[|\sum_{t:(i,t)\in\Omega} v_{i,t}\epsilon_{i,t}|^p|\mathbb{D}\right] \\
&\leq E\left[\left(\sum_{t:(i,t)\in\Omega} |v_{i,t}\epsilon_{i,t}|\right)^p |\mathbb{D}\right] \\
&= E\left[\left(\sum_{t:(i,t)\in\Omega} |v_{i,t}|^{1/q}|v_{i,t}|^{1/p}|\epsilon_{i,t}|\right)^p |\mathbb{D}\right] \\
&\leq \left(\sum_{t:(i,t)\in\Omega} |v_{i,t}|\right)^{p/q} \sum_{t:(i,t)\in\Omega} |v_{i,t}| E\left[|\epsilon_{i,t}|^p|\mathbb{D}\right] \\
&\leq \left(\sum_{t:(i,t)\in\Omega} |v_{i,t}|\right)^{p/q+1} C \\
&= \left(\sum_{t:(i,t)\in\Omega} |v_{i,t}|\right)^{p} C
\end{aligned}$$

Then:

$$\begin{aligned}
\sum_{i\in\Omega_N} E[|\xi_i|^{2+\kappa}|\mathbb{D}] \left(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right)^{(2+\kappa)/2} &\leq \sum_{i\in\Omega_N} \left(\sum_{t:(i,t)\in\Omega} |v_{i,t}|\right)^{2+\kappa} \left(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right)^{(2+\kappa)/2} C \\
&= \sum_{i\in\Omega_N} \left(\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}} \sum_{t:(i,t)\in\Omega} |v_{i,t}|\right)^{2+\kappa} C
\end{aligned}$$

Since, by assumption, $\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}} / \left(\sum_{g=1}^{G'} N_g N_g^0 / N_{CCS^e_{g(\mathcal{G}')}}\right) \to 0$, one also has that $\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}} |v_{i,t}| \to 0$ for all weight $v_{i,t}$. It follows that:

$$\sum_{i\in\Omega_N} E[|\xi_i|^{2+\kappa}|\mathbb{D}] \left(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right)^{(2+\kappa)/2} \to 0$$

Moreover, the assumption $\sigma_e^2 \sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} > 0$ is equivalent to $1/\left(\sigma_e^2 \sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right) <$

∞, which implies $1/\left(\sigma_e^{2+\kappa}(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}})^{(2+\kappa)/2}\right) < \infty$. Hence:

$$\frac{\sum_{i\in\Omega_N} E[|\xi_i|^{2+\kappa}]}{\sigma_e^{2+\kappa}} = \frac{\sum_{i\in\Omega_N} E[|\xi_i|^{2+\kappa}](\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}})^{(2+\kappa)/2}}{\sigma_e^{2+\kappa}\left(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right)^{(2+\kappa)/2}} \to 0$$

Then, by the Lyapunov central limit theorem: $\sigma_e^{-1}(\hat{\beta}^{FD-DiD,e} - \tau^e) \xrightarrow{d} \mathcal{N}(0,1)$. The proof of Proposition 4 is completed.

F Asymptotic properties with errors clustered at the CCS level

Assumption 5". *(Clustered errors at the clean control sample level)*
Error terms $\epsilon_{i,t}$ are independent across periods t and have bounded variance $Var(\epsilon_{i,t}|\mathbb{D}) \leq \bar{\sigma}^2$ for all $(i,t) \in \Omega_N \times \Omega_T$ uniformly.

$N_g/N_{CCS^e_{g(\mathcal{G}')}}$ is assumed to be bounded away from 0 and 1.

Theorem 3'. *Denote $\omega_g = (1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}})/(\sum_{g=1}^{G'} N_{g(\mathcal{G}')}(1 - N_{g(\mathcal{G}')}/N_{CCS^e_{g(\mathcal{G}')}}))$. Assume that A1-A4 and A5' hold and that, $\forall g \in \mathcal{G}'$, $(N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}})^2/(\sum_{g=1}^{G'} N_g N_g^0/N_{CCS^e_{g(\mathcal{G}')}})^2 \to 0$. Then: $\hat{\beta}^{FD-DiD,e} - \tau^e \xrightarrow{L_2} 0$.*

Proof of Proposition 3'.

$$\begin{aligned}\text{Var}(\hat{\beta}^{FD-DiD,e}|\mathbb{D}) &= \text{Var}\left(\sum_{g=1}^{G'}\sum_{i=1}^{N_{CCS^e_{g(\mathcal{G}')}}} v_{i,E_i^e}\epsilon_{i,E_i^e}|\mathbb{D}\right)\\ &\leq \sum_{g=1}^{G'}\left(\sum_{i:(i,t)\in\Omega} |v_{i,E_g^e}|\right)^2 \bar{\sigma}^2\end{aligned}$$

Moreover:

$$
\begin{aligned}
\sum_{g=1}^{G'}\left(\sum_{i:(i,t)\in\Omega}|v_{i,E_g^e}|\right)^2 &\leq 2\sum_{g=1}^{G'}\left(\sum_{i:(i,t)\in\Omega,\Delta D^e_{i,E_g^e}=0}|v_{i,E_g^e}|\right)^2+2\sum_{g=1}^{G'}\left(\sum_{i:(i,t)\in\Omega,\Delta D^e_{i,E_g^e}=1}|v_{i,E_g^e}|\right)^2\\
&\leq 2\sum_{g=1}^{G'}N_g^0\sum_{i:(i,t)\in\Omega,\Delta D^e_{i,E_g^e}=0}v^2_{i,E_g^e}+2\sum_{g=1}^{G'}N_g\sum_{i:(i,t)\in\Omega,\Delta D^e_{i,E_g^e}=1}v^2_{i,E_g^e}\\
&=2\sum_{g=1}^{G'}(N_g^0)^2\left(\omega_g N_g/N_g^0\right)^2+2\sum_{g=1}^{G'}(N_g\omega_g)^2\\
&=4\sum_{g=1}^{G'}\frac{(N_gN_g^0/N_{CCS^e_{g(\mathcal{G}')}})^2}{(\sum_{g=1}^{G'}N_gN_g^0/N_{CCS^e_{g(\mathcal{G}')}})^2}
\end{aligned}
$$

Therefore, using the assumption of Proposition 1': $\text{Var}(\hat{\beta}^{FD-DiD,e}|\mathbb{D})\to 0$, and by the law of total variance: $\text{Var}(\hat{\beta}^{FD-DiD,e})\to 0$. $\hat{\beta}^{FD-DiD,e}$ converges in quadratic mean, and the proof of Proposition 1' is completed.

Proposition 4'. *Under A1-A4 and A5', if there exists $\kappa>0$ such that $E[|\epsilon_{i,t}|^{2+\kappa}|\mathbb{D}]$ is uniformly bounded, that, $\forall g\in\mathcal{G}'$, $\sqrt{\sum_{g=1}^{G'}N_{CCS^e_{g(\mathcal{G}')}}}N_gN_g^0/N_{CCS^e_{g(\mathcal{G}')}}/\left(\sum_{g=1}^{G'}N_gN_g^0/N_{CCS^e_{g(\mathcal{G}')}}\right)\to 0$, and that: $\sigma_e^2\sum_{g=1}^{G'}N_{CCS^e_{g(\mathcal{G}')}}>0$ with $\sigma_e^2=Var(\hat{\beta}^{FD-DiD,e})$, then: $\sigma_e^{-1}(\hat{\beta}^{FD-DiD,e}-\tau^e)\xrightarrow{d}\mathcal{N}(0,1)$.*

Proof of Proposition 4. One has:

$$\hat{\beta}^{FD-DiD,e}-\tau^e=\sum_{g=1}^{G'}\xi_g$$

with $\xi_g=\sum_{i:(i,t)\in\Omega}v_{i,t}\epsilon_{i,t}$, $E[\xi_g|\mathbb{D}]=0$, $\text{Var}(\xi_g|\mathbb{D})=(\sum_{i:(i,t)\in\Omega}v_{i,t}\epsilon_{i,t})^2$. Write $p=2+\kappa$ and let q be the solution to $1/p+1/q=1$ (so $1<q<2<p$). Using Hölder's inequality, $E[|\epsilon_{i,t}|^p|\mathbb{D}]\leq C$ and $p/q+1=p$, one gets:

$$E[|\xi_g|^{2+\kappa}|\mathbb{D}]\leq\left(\sum_{t:(i,t)\in\Omega}|v_{i,t}|\right)^p C$$

Then:

$$\sum_{g=1}^{G'} E[|\xi_g|^{2+\kappa}|\mathbb{D}]\left(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right)^{(2+\kappa)/2} \leq \sum_{g=1}^{G'}\left(\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}}\sum_{i:(i,t)\in\Omega}|v_{i,t}|\right)^{2+\kappa} C$$

Moreover:

$$\begin{aligned}\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}}\sum_{i:(i,t)\in\Omega}|v_{i,t}| &= \sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}}\left(\sum_{i:(i,t)\in\Omega,\Delta D^e_{i,t}=0}|v_{i,t}| + \sum_{i:(i,t)\in\Omega,\Delta D^e_{i,t}=1}|v_{i,t}|\right)\\ &= 2\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} N_g\omega_g}\\ &= 2\frac{\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}} N_g N^0_g/N_{CCS^e_{g(\mathcal{G}')}}}{\sum_{g=1}^{G'} N_g N^0_g/N_{CCS^e_{g(\mathcal{G}')}}}\end{aligned}$$

Since, by assumption, $\forall g \in \mathcal{G}'$, $\sqrt{\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}} N_g N^0_g/N_{CCS^e_{g(\mathcal{G}')}}/\left(\sum_{g=1}^{G'} N_g N^0_g/N_{CCS^e_{g(\mathcal{G}')}}\right) \to 0$, it follows that:

$$\sum_{g=1}^{G'} E[|\xi_g|^{2+\kappa}|\mathbb{D}]\left(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right)^{(2+\kappa)/2} \to 0$$

Moreover, the assumption $\sigma^2_e\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}} > 0$ is equivalent to $1/\left(\sigma^2_e\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right) < \infty$, which implies $1/\left(\sigma^{2+\kappa}_e(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}})^{(2+\kappa)/2}\right) < \infty$. Hence:

$$\frac{\sum_{g=1}^{G'} E[|\xi_g|^{2+\kappa}]}{\sigma^{2+\kappa}_e} = \frac{\sum_{g=1}^{G'} E[|\xi_g|^{2+\kappa}](\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}})^{(2+\kappa)/2}}{\sigma^{2+\kappa}_e\left(\sum_{g=1}^{G'} N_{CCS^e_{g(\mathcal{G}')}}\right)^{(2+\kappa)/2}} \to 0$$

Then, by the Lyapunov central limit theorem: $\sigma^{-1}_e(\hat{\beta}^{FD-DiD,e} - \tau^e) \xrightarrow{d} \mathcal{N}(0,1)$, and the proof of Proposition 2 is completed.