Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects

Sarah Abraham and Liyang Sun*

April 17, 2018

Abstract

Event studies are frequently used to estimate average treatment effects on the treated (ATT). In estimating the ATT, researchers commonly use fixed effects models that implicitly assume constant treatment effects across cohorts. We show that this is not an innocuous assumption. In fixed effect models where the sole regressor is treatment status, the OLS coefficient is a non-convex average of the heterogeneous cohort-specific ATTs. When regressors containing lags and leads of treatment are added, the OLS coefficient corresponding to a given lead or lag picks up spurious terms consisting of treatment effects from other periods. Therefore, estimates from these commonly used models are not causally interpretable. We propose alternative estimators that identify certain convex averages of the cohort-specific ATTs, hence allowing for causal interpretation even under heterogeneous treatment effects. To illustrate the empirical content of our results, we show that the fixed effects estimators and our proposed estimators differ substantially in an application to the economic consequences of hospitalization.

1 Introduction

Rich panel data has encouraged a growing literature of empirical work using difference-in-differences (DID) designs to estimate the average treatment effects among units that receive the treatment (ATT) in the treatment period. The key identifying assumption is that in the absence of treatment, treated units would have experienced the same trends in average outcomes as the control units (i.e. a parallel trends assumption). An extension to DID, frequently called an "event study", exploits variation in the timing of a unit's treatment to estimate the ATT. Examples of treatments include adverse health shocks to individuals, and policies being rolled out to different states.

*Email lsun20@mit.edu. We are grateful to Isaiah Andrews, Amy Finkelstein, Anna Mikusheva, and Heidi Williams for their guidance and support. We thank Alberto Abadie, Peter Hull, Kevin Kainan Li, Whitney Newey, and James Poterba for helpful comments. This research was supported by the National Institute on Aging, Grant Number T32-AG000186. This draft is a work in progress and should not be cited without prior approval. Replication code is available on `http://economics.mit.edu/grad/lsun20/` Comments are welcome; all errors are our own.

Event study differs from DID design in that treatment is no longer uniquely characterized by a binary variable. In event studies (where treatment is an absorbing state), treatment is uniquely characterized by the timing of treatment, and we can categorize treated units into cohorts based on when they first receive the treatment. For each cohort, we can define relative time as the time relative to the initial treatment. Within a cohort, calendar time and relative time are collinear, so we cannot identify dynamic treatment effects in the presence of time trends. However, with multiple cohorts, as in DID, we can separate these two effects by comparing the trends in average outcomes between treated cohorts and cohorts yet to be treated (given a parallel trends assumption). As a result, we can identify causal treatment effects for a given cohort at many different relative times, provided that this cohort is treated for more than one period. We call these cohort-specific average treatment effects on the treated $CATT$s. Researchers are often interested in some weighted averages of $CATT$s, and their estimates are typically obtained from a linear two-way fixed effects (FE) regression, with both unit fixed effects and time fixed effects. To estimate dynamic treatment effects, researchers add lags of treatment to the regression and interpret the coefficient on a specific lag as a weighted average of $CATT$s from the given lag. To gauge whether the parallel trends assumption is plausible, researchers add leads of treatment to the regression and take the coefficients on these leads as evidence for lack of or existence of pre-trends.[1]

In this paper, we make the following three contributions. First, we cast event studies in a potential outcome framework to discipline possible forms of heterogeneity in $CATT$s. Second, we examine the behavior of the linear two-way fixed effects regressions for event studies when $CATT$s are allowed to differ across cohorts. We show that the FE estimators can be non-convex averages of $CATT$s, and thus not causally interpretable. Lastly, we propose alternative estimators that are guaranteed to estimate convex averages of $CATT$s. We close with an empirical example that demonstrates both how and why FE estimators can be misleading in comparison to our suggested alternative estimators.

We begin by providing a potential outcome framework for event studies, and clarify the notion of treatment effects heterogeneity in a dynamic setting, which is specific to event studies. We define the underlying causal parameter of interest as the cohort-specific average treatment effect on the treated ($CATT$). Since cohorts are the level used to identify time trends separately from dynamic treatment effects, they are also the key level of concern for confounding heterogeneity. We can categorize a single cohort's effects as either static or dynamic and the overall effects as either stationary or non-stationary. A cohort has "static" treatment effects when the $CATT$ is constant over time; in contrast, it is "dynamic" when the treatment effect evolves over time. Treatment effects are "stationary" when the the $CATT$ is the same across all cohorts for a given relative time; otherwise the effects are said to be non-stationary, which can be interpreted as heterogeneity

[1] Such practice can be considered as an extension to the Granger causality model popular in DID designs.

in a dynamic setting.

Building on the potential outcome framework, we derive the estimands for conventionally used FE estimators and find that they do not necessarily identify convex averages of $CATT$s, and therefore do not identify an interpretable causal effect. These two-way fixed effects models come in two varieties: static and dynamic. In the static model, the only regressor other than fixed effects is an indicator for a unit being treated at time t. Similar to Borusyak and Jaravel (2017) and de Chaisemartin and D'Haultfœuille (2018), we find that while this FE estimator estimates a weighted average of $CATT$s across all cohorts and relative times, the weights can be non-convex and thus do not correspond to an overall causal effect.

In the dynamic model, researchers include leads and lags of the treatment indicator. When treatment effects are dynamic and heterogenous, we find the dynamic model also does not return interpretable estimates of causal effects. The estimator associated with a particular lead (or lag) l, does not necessarily estimate a weighted average of $CATT$s associated with l; instead, it may pick up spurious terms consisting of treatment effects from periods other than l.[2] This cautions against interpreting the lead coefficient estimates as (lack of) a pre-trend, and the lag coefficient estimates as lagged treatment effects. We propose an alternative estimator that is guaranteed to estimate a convex average of $CATT$s from the corresponding relative time l. These estimators are based on an interacted model that is saturated in relative time and cohort indicators which is first used to estimate $CATT$s. Then we pick the $CATT$s associated with l to form a weighted average, with weights equal to the share of each cohort. These alternative estimates are thus causally interpretable. To our knowledge, the negative results regarding the dynamic model have not been documented.

We demonstrate the empirical relevance of our negative finding regarding the dynamic FE model by re-estimating the economic consequences of hospitalization seen in Dobkin et al. (2018) for an elderly sample. We study hospitalization in a subsample of the Health and Retirement Survey (HRS), consisting of 2,813 patients across five waves of survey (roughly 2004-2012, biennially), and estimate the effect of hospitalization on out-of-pocket medical spending and on labor earnings. In this context we think stationarity might be violated because individuals gradually age into Medicare and retirement, thus we expect the treatment effect may change over time. Using our alternative estimator we find that hospitalization increases out-of-pocket medical spending and decreases labor earnings, similar to the results of Dobkin et al. (2018). In contrast, we show that the FE estimator gives results of the opposite sign, suggesting hospitalization decreases medical spending and increases labor earnings. This example illustrates how FE estimates can lead to misleading results in the presence of dynamic and heterogeneous treatment effects.

[2]Hull (2017) discusses identification of mover average treatment effects in mover regressions. Similar to the dynamic model we consider, mover regressions are two-way fixed effects models using multiple treatment indicators as regressors. Each treatment indicator corresponds to a unique treatment, unlike the present setting where the relevant "treatments" are lags and leads of the current treatment. Hull also finds the multi-valued treatment coefficients identify a linear combination of mover treatment effects and a set of non-causal terms.

In the next section, we relate our findings to the previous literatures on estimating treatment effects of a time-varying treatment. Section 2 introduces the potential outcome framework for event studies. Following that, we derive the asymptotic behavior of the FE estimators and present the negative results on them. Section 4 develops our alternative estimators. Section 5 discusses applications of our results and Section 6 concludes.

Literature

Our results relate to a number of previous literatures. The literature on estimating treatment effects of a time-varying treatment starts with the marginal structural models developed by Hernán et al. (2001). They take a selection-on-observables approach, by assuming sequential randomization conditional on time-varying observable confounders. Event studies typically allow for unobserved time-invariant and separable confounders. By assuming the timing of treatment is exogenous to the growth in outcome if never treated, the timing can depend on unobserved time-invariant confounders in an unrestricted form, as long as they are separable and can be differenced out.[3]

Our analysis contributes to the literature analyzing linear two-way fixed effects models (de Chaisemartin and D'Haultfœuille, 2018 and Kim and Imai, 2017). de Chaisemartin and D'Haultfœuille (2018) focus on a setting where treatment groups are defined by the evolution of treatment rate. For example, the treated group experiences a larger increase in treatment rate than the control group between two time periods. They find that the FE estimator estimates a weighted average of local average treatment effects ($LATE$) in each treatment group, under the assumption that treatment effects are not dynamic. That is, they assume the $LATE$ for each treatment group does not vary after treatment. Kim and Imai (2017) focus on a setting where treatment groups are defined by their treatment history. Under the assumption that treatment has no lasting effect beyond the period of initial treatment, they find that the FE estimator estimates a weighted average of average treatment effects in each treatment group. While we show similar results that the FE estimator estimates a weighted average of average treatment effects in each cohort, due to the structure of event studies, we are able to relax their assumptions on treatment effects.

In event studies, we are able to identify average treatment effects on the treated in each cohort ($CATT$) instead of $LATE$ because cohorts experience sharp increases in treatment rate. Cohorts that receive treatment in the future serve as control group for cohorts that have already received treatment. Unlike de Chaisemartin and D'Haultfœuille (2018), we are thus able to allow for dynamic treatment effects, meaning treatment effects can vary with time since treatment. We can also allow for lasting treatment effects. Kim and Imai

[3]Chernozhukov et al. (2013) studies the case with non-separable unobserved time-invariant confounders. Our results can be extended to non-separable models under the assumption of no time effects or a restricted form of time effects.

(2017) requires that past treatment cannot affect current outcome because it is difficult to adjust for all treatment history and time-invariant unobserved confounders at the same time. To adjust for time-invariant unobserved and separable confounders, we compare treated and control observations across different time periods within the same unit. Had past treatment affected current outcome, then observations in control status of current period might not have the outcome as if the unit is never treated, if the unit was treated in the past. In the absence of lasting treatment effects, conditioning on the contemporaneous treatment status is sufficient for the comparison. Event studies are an exception because treatment is an absorbing state and treatment status changes only once from the control to treatment condition. Observations in control status of current period thus have the outcome as if the unit is never treated (under the assumption of no anticipatory behavior, as we discuss in Section 2. We can thus estimate $CATT$s at lagged time periods. We think allowing for dynamic and lasting treatment effects is important in event studies, and particularly for our application.

Borusyak and Jaravel (2017) focus on event studies and find that when treatment effects are dynamic but homogeneous across cohorts, the FE estimator in the static model could weight long-run treatment effects negatively when no leads or lags of treatment are included in the specification. We show a similar result when treatment effects are allowed to be heterogeneous, and our negative results extend to the specification with leads and lags of treatment. The FE estimators for pre-trends are likely to assign non-zero weight on treatment effects post-treatment, which can yield non-zero pre-trends estimates even though there are no pre-trends. The FE estimators for l lags to treatment are likely to be a non-convex average of $CATT$s from l periods after initial treatment, as well as $CATT$s from other periods, which can yield an estimate with the opposite sign to any $CATT$s from l periods after initial treatment. These negative results have not been documented to our knowledge, and we illustrate their relevance using both a simulation example and an empirical application.

2 Event studies in a potential outcome framework

We assume a balanced panel of N units and $T + 1$ time periods with no missing data. We also assume a simple random sampling of units with T fixed. For each unit $i = 1, \dots, N$ at time $t = 0, \dots, T$, we observe an outcome variable $Y_{i,t}$ and a binary treatment status variable $D_{i,t} \in \{0, 1\}$: $D_{i,t} = 0$ if i has not been treated by period t and $D_{i,t} = 1$ if i has been treated by period t.

For each unit we observe a treatment path $\{D_{i,t}\}_{t=0}^{T}$. In event studies, treatment is an absorbing state and the treatment path is a non-decreasing sequence of zeros and then ones. Thus, the treatment path of a unit can be uniquely characterized by the time period of the initial treatment, identified with the scalar random variable $E_i = \min\{t : D_{i,t} = 1\}$. We call this random variable E_i "event time" as it refers to the onset of the treatment. In our notation, cohort e is the set of units for which $E_i = e$. Denote $E_i = \infty$ for units never treated.

We assume that in this panel, E_i is supported on $\{1, 2, 3, \ldots, T\}$, so no units are treated in the first period ($t = 0$) and all units are treated by the last period ($t = T$). We assume $Pr\{E_i = 0\} = 0$ since we cannot apply DID estimators for $E_i = 0$ as we see later.[4] Our results extend to the case where we observe units never treated in the panel (possibly treated after the panel $E_i > T$, or a pure control group that is never treated $E_i = \infty$).

Potential outcomes describe realizations of the outcome variable $Y_{i,t}$ that arise in response to a hypothetical treatment path e. We denote the potential outcome in period t under treatment path e by $Y_{i,t}^e$. In particular, $Y_{i,t}^\infty$ is the potential outcome if unit i never receives treatment, which we call "baseline outcome". We observe each unit under only a single treatment path $E_i = e$, so the observed outcome for unit i is

$$Y_{i,t} = Y_{i,t}^\infty + \sum_{1 \le e \le T} \left(Y_{i,t}^e - Y_{i,t}^\infty\right) \cdot \mathbf{1}\{E_i = e\} \tag{1}$$

2.1 Identifying assumptions

We maintain the following assumption which makes growth in baseline outcome mean independent of the event time E_i.

Assumption 1. (Parallel trends in baseline outcome.) *$E\left[Y_{i,t}^\infty - Y_{i,s}^\infty | E_i = e\right]$ is the same for all $e \in supp(E_i)$ $\forall s,\ t$ and in particular, equal to $E\left[Y_{i,t}^\infty - Y_{i,s}^\infty\right]$.*

Another assumption necessary for identification of the causal parameters we are interested in precludes treatment from influencing outcomes before treatment takes place.

Assumption 2. (No anticipatory behavior.) *$Y_{i,t}^e = Y_{i,t}^\infty$ for all $t < e$. In other words, potential outcomes before treatment happens are equal to the baseline outcome. Equivalently, $Y_{i,e+l}^e = Y_{i,e+l}^\infty$ for all $l < 0$.*

This assumption is plausible if the full treatment paths are not known to units, so that future treatment cannot affect current potential outcome. However, if the full treatment path is known to the units, then the potential outcome prior to treatment may be affected by future treatments.[5]

We maintain both Assumption 1 and Assumption 2 throughout our analysis.

[4]Intuitively, if a cohort is treated in all periods we cannot identify dynamic treatment effects separately from time effects.

[5]An example is consumption smoothing. Let $Y_{i,t}$ be consumption and $D_{i,t}$ be an indicator of increase in sales tax. Future increase in sales tax $D_{i,1} = 1$ may lower current level of consumption $Y_{i,0}^1$ and we cannot assume $Y_{i,0}^1$ equals $Y_{i,0}^\infty$. Malani and Reif (2015) also gave an example of anticipated tort reforms impacting physicians supply before the adoption of these reforms.

Under these two assumptions, we can identify causal parameters $CATT_{e,l} \coloneqq E\left[Y^{e}_{i,e+l} - Y^{\infty}_{i,e+l} \mid E_i = e\right]$, the cohort-specific average treatment effects on the treated l periods relative to the initial treatment for cohort e as shown later in Proposition 3. For cohort e, l ranges from $-e$ to $T-e$ because at most we observe e periods before initial treatment and $T-e$ periods after initial treatment. Note that we cannot identify unit treatment effect $Y^{E_i}_{i,E_i+l} - Y^{\infty}_{i,E_i+l}$ as in Chernozhukov et al. (2013) without further assumptions.[6]

This assumption also implies no pre-trends since $CATT_{e,l} = 0$ for all $l < 0$.

2.2 Additional assumptions on potential outcomes

We state two additional assumptions that may be imposed on potential outcomes. As we show later, these additional assumptions are necessary for the FE estimators to yield interpretable estimates.

Assumption 3. Static treatment effects. *If $CATT_{e,l} = CATT_{e,l'}$ for all $l \neq l'$, $l, l' \geq 0$, then we say treatment effect for cohort e is static; in contrast, if $CATT_{e,l} \neq CATT_{e,l'}$ for any $l \neq l'$, $l, l' \geq 0$, then we say treatment effect for cohort e is dynamic.*

Within a cohort, we can categorize treatment effects as static or dynamic. Static effects impact units immediately upon treatment and on average, persist at the same level for all treated periods. We call the opposite of static effects dynamic. Treatment effects are likely to be dynamic when the treatment of interest is a transient shock (like an adverse health shock which we will explore in our application) or when there is learning from and adaptation to the treatment over time (like switching to a new firm).

Assumption 4. Stationary treatment effects. *If $CATT$ for all $l \geq 0$ depends only on l and not on e, then we say treatment effects are stationary; in contrast, if for any $l \geq 0$, $CATT_{e,l}$ varies by e, then we say treatment effects are non-stationary.*

We can categorize the overall treatment effects as stationary or non-stationary. Stationary effects mean that on average, each cohort experiences the same effects at any given relative time period. We call the opposite of stationary effects non-stationary, which reflects heterogeneity across cohorts. In an event study, this sort of heterogeneity across cohorts could be driven by time-varying effect. An example of this, which we explore in our application, is that aging into Medicare and retirement could mean the same adverse health shock has a different impact on spending and labor earnings in later waves of the HRS panel.

[6] Chernozhukov et al. (2013) shows that when outcomes with and without treatment have the same distribution in each time period, we can identify unit treatment effect. This assumption would further restrict trends in baseline outcomes $E\left[Y^{\infty}_{i,t} - Y^{\infty}_{i,s}\right] = 0 \forall s, t$ and treatment effects to be static (see Assumption 3 below).

3 Linear two-way FE regression models for event studies

Treatment effects in event studies are often estimated by a linear regression with two-way fixed effects containing unit and time fixed effects. There are two common specifications: static and dynamic. The static specification is

$$Y_{i,t} = \alpha_i + \lambda_t + \gamma D_{i,t} + \xi_{i,t} \tag{2}$$

where $D_{i,t}$ is the treatment indicator in time period t. The regression coefficient estimator $\widehat{\gamma}$ is commonly interpreted as an estimator for some weighted average of $CATT_{e,l}$'s across all cohorts and all lagged time periods. The other common specification, which we refer to as the dynamic specification, adds leads and lags of treatment to the above specification

$$Y_{i,t} = \alpha_i + \lambda_t + \mathbf{D}_{i,t}^{\intercal}\boldsymbol{\mu} + \upsilon_{i,t}\,. \tag{3}$$

Here $\mathbf{D}_{i,t}$ is a column vector collecting $D_{i,t}^{l}$ for $-T \leq l \leq T-1$, where $D_{i,t}^{l}$ is an indicator for being l time periods relative to i's initial treatment ($l = 0$ is the year of initial treatment).[7] The lagged coefficient estimator $\widehat{\mu}_l$ for some $l \geq 0$ is commonly interpreted as an estimator for some weighted average of $CATT_{e,l}$'s across all cohorts e that experience at least l periods of treatment. The regression coefficient estimators $\widehat{\mu}_l$ for $l < 0$ are commonly interpreted as estimators for pre-trends.

Below we show that while these FE estimators do estimate weighted averages of causal parameters $CATT_{e,l}$'s, the weights are in general unreasonable and need not be positive without additional assumptions that restrict treatment effects to be static or stationary. These negative results caution us against always interpreting the FE estimates as estimates for some convex averages of $CATT_{e,l}$'s. We first focus on the static FE estimator and present a result similar to Borusyak and Jaravel (2017) and de Chaisemartin and D'Haultfœuille (2018).

Proposition 1. (Probability limit of the static FE estimator). *Denote* $\bar{D}_{\cdot,e+l} = Pr\left\{E_i \leq e+l\right\}$, $\bar{D}_{e,\cdot} = \frac{T-e}{T+1}$ *and* $\bar{D} = \frac{T-E[E_i]}{T+1}$. *Under Assumption 1 and 2, the regression coefficient estimator* $\widehat{\gamma}$ *from (2) converges in probability to*

$$\widehat{\gamma} \rightarrow_p \sum_{e=1}^{T} \sum_{0 \leq l \leq T-e} \omega_{e,l} ATET_{e,l}$$

where the weight is $\omega_{e,l} \coloneqq Pr\left\{E_i = e\right\} \frac{\left(1-\bar{D}_{\cdot,e+l}-\bar{D}_{e,\cdot}+\bar{D}\right)}{\sum_{t=0}^{T} E\left[\ddot{D}_{i,t}^{2}\right]}$ *and these weights sum up to one.*

[7] Even though we sum l over its possible range, we actually need to exclude at least two relative time indicators due to multicolinearities.

Proposition 1 reveals that, while the static FE estimator converges to a weighted average of $CATT_{e,l}$'s, the weights generally do not coincide with sample frequencies. The numerator of the weight, $1 - \bar{D}_{\cdot,e+l} - \bar{D}_{e,\cdot} + \bar{D}$, is the residual from predicting treatment status $D_{i,t}$ with unit and time fixed effects. It thus downweights (to the point of potentially negatively weighting) the long-run treatment effects (large l) for cohorts with an early onset of treatment (small e) because the fixed effects would overpredict their treatment probability.

As a concrete example, if all units are treated at some time within the panel, some weights will necessarily be negative. Since all units have been treated by the final period, we have $\bar{D}_{\cdot,T} = 1$. Furthermore, $\bar{D}_{e,\cdot} > \bar{D}$ for cohorts treated in the early half of the panel ($e < E[E_i]$). So the weights associated with $CATT_{e,T-e}$'s for these cohorts in the last period are all negative. $\widehat{\gamma}$ would then estimate a non-convex average of $CATT_{e,l}$'s. Its probability limit will be further from a convex average when the $CATT_{e,l}$'s associated with negative weights are of large magnitude, and can even drive the the sign of $\widehat{\gamma}$ to be the opposite to the majority of $CATT_{e,l}$'s.

Special Case: Static and Stationary Treatment Effects. From Proposition 1, we can also see that when treatment effects are static and stationary for all cohorts, the FE estimator $\widehat{\gamma}$ will estimate the causal effect of interest. With those strict assumptions, the $CATT_{e,l}$ does not depend on e nor l so we can move it outside of the summation and we know the weights sum to one; thus the static FE estimator would give a reasonable result as it converges to a constant $CATT$.

Next we present the probability limits of dynamic FE estimators without any restriction on treatment effects. To our knowledge, no paper has yet to study the dynamic specification when treatment effects are dynamic and heterogenous.

Proposition 2. (Probability limits of the dynamic FE estimators). *Under Assumption 1 and 2, the lagged FE estimator $\widehat{\mu}_l$ for some $l \geq 0$ converges in probability to a weighted average of post-treatment $CATT_{e,l'}$ for all $l' \neq l$,*

$$\widehat{\mu}_l \to_p \sum_{1 \leq e \leq T-l} \omega_{e,l}^{l} CATT_{e,l} + \sum_{l' \neq l} \sum_{1 \leq e \leq T-l'} \omega_{e,l'}^{l} CATT_{e,l'},$$

where $\sum_{-l \leq e \leq T-l} \omega_{e,l}^{l} = 1$ and for each $l' \neq l$, $\sum_{-l' \leq e \leq T-l'} \omega_{u,l'}^{l} = 0$. The lead FE estimator $\widehat{\mu}_l$ for some $l < 0$ converges in probability to a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$,

$$\widehat{\mu}_l \to_p \sum_{l' \geq 0} \sum_{1 \leq e \leq T-l'} \omega_{e,l'}^{l} CATT_{e,l'},$$

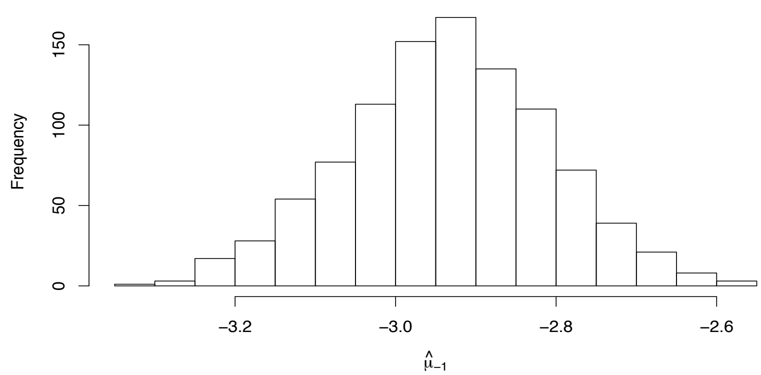
where for each l', $\sum_{-l' \leq e \leq T-l'} \omega_{e,l'}^{l_0} = 0$. The weight $\omega_{e,l'}^{l}$ is the projection coefficient of $D_{i,t}^{l'}$ from regressing $\mathbf{1}\{E_i = e\} \cdot D_{i,t}^{l}$ on the rest of relative time indicators in (3) i.e. $\left\{D_{i,t}^{l'}\right\}_{l' \neq l}$ and two-way fixed effects.

The above proposition yields two negative results. First, Proposition 2 shows that the lagged FE estimators $\widehat{\mu}_l$ for $l \geq 0$ average treatment effects across cohorts from multiple times periods: $CATT_{e,l}$'s as well as $CATT_{e,l'}$'s for $l' \neq l$. Therefore, $\widehat{\mu}_l$ is not isolating the average treatment effects l periods after initial treatment, but rather estimating an affine combination of effects from multiple periods.

Second, lead FE estimators $\widehat{\mu}_l$ for $l < 0$ do not converge to zero even when there are no pre-trends by assumption. Due to the influence of post-treatment $CATT_{e,l'}$'s for $l' \geq 0$, $\widehat{\mu}_l$ does not necessarily converge to zero even though under the no anticipation assumption, $CATT_{e,l} = 0$ for all $l < 0$ and all e so any weighted average of $CATT_{e,l}$'s should be zero.

We demonstrate the negative result on lead FE estimators $\widehat{\mu}_l$ for $l < 0$ through simulation.[8] We estimate (3) on 1000 simulated panel datasets with $T = 3$ and uniformly distributed $E_i \in \{1, 2, 3\}$. We exclude anticipatory behavior by setting $Y_{i,t}^{E_i} = Y_{i,t}^{\infty}$ for all $t < E_i$. We set post-treatment treatment effects to be all positive and grow in time. Even though pre-treatment treatment effect is set to zero, the estimates $\hat{\mu}_{-1}$'s are usually negative as shown in Figure 1, which would suggest a negative pre-trend. Even though here we picked an extreme case where both post-treatment $CATT_{e,l}$'s and $\left|\omega_{u,l'}^l\right|$ are set to be large to illustrate the negative result, we show that in our application, $\left|\omega_{u,l'}^l\right|$ are large as well and cannot be ignored.

Figure 1: Simulation example



Notes: This figure plots the histogram of $\hat{\mu}_{-1}$, the FE estimate for the effect of treatment one period before treatment, across 1000 simulated samples.

Special Case: Stationary Treatment Effects. In the special case of stationary effects, $CATT_{e,l}$ is constant across e for a given l. The FE dynamic estimators thus recover causally interpretable estimates: the lagged FE estimators $\widehat{\mu}_l$ for $l \geq 0$ converges to $CATT_{e,l}$ and the lead FE estimators $\widehat{\mu}_l$ for $l < 0$ converges to zero. Otherwise, it is problematic to interpret a non-zero estimates on $D_{i,t}^l$ for $l < 0$ as (lack of) evidence for pre-trends because non-zero post-treatment $CATT_{e,l}$ can affect the lead FE estimators.

[8]See Appendix B for more details on the simulation design.

4 Alternative method for estimating dynamic treatment effects

We propose a new specification for estimating the dynamic treatment effects in event studies as an alternative to Regression (3). Unlike the dynamic FE estimators, these alternative estimators are weighted averages of $CATT_{e,l}$'s with reasonable weights.

Recall that Proposition (2) implies that under stationary treatment effects when $CATT_{e,l}$'s are constant for a given l, Regression (3) is the correct specification and $\widehat{\mu}_l$ consistently estimates $CATT_{e,l}$. Without stationarity when $CATT_{e,l}$ varies by e, the error term $\epsilon_{i,t}$ involves the difference between $CATT_{e,l}$ and $CATT_{e',l}$ for $e \neq e'$, which is correlated with the unit fixed effects. This hints at using an interacted model, saturated in relative time indicators $D_{i,t}^{l}$ and cohort indicators $\mathbf{1}\{E_i = e\}$, to estimate each of the $CATT_{e,l}$'s; the resulting $CATT_{e,l}$'s are averaged to provide the final estimate. Following Gibbons et al. (2018), we call these estimators "interaction-weighted" (IW) estimators.

Before introducing the IW estimators formally, we define the DID estimator for $CATT_{e,l}$ as

$$\widehat{\delta}_{e,l} = \underbrace{\frac{\mathbb{E}_N\left[(Y_{i,e+l} - Y_{i,s}) \cdot \mathbf{1}\{E_i = e\}\right]}{\mathbb{E}_N\left[\mathbf{1}\{E_i = e\}\right]}}_{\widehat{g}^{l}} - \underbrace{\frac{\mathbb{E}_N\left[(Y_{i,e+l} - Y_{i,s}) \cdot \mathbf{1}\{E_i \in C\}\right]}{\mathbb{E}_N\left[\mathbf{1}\{E_i \in C\}\right]}}_{\widehat{g}^{\infty}} \tag{4}$$

for some $s < e$ and $C \subseteq \{c : e + l < c \leq T\}$. Here we use the notion $\mathbb{E}_N$ to abbreviate the symbol $\frac{1}{N}\sum_{i=1}^{N}$. This estimator is the difference between the average change in outcomes for cohort e, which is exactly l periods relative to treatment, and average change in outcomes for cohorts that have not been treated by $t = e + l$. We refer to this set of cohorts as the control cohorts.[9]

The DID estimator is an unbiased and consistent estimator for $CATT_{e,l}$, a fact that we build on in showing the probability limit of the IW estimator. We state this in the following proposition.

Proposition 3. *Under Assumption 1 and 2, the DID estimator is an unbiased and consistent estimator for* $CATT_{e,l}$.

Proof. Provided that the conditional expectations exist, the DID estimator is an unbiased and consistent estimator for $E[Y_{i,e+l} - Y_{i,s} \mid E_i = e] - E[Y_{i,e+l} - Y_{i,s} \mid E_i \in C]$. To show that it is an unbiased and consistent estimator for $CATT_{e,l}$, we show $E[Y_{i,e+l} - Y_{i,s} \mid E_i = e] - E[Y_{i,e+l} - Y_{i,s} \mid E_i \in C] = CATT_{e,l}$.

[9] When we do not observe $E_i = \infty$, we cannot estimate $E[Y_{i,t} - Y_{i,t}^{\infty}|E_i]$ for $t = T$ or $E_i = T$ using DID estimators because everyone is treated in the last period and $C = \emptyset$. Similarly, we cannot estimate $E[Y_{i,t} - Y_{i,t}^{\infty}|E_i = 0]$ using DID estimators because we do not observe $s < 0$.

Since $s < e$ and $c > e + l$, we have

$$
\begin{aligned}
& E[Y_{i,e+l} - Y_{i,s} \mid E_i = e] - E[Y_{i,e+l} - Y_{i,s} \mid E_i \in C] \\
=& E[Y^e_{i,e+l} - Y^e_{i,s} \mid E_i = e] - \sum_{c \in C} Pr\{E_i = c \mid E_i \in C\} E[Y^c_{i,e+l} - Y^c_{i,s} \mid E_i = c] \\
=& E[Y^e_{i,e+l} - Y^\infty_{i,s} \mid E_i = e] - \sum_{c \in C} Pr\{E_i = c \mid E_i \in C\} E[Y^\infty_{i,e+l} - Y^\infty_{i,s} \mid E_i = c] \\
=& E[Y^e_{i,e+l} - Y^\infty_{i,e+l} \mid E_i = e] + E[Y^\infty_{i,e+l} - Y^\infty_{i,s} \mid E_i = e] - \sum_{c \in C} Pr\{E_i = c \mid E_i \in C\} E[Y^\infty_{i,e+l} - Y^\infty_{i,s} \mid E_i = c] \\
=& E[Y^e_{i,e+l} - Y^\infty_{i,e+l} | E_i = e] + E[Y^\infty_{i,e+l} - Y^\infty_{i,s}] - E[Y^\infty_{i,e+l} - Y^\infty_{i,s}] \\
=& E[Y^e_{i,e+l} - Y^\infty_{i,e+l} | E_i = e]
\end{aligned}
$$

where the second equality follows from Assumption 2 and the fourth equality follows from Assumption 1. □

Note that under Assumption 2, Assumption 1 has many testable implications. There are more than one DID estimators for $CATT_{e,l}$ since we can set different $s < e$ and take subsets of $\{c : e + l < c \leq T\}$ as the control cohorts. One can thus form an omnibus test for the validity of these two assumptions. See Callaway and Sant'Anna (2018) for one such test.

It is also possible to relax the parallel trends assumption to allow the timing of treatment to depend on covariates. We state the conditional parallel trends assumption in Appendix C, and present the doubly robust scores for estimating $CATT_{e,l}$ consistently and efficiently as an extension to the inverse propensity score reweighted estimator proposed by Abadie (2005) and Callaway and Sant'Anna (2018).

4.1 Interaction-weighted estimators

Definition 1. The dynamic interaction-weighted (IW) estimator for a weighted average of $CATT_{e,l}$'s is found by first estimating the following specification

$$Y_{i,t} = \alpha_u + \lambda_t + \mathbf{B}^\intercal_{i,t} \boldsymbol{\delta} + \epsilon_{i,t} \tag{5}$$

on $t = 0, \ldots, T - 1$.[10] Here $\mathbf{B}_{i,t}$ is a column vector collecting $\mathbf{1}\{E_i = e\} \cdot D^l_{i,t}$, for $1 \leq e \leq T - 1$ and for each e, $1 - e \leq l \leq T - e$. Similarly, $\boldsymbol{\delta}$ is a column vector collecting the coefficients $\delta_{e,l}$ on $\mathbf{1}\{E_i = e\} \cdot D^l_{i,t}$. The IW estimator is given by a weighted average of the estimates $\widehat{\delta}_{e,l}$'s:

$$\widehat{\nu}_l := \sum_{e=1}^{T-1-l} \frac{N_e}{\sum_{e=1}^{T-1-l} N_e} \widehat{\delta}_{e,l} = vec\left(\widehat{\mathbf{f}}^l\right)^\intercal \widehat{\boldsymbol{\delta}}$$

[10]We need to drop time period T because $CATT_{e,T}$'s are not identified. See footnote 9 for more details.

where $\widehat{\mathbf{f}}^l$ is a matrix with its $(t,e)^{th}$ entry equal to $\left(\widehat{P}_{t,e}\right)$. Here $\widehat{P}_{t,e} = \mathbf{1}\{t-e=l\} \frac{N_e}{\sum_{e=1}^{T-1-l} N_e}$ and $N_e = \sum_{i=1}^{N} \mathbf{1}\{E_i = e\}$. For example, for $T=3$ and $l=0$,

$$\widehat{\mathbf{f}}^0 = \begin{pmatrix} \frac{N_1}{N_1+N_2} & 0 \\ 0 & \frac{N_2}{N_1+N_2} \end{pmatrix} \to_p \mathbf{f}^0 = \begin{pmatrix} Pr\{E_i = 1 | 1 \leq E_i \leq 2\} & 0 \\ 0 & Pr\{E_i = 2 | 1 \leq E_i \leq 2\} \end{pmatrix}.$$

Lastly, we collect all vectorized $\widehat{\mathbf{f}}^l$'s in $\widehat{\mathbf{f}}$. We can thus write $\widehat{\boldsymbol{\nu}} = (\widehat{\nu}_{1-T}, \ldots, \widehat{\nu}_{T-1})^{\intercal} = \widehat{\mathbf{f}}^{\intercal}\widehat{\boldsymbol{\delta}}$, which converges to $\boldsymbol{\nu} = \mathbf{f}^{\intercal}\boldsymbol{\delta}$.

Proposition 4. (Probability of the IW estimators for dynamic treatment effects). *Under Assumption 1 and 2, the IW estimator converges in probability to*

$$\widehat{\nu}_l \to_p \sum_{e=1}^{T-1} Pr\{E_i = e \mid 1 \leq E_i \leq T-1-l\} \, CATT_{e,l}$$

Note that $\widehat{\delta}_{e,l}$ from Regression (5) is in essence a DID estimator for $CATT_{e,l}$ in the form of (4) with $s=0$ and $C=\{T\}$, which is consistent. Thus, $\widehat{\nu}_l$ is consistent for a weighted average of $CATT_{e,l}$ with weights equal to the share of cohort e across cohorts that have experienced at least l periods of treatment. These weights are guaranteed to be convex and have a reasonable interpretation.

Among all possible DID estimators for $CATT_{e,l}$, $\widehat{\delta}_{e,l}$ might not be the most efficient one: suppose $Y_{i,t}^{\infty}$ follows an AR(1) process with positive correlation, then a possibly more efficient DID estimator sets $s=u-1$, $C=\{c, u+l<c\}$, and weights each cohort c by their precision.[11] That said, we recognize that Regression (5) is easy to implement compared to other more efficient estimators.

We can similarly form an estimator alternative to the static FE estimator from Regression (2) by

$$\widehat{\kappa} := \sum_{e=1}^{T-1} \frac{N_e}{\sum_{e=1}^{T-1} N_e} \sum_{0 \leq l \leq T-1-e} \frac{1}{T-e} \widehat{\delta}_{e,l}.$$

This estimator consistently estimates a weighted average of $CATT_{e,l}$'s, with weights proportional to the representativeness of each $CATT_{e,l}$, a result stated in the following proposition.

Proposition 5. (Probability of the IW estimator for overall treatment effects). *Under Assumption 1 and 2, the IW estimator converges in probability to*

$$\widehat{\kappa} \to_p \sum_{e=1}^{T-1} Pr\{E_i = e \mid E_i \leq T-1\} \sum_{0 \leq l \leq T-1-e} \frac{1}{T-e} CATT_{e,l}$$

[11] Under random sampling, the variance of the DID estimator is the sum of variance of $\widehat{g}^l$ and $\widehat{g}^{\infty}$. Note that the variance of $Y_{i,t}^{\infty} - Y_{i,s}^{\infty}$ decrease in $|t-s|$, which suggests we set $s=u-1$. Suppose the variance of $Y_{i,t}^{\infty} - Y_{i,s}^{\infty}$ differs across cohorts, then the variance of $\widehat{g}^{\infty}$ is minimized when we weight each cohort c by the inverse of its variance.

Finally, we derive the asymptotic distribution for the dynamic FE and IW estimators. For simplicity, we assume that like Regression (5), Regression (3) is also estimated on $t = 0, \dots, T-1$.

Proposition 6. (Asymptotic distribution of the dynamic FE and IW estimators). *Under standard assumptions, the asymptotic distribution of the estimators is*

$$\sqrt{N}\begin{pmatrix} \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \\ \hat{\boldsymbol{\nu}} - \boldsymbol{\nu} \end{pmatrix} \to_d N\Big(\mathbf{0}, \begin{pmatrix} \Sigma_{FE} & \Sigma_{12} \\ \Sigma_{12}^\intercal & \Sigma_{IW} \end{pmatrix}\Big) \tag{6}$$

where

$$\boldsymbol{V}_{\ddot{\mathbf{D}}} = \left(\sum_{t=0}^{T-1} E\left[\ddot{\mathbf{D}}_{i,t}^\intercal \ddot{\mathbf{D}}_{i,t}\right]\right) \qquad \boldsymbol{V}_{\ddot{\mathbf{B}}} = \left(\sum_{t=0}^{T-1} E\left[\ddot{\mathbf{B}}_{i,t}^\intercal \ddot{\mathbf{B}}_{i,t}\right]\right)$$

$$\Sigma_{FE} = \boldsymbol{V}_{\ddot{\mathbf{D}}}^{-1}\left(\sum_{t=0}^{T-1} E\left[\ddot{\mathbf{D}}_{i,t}^\intercal \ddot{v}_{i,t}^2 \ddot{\mathbf{D}}_{i,t}\right]\right)\boldsymbol{V}_{\ddot{\mathbf{D}}} \qquad \Sigma_{IW} = \mathbf{f}^\intercal \boldsymbol{V}_{\ddot{\mathbf{B}}}^{-1}\left(\sum_{t=0}^{T-1} E\left[\ddot{\mathbf{B}}_{i,t}^\intercal \ddot{\epsilon}_{i,t}^2 \ddot{\mathbf{B}}_{i,t}\right]\right)\boldsymbol{V}_{\ddot{\mathbf{B}}}\mathbf{f}$$

$$\Sigma_{12} = \boldsymbol{V}_{\ddot{\mathbf{D}}}^{-1}\left(\sum_{t=0}^{T-1} E\left[\ddot{\mathbf{D}}_{i,t}^\intercal \ddot{v}_{i,t} \ddot{\epsilon}_{i,t} \ddot{\mathbf{B}}_{i,t}\right]\right)\boldsymbol{V}_{\ddot{\mathbf{B}}}\mathbf{f}$$

We use the notation $\ddot{X}_{i,t}$ to denote $X_{i,t} - \overline{X}_{i,\cdot} - \overline{X}_{\cdot,t} + \overline{X}$, where $\overline{X}_{i,\cdot} = \frac{1}{T}\sum_{t=0}^{T-1} X_{i,t}$, $\overline{X}_{\cdot,t} = E\left[X_{i,t}\right]$ and $\overline{X} = \frac{1}{T}\sum_{t=0}^{T-1} E\left[X_{i,t}\right]$.

In the presence of non-stationary treatment effects, the saturated specification (5) may reduce standard errors of residuals due to better fit. The IW estimators may thus be more precise. Under stationary treatment effects when both the dynamic FE and IW estimators are consistent for $CATT_{e,l}$, which are constant across e for a given l. The dynamic FE estimator might have smaller variance because specification (3) estimates less coefficients. There is established efficiency gain when specification (3) is a "grouped" version of specification (5) and errors are homoskedastic and serially uncorrelated. Specification (3) is a "grouped" version if all included $D_{i,t}^l$ can be written as $D_{i,t}^l = \sum_{e\geq 1}^{T-1-l} \mathbf{1}\{E_i = e\} \cdot D_{i,t}^l$, and all such $\mathbf{1}\{E_i = e\} \cdot D_{i,t}^l$ are included in specification (5). In this case, $\mathbf{D}_{i,t}$ is a linear transformation of $\mathbf{B}_{i,t}$. To construct such specifications, note that if there is no pre-trends, then we only need to include lags of treatments in (3) and $\mathbf{1}\{E_i = e\} \cdot D_{i,t}^l$ for $l \geq 0$ can all be included in the saturated model. If further error terms are homoskedastic and serially uncorrelated, then by Gauss-Markov theorem the dynamic FE estimators are more efficient.

5 Applications

We illustrate the empirical importance of our findings by extending the analysis by Dobkin et al. (2018). We use both FE and IW estimators to estimate the impact of hospital admissions on out-of-pocket medical spending and labor earnings using a subsample of respondents from the Health and Retirement Survey (HRS). We show that the FE estimators produces statistically insignificant results with point estimates of the "wrong" sign; the FE estimates suggest that hospital admissions may decrease out-of-pocket medical spending and increase earnings for the elderly. In contrast, we show that estimates of $CATT_{e,l}$'s are of the expected signs and consequently the IW estimator recovers that hospitalization increases out-of-pocket medical spending right after hospitalization and decreases long-run labor earnings.

5.1 Settings

Our sample selection follows Dobkin et al. (2018) but we include a cursory explanation here for completeness. Our primary source of data is the biennial Health and Retirement Survey (HRS). We identify the sample of individuals who appear in two sequential waves of surveys and newly report having a hospital admission over the last two years (the "index" admission) at the second survey. To focus on health "shocks", we restrict attention to non-pregnancy-related hospital admissions. We restrict our analysis to a subsample of these individuals who appear throughout waves 7-11 (roughly 2004-2012) to maintain a balanced panel with a reasonable sample size.[12] Our sample of analysis therefore includes HRS respondents with index hospitalization during wave 8-11. Here i naturally indexes individual, and t indexes survey wave ($T = 4$) and is normalized to zero for wave 7, the first wave in our sample.

The treatment status $D_{i,t}$ is thus an indicator for whether a respondent has had an unexpected hospitalization by wave t. We think hospitalization is an absorbing state because it approximates the start of a sick state (as opposed to a healthy state). Thus, event study is an appropriate research design as it allows for lasting treatment effect of hospitalization. Our outcomes of interest $Y_{i,t}$ include out-of-pocket medical spending and labor earnings. They are derived from self-reports, adjusted to 2005 real dollar value and censored at the 99.95th percentile.

In our terminology, we categorize individuals into cohorts based on E_i, which is defined as the survey wave of their index hospitalization. Since we restrict the sample to individuals who were ever hospitalized between wave 8-11, $E_i \in \{1, 2, 3, 4\}$.

Parallel trend assumption Hospitalization is likely to be earlier among sicker individuals with high out-of-pocket medical spending and low labor earnings, even when restricted to individuals who were ever hospitalized. Thus, it is not plausible that the baseline outcome $Y_{i,t}^{\infty}$ is mean independent of the timing of hospitalization. Parallel trends assumption is more plausible as it allows the timing to depend on unobserved time-invariant characteristics such as chronic disease.

[12] For their main results, Dobkin et al. (2018) focus on adults with health insurance who are hospitalized at ages 50-59. We include adults of all ages.

No anticipation assumption Given that the treatment is restricted to be unexpected hospitalization, it is also plausible that there is no anticipatory behavior.

Potential treatment effects heterogeneity The effect on out-of-pocket medical spending is largely determined by generosity of health insurance, which has possibly decreased as individuals age into Medicare. The effect on labor earnings is affected by the labor market: for example, individuals hospitalized during the financial crisis may find it more difficult to return to the labor force, and suffer a more grave decrease in earnings. We suspect these sources of non-stationarity may be exacerbated in this sample because the elderly population has experienced the especially rapid change in healthcare spending and labor participation over time.

Summary Statistic Table 1 presents basic summary statistics for our analysis sample before hospitalization. By design, our sample is older than that of Dobkin et al. (2018) because Dobkin et al. (2018) restricts to non-elderly adults with ages 50-59 at the time of index hospitalization. Other demographic differences between our sample and that of Dobkin et al. (2018) align with expectations based on the differences in age. We have a slightly lower fraction of men in our sample, as well as a higher fraction white and lower fraction black. Many individuals have also aged into Medicare coverage (for which the qualifying age is 65).

In panel D, we compare means of the cross-sectional distributions of outcomes for individuals who have not been hospitalized by each wave. The number of respondents hospitalized in each cohort is $N_1 = 833$, $N_2 = 711$, $N_3 = 733$, and $N_4 = 536$. Thus, the size of the sample conditional on not having been hospitalized strictly decreases each wave. There are apparent time trends in our outcomes of interest prior to hospitalization as we observe distributional changes across waves. Out-of-pocket medical spendings decrease first on average, possibly due to coverage by Medicare, but eventually increase. Earnings decrease each wave on average as more individuals are retired each wave.

Table 1: Sample characteristics

	N	Mean	Std. Dev
Panel A. Demographics			
Age at admission	2,813	68	9
Male	2,813	0.425	0.494
Year of admission	2,813	2,007	2
Panel B. Race/ethnicity			
Hispanic	2,813	0.097	0.296
Black	2,813	0.126	0.332
White	2,813	0.817	0.387
Other race	2,813	0.057	0.232
Panel C. Index hospitalization			
Medicaid	2,813	0.057	0.232
Private	2,813	0.508	0.5
Medicare	2,813	0.57	0.495
Panel D. Pre-hospitalization outcome			
Out-of-pocket medical spending			
Wave 7	2,813	3,143	8,797
Wave 8	1,980	2,339	5,767
Wave 9	1,269	1,858	2,746
Wave 10	536	2,412	4,105
Earnings			
Wave 7	2,813	23,656	49,163
Wave 8	1,980	20,539	43,201
Wave 9	1,269	17,723	40,486
Wave 10	536	14,773	39,246

We estimate the following two specifications without survey weights. For the linear two-way FE regression in the spirit of equation (3), we estimate

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{l=-2}^{3} D_{i,t}^l + \upsilon_{i,t} \tag{7}$$

for $t = 0, \ldots, 4$. For the saturated model in the spirit of (5), we estimate

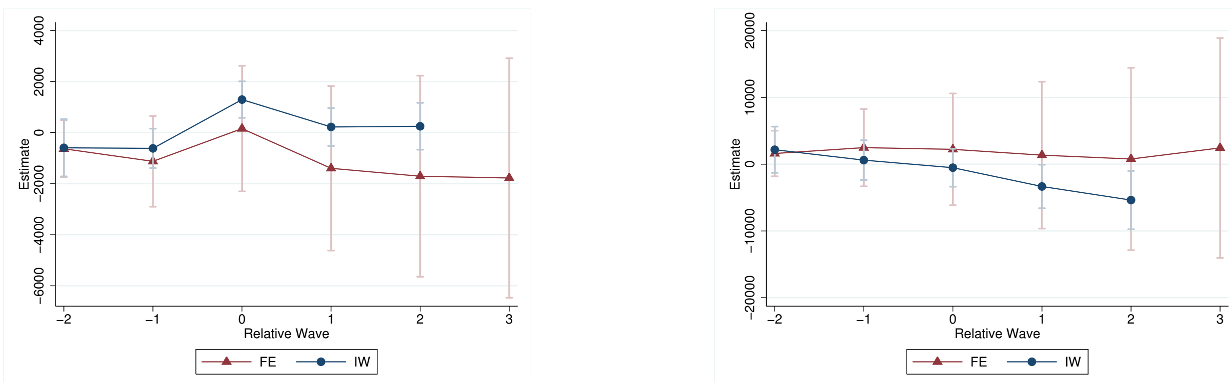
$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \in \{1,2,3\}} \sum_{l=1-e}^{T-e} \nu_{e,l} \mathbf{1}\{E_i = e\} \cdot D_{i,t}^l + \epsilon_{i,t} \tag{8}$$

for $t = 0, \ldots, 3$. We drop $t = 4$ from (8) because everyone has had hospitalization by $t = 4$, and $CATT_{e,l}$'s in $t = 4$ are not identified. We estimate the above specifications with $Y_{i,t}$ equal to out-of-pocket medical spending as well as labor earnings.

5.2 Results and Discussion

Figure 2 compares the dynamic FE estimates of the impact of hospitalization with the IW estimates for two outcomes: out-of-pocket medical spending and labor earnings. The IW estimates show qualitatively similar results to Dobkin et al. (2018). The effect on out-of-pocket spending is significant and positive for $l = 0$ (roughly 1 year after hospitalization). The effect on earnings becomes significantly negative starting in $l = 1$ (roughly 3 years after hospitalization) and remains large for $l = 2$. In contrast, the dynamic FE estimates are all statistically insignificant, even though the underlying $CATT_{e,l}$'s are mostly precisely estimated. Furthermore, the FE point estimates are of the opposite signs of the IW estimates, which are unreasonable as they suggest that out-of-pocket medical spending decrease while earnings increase after hospitalization.

Figure 2: FE vs IW Estimates of the Effect of Hospitalization on Outcomes



(a) Out-of-pocket Medical Spending (b) Labor Earnings

Notes: Each figure plots FE estimates $\widehat{\mu}_l$ (from (7)) in maroon triangles and IW estimates $\widehat{\nu}_l$ (from (8)) in blue circles against relative time l, with their respective 95% confidence intervals. Recall that $\widehat{\nu}_l$ are constructed as the weighted average of $\widehat{\delta}_{e,l}$'s. The outcome variable is out-of-pocket medical spending in Panel A and labor earnings in panel B.

In Table 2, we report the FE estimates $\widehat{\mu}_l$ and the IW estimates $\widehat{\nu}_l$, as well as the underlying $CATT_{e,l}$ estimates $\widehat{\delta}_{e,l}$. While not reported, we fail to reject the joint hypothesis of parallel trends and no anticipation based on a Wald test for $H_0 : \mu_{-2} = \mu_{-1} = 0$ or $H_0 : \delta_{3,-2} = \delta_{3,-1} = \delta_{2,-1} = 0$.

The IW estimates $\widehat{\nu}_l$ are weighted averages of estimates for $CATT_{e,l}$, with weights equal to the share of cohort e across cohorts that experience at least l periods of treatment. Therefore by construction, the IW estimate falls within the convex hull of the $CATT_{e,l}$ estimates and has an interpretation as an average effect of the treatment l periods after initial treatment. In contrast, the FE estimates are not within the convex hull of the $CATT_{e,l}$ estimates and thus do not have a causal interpretation.

Table 2: Estimates for the Effect of Hospitalization on Outcomes

(a) Out-of-pocket Medical Spending

l wave relative to index hospitalization	FE estimates $\widehat{\mu}_l$	IW estimates $\widehat{\nu}_l$	Estimates for $CATT_{e,l}$ $\widehat{\delta}_{1,l}$	$\widehat{\delta}_{2,l}$	$\widehat{\delta}_{3,l}$
-2	-630	-592	-	-	-592
	(571)	(572)			(572)
-1	-1,122	-614	-	-605	-613
	(907)	(394)		(509)	(478)
0	160	1,297	676	1,691	1,619
	(1255)	(365)	(506)	(533)	(513)
1	-1,396	224	35	445	-
	(1644)	(379)	(404)	(536)	
2	-1,706	251	251	-	-
	(2010)	(468)	(468)		
3	-1,775	-	-	-	-
	(2394)				

(b) Labor Earnings

l wave relative to index hospitalization	FE estimates $\widehat{\mu}_l$	IW estimates $\widehat{\nu}_l$	Estimates for $CATT_{e,l}$ $\widehat{\delta}_{1,l}$	$\widehat{\delta}_{2,l}$	$\widehat{\delta}_{3,l}$
-2	1,600	2,165	-	-	2,165
	(1743)	(1766)			(1766)
-1	2,475	605	-	1,879	-620
	(2953)	(1519)		(1888)	(1811)
0	2,221	-527	-557	439	-1,429
	(4268)	(1449)	(1866)	(2147)	(1981)
1	1,352	-3,337	-3,518	-3,124	-
	(5609)	(1663)	(2015)	(2188)	
2	773	-5,375	-5,375	-	-
	(6965)	(2229)	(2229)		
3	2,434	-	-	-	-
	(8396)				

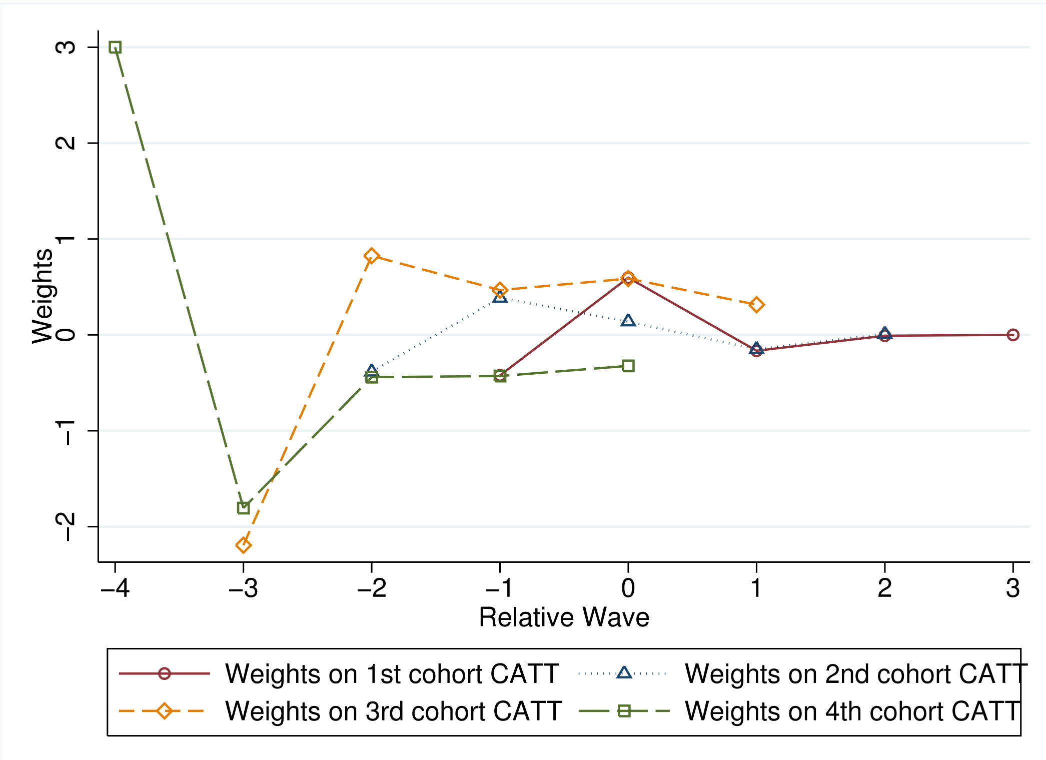
Notes: This table reports three different sets of estimates for the dynamic effects of hospitalization on out-of-pocket medical spending (panel A) and labor earnings (panel B). The first column reports estimates from the FE estimator $\widehat{\mu}_l$ as in (7). Columns 3-5 report the estimates for $CATT_{e,l}$ from $\widehat{\delta}_{e,l}$ as in (8). Column 2 reports the IW estimates which are constructed as the weighted average of $\widehat{\delta}_{e,l}$ across cohorts e who have experienced l periods of treatment. Standard errors (clustered on the individual) are shown in parentheses.

The reason for the dynamic specification to produce wrong-signed and imprecise estimates is that $\hat{\mu}_l$'s are sensitive to effects from other time periods. Recall by Proposition (2), the probability limit of $\hat{\mu}_l$ is

$$\hat{\mu}_l \rightarrow_p \sum_{1 \leq e \leq T-l} \omega^l_{e,l} CATT_{e,l} + \sum_{l' \neq l} \sum_{-l' \leq e \leq T-l'} \omega^l_{e,l'} CATT_{e,l'} \tag{9}$$

and the point estimate $\hat{\mu}_l$ is the sample analog of (9). For $l = 0$, we can estimate the weights $\omega^0_{e,l}$ by regressing $\mathbf{1}\{E_i = e\} \cdot D^l_{i,t}$ on the relative time indicators in (7) i.e. $\left\{D^l_{i,t}\right\}^3_{l=-2}$ and two-way fixed effects. The regression coefficient estimator of $D^0_{i,t}$, $\widehat{\omega}^0_{e,l}$, consistently estimates $\omega^0_{e,l}$. The point estimate $\hat{\mu}_0$ is thus the inner product of these estimated weights $\widehat{\omega}^0_{e,l}$ and $\widehat{\delta}_{e,l}$. Figure 3 plots these estimated weights. The weights are large for leads of treatments (negative relative waves), which suggest that the point estimate $\widehat{\mu}_0$ is particularly sensitive to estimates of pre-trends, and does not isolate the contemporaneous effect of hospitalizations.[13] While we just showed that there is likely no anticipation and in the limit, $\hat{\mu}_0$ is not affected by pre-trends since all pre-trends $CATT_{e,l} = 0\ \forall l < 0$, in finite samples, their estimates $\widehat{\delta}_{e,l}\ \forall l < 0$ are not necessarily zero and can influence $\hat{\mu}_0$ if their weights are non-negligible as shown in Figure 3.

Figure 3: Weights $\widehat{\omega}^0_{e,l}$ on cohort specific effect estimates $\widehat{\delta}_{e,l}$ in forming point estimate $\widehat{\mu}_0$



Notes: (9) implies that $\widehat{\mu}_0$ is a weighted average of $\widehat{\delta}_{e,l}$, estimates for $CATT_{e,l}$. This figure plots these weights $\widehat{\omega}^0_{e,l}$ in forming the FE estimate $\widehat{\mu}_0$.

6 Conclusions

Linear two-way fixed effects (FE) regressions are commonly used to estimate dynamic treatment effects in event studies. Their behavior has not been closely studied when treatment effects are non-stationary, i.e. there is heterogeneity in treatment effects across cohorts. The main goal of this paper is to analyze the

[13]To see this, note the DID estimator $\widehat{\delta}_{4,l}$ and $\widehat{\delta}_{e,-e}$ are all normalized to zero. While $\widehat{\delta}_{e,T-3}$ are not shown in Table 2 above, they receive small weights in forming $\hat{\mu}_0$. Thus, one can take the inner product of weight estimates $\widehat{\omega}^0_{e,l}$ from Figure 3 and $\widehat{\delta}_{e,l}$ from Table 2 to confirm that the result is similar to $\hat{\mu}_0$. In particularly, $\widehat{\delta}_{3,-2}$, $\widehat{\delta}_{2,-1}$ and $\widehat{\delta}_{2,-1}$ receive positive weights, and $\widehat{\delta}_{1,1}$ and $\widehat{\delta}_{2,1}$ receive negative weights, which explains why $\hat{\mu}_0$ are likely to have signs closer to the pre-trends estimates and to the negative of estimates of long-run effects.

behavior of FE estimators under heterogeneous treatment effects. We first cast event studies in a potential outcome framework and show the causal parameters, cohort-specific average treatment effects on the treated ($CATT_{e,l}$), are identified under parallel trends and no anticipation assumptions. We clarify the notion of heterogeneity in a dynamic setting: $CATT_{e,l}$'s evolves differently across cohorts, such that at a given lag, $CATT_{e,l}$ varies by e.

Building on the potential outcome framework, we derive asymptotic distribution of FE estimators and show they converge to weighted averages of $CATT_{e,l}$'s. The weights sum to one but need not be positive. This means when treatment effects are dynamic or heterogenous, the static FE estimate may not correspond to a causal effect as its estimand may fall outside the convex hull of $CATT_{e,l}$'s. In the dynamic specification, the situation is even worse: under heterogeneous treatment effects, in addition to non-convex weights, the FE estimator associated with l periods relative to initial treatment may pick up spurious terms consisting of treatment effects form periods other than l. Researchers can easily verify the robustness of their FE estimators by estimating these underlying weights.

Given the negative results on FE estimators, we propose "interaction-weighted" (IW) estimators for estimating dynamic treatment effects in event studies. These estimators are formed by first estimating $CATT_{e,l}$'s with a regression saturated in cohort and relative time indicators, and then average estimates of $CATT_{e,l}$'s across e at different l's. These estimators are easy to implement and robust to heterogenous treatment effects: the IW estimator associated with relative time l is guaranteed to estimate a convex average of $CATT_{e,l}$'s, with weighting being share of each cohort e.

Finally, we illustrate the empirical relevance of our results by estimating the economic consequence of hospitalization on the elderly using the Health and Retirement Survey. We compare FE estimates with IW estimates of the dynamic effects of hospitalization. We find that IW estimates are of the correct sign and more precise relative to FE estimates. We estimate the weights in forming FE estimates and show they are non-convex. As an example, the FE estimate of contemporaneous effects is sensitive to pre-trends estimates and negatively influenced by long-run effects, which makes it causally uninterpretable.

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 2005, *72* (1), 1–19.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," Working Paper, Harvard University May 2017.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences With Multiple Time Periods and an Application on the Minimum Wage and Employment," March 2018.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey, "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, May 2017, *107* (5), 261–65.

_ , Ivan Fernandez-Val, Jinyong Hahn, and Whitney Newey, "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 2013, *81* (2), 535–580.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-way fixed effects estimators with heterogeneous treatment effects," Working Paper, UCSB April 2018.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo, "The Economic Consequences of Hospital Admissions," *American Economic Review*, February 2018, *108* (2), 308–52.

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic., "Broken or Fixed Effects?," *Journal of Econometric Methods*, 2018, *forthcoming.*

Heckman, James, Hidehiko Ichimura, and Petra E. Todd, "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 1997, *64* (4), 605–654.

Hernán, Miguel A, Babette Brumback, and James M Robins, "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments," *Journal of the American Statistical Association*, 2001, *96* (454), 440–448.

Hull, Peter, "Estimating Treatment Effects in Mover Designs," Working Paper, Microsoft Research October 2017.

Kim, In Song and Kosuke Imai, "When Should We Use Fixed Effects Regression Models for Causal Inference with Longitudinal Data?," Working Paper, Princeton University 2017.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, 2015, *124*, 1 – 17.

A Proofs

Proof of Proposition 1

Proof. Partialling out the unit and time fixed effects, Regression (2) is

$$\ddot{Y}_{i,t} = \gamma \ddot{D}_{i,t} + \epsilon_{i,t}$$

where $\ddot{X}_{i,t}$ is time- and cross-sectional demeaned version of $X_{i,t}$ i.e. $\ddot{X}_{i,t} = X_{i,t} - \bar{X}_{\cdot,t} - \bar{X}_{i,\cdot} + \bar{X}$, and $\bar{X}_{\cdot,t} = E[X_{i,t}]$, $\bar{X}_{i,\cdot} = \frac{1}{T+1}\sum_{t=0}^{T} X_{i,t}$, $\bar{X} = \frac{1}{T+1}\sum_{t=0}^{T} E[X_{i,t}]$. It is easy to see that for $\ddot{X}_{i,t}$ and $\ddot{Z}_{i,t}$, demeaned versions of $X_{i,t}$ and $Z_{i,t}$ respectively, we have $E\left[\ddot{X}_{i,t}\ddot{Z}_{i,t}\right] = E\left[X_{i,t}\ddot{Z}_{i,t}\right] = E\left[\ddot{X}_{i,t}Z_{i,t}\right]$. The projection coefficient on $D_{i,t}$ from (2), i.e. the probability of the coefficient estimator $\widehat{\gamma}$, is then

$$\gamma = \frac{\frac{1}{T+1}\sum_{t=0}^{T} E\left[\ddot{Y}_{i,t}\ddot{D}_{i,t}\right]}{\frac{1}{T+1}\sum_{t=0}^{T} E\left[\ddot{D}_{i,t}\ddot{D}_{i,t}\right]} = \frac{\sum_{t=0}^{T} E\left[Y_{i,t}\ddot{D}_{i,t}\right]}{\sum_{t=0}^{T} E\left[D_{i,t}\ddot{D}_{i,t}\right]} = \frac{\sum_{t=0}^{T} E\left[E\left[\left(Y_{i,t} - Y_{i,0}\right)\ddot{D}_{i,t} \mid E_i\right]\right]}{\sum_{t=0}^{T} E\left[D_{i,t}\ddot{D}_{i,t}\right]} \tag{10}$$

Note that $\ddot{D}_{i,t}$ is a function of E_i. Conditioning on $E_i = e$, $\ddot{D}_{i,t} = D_{i,t} - \bar{D}_{\cdot,t} - \bar{D}_{u,\cdot} + \bar{D}$ is non-random: when $t \geq e, D_{i,t} = 1$ and when $t < e$, $D_{i,t} = 0$; $\bar{D}_{\cdot,t} = Pr\left\{D_{i,t} = 1\right\} = Pr\left\{E_i \leq t\right\}$ is the share of treated units at time t; $\bar{D}_{e,\cdot} = \frac{T-e}{T+1}$ is the share of treated periods; and $\bar{D} = \frac{T-E[E_i]}{T+1}$ is fixed for all i. We thus have

$$\begin{aligned}
\gamma =& \frac{\sum_{t=0}^{T} E\left[\ddot{D}_{i,t} E\left[Y_{i,t}^{E_i} - Y_{i,0}^{E_i} \mid E_i\right]\right]}{\sum_{t=0}^{T} E\left[D_{i,t}\ddot{D}_{i,t}\right]} \\
=& \frac{1}{\sum_{t=0}^{T} E\left[D_{i,t}\ddot{D}_{i,t}\right]} \sum_{t=0}^{T} \Big\{ \sum_{e \leq t} Pr\left\{E_i = e\right\}\left(1 - \bar{D}_{\cdot,t} - \bar{D}_{e,\cdot} + \bar{D}\right) \overbrace{E\left[Y_{i,t}^{e} - Y_{i,0}^{e} \mid E_i = e, D_{i,t} = 1\right]}^{=E\left[Y_{i,t}^{e} - Y_{i,0}^{\infty} \mid E_i = e\right]} && (11) \\
&+ \sum_{e > t} Pr\left\{E_i = e\right\}\left(0 - \bar{D}_{\cdot,t} - \bar{D}_{e,\cdot} + \bar{D}\right) \overbrace{E\left[Y_{i,t}^{e} - Y_{i,0}^{e} \mid E_i = e, D_{i,t} = 0\right]}^{=E\left[Y_{i,t}^{\infty} - Y_{i,0}^{\infty}\right]} \Big\} && (12)
\end{aligned}$$

The equality inside the braces of (11) follows from no anticipation (Assumption 2). The equality inside the braces of (12) follows from parallel trends and no anticipation (Assumption 1 and 2).

Furthermore, we have

$$\begin{aligned}
&\sum_{e \leq t} Pr\left\{E_i = e\right\}\left(1 - \bar{D}_{\cdot,t} - \bar{D}_{e,\cdot} + \bar{D}\right) + \sum_{e > t} Pr\left\{E_i = e\right\}\left(0 - \bar{D}_{\cdot,t} - \bar{D}_{e,\cdot} + \bar{D}\right) \\
&= Pr\left\{D_{i,t} = 1\right\} - \bar{D}_{\cdot,t} - \sum_{e} \bar{D}_{e,\cdot} + \bar{D} = \bar{D}_{\cdot,t} - \bar{D}_{\cdot,t} - \bar{D} + \bar{D} = 0
\end{aligned}$$

With this, we can distribute $E\left[Y_{i,t}^{\infty}-Y_{i,0}^{\infty}\right]$ across the treated units and obtain the following expression:

$$\frac{1}{\sum_{t=0}^{T}E\left[D_{i,t}\ddot{D}_{i,t}\right]}\sum_{t=0}^{T}\{\sum_{e\leq t}Pr\left\{E_i=e\right\}\left(1-\bar{D}_{\cdot,t}-\bar{D}_{e,\cdot}+\bar{D}\right)\left(E\left[Y_{i,t}^{e}-Y_{i,0}^{\infty}\mid E_i=e\right]-E\left[Y_{i,t}^{\infty}-Y_{i,0}^{\infty}\right]\right)\}$$

$$=\sum_{e=1}^{T}\sum_{l=0}^{T-e}\underbrace{\frac{Pr\left\{E_i=e\right\}\left(1-\bar{D}_{\cdot,e+l}-\bar{D}_{e,\cdot}+\bar{D}\right)}{\sum_{t=0}^{T}E\left[D_{i,t}\ddot{D}_{i,t}\right]}}_{=:\omega_{e,l}}E\left[Y_{i,t}^{e}-Y_{i,t}^{\infty}|E_i=e\right] \tag{13}$$

To see that the weights sum up to one, note that the denominator of the weights can be written as

$$\begin{aligned}\sum_{t=0}^{T}E\left[D_{i,t}\ddot{D}_{i,t}\right]&=\sum_{t=0}^{T}\sum_{e\leq t}Pr\left\{E_i=e\right\}\left(1-\bar{D}_{\cdot,t}-\bar{D}_{e,\cdot}+\bar{D}\right)\\&=\sum_{e=1}^{T}\sum_{l=0}^{T-e}Pr\left\{E_i=e\right\}\left(1-\bar{D}_{\cdot,e+l}-\bar{D}_{e,\cdot}+\bar{D}\right)\end{aligned}$$

□

Proof of Proposition 2

Proof. Following a similar argument in the proof for Proposition 1, the projection coefficients μ_l's from regression (3) are:

$$(\mu_l)_{-T\leq l\leq T-1}=\left(\sum_{t=0}^{T}E\left[\ddot{\mathbf{D}}_{i,t}\mathbf{D}_{i,t}^{\intercal}\right]\right)^{-1}\left(\sum_{t=0}^{T}E\left[\ddot{\mathbf{D}}_{i,t}Y_{i,t}\right]\right)$$

where $\ddot{\mathbf{D}}_{i,t}=\begin{pmatrix}\ddot{D}_{i,t}^{-T}\\ \vdots\\ \ddot{D}_{i,t}^{T-1}\end{pmatrix}$ and again, $\ddot{X}_{i,t}$ is time- and cross-sectional demeaned version of $X_{i,t}$.[14]

To further develop the expression for the projection coefficients μ_l's, we note that

[14]Note that $\sum_{l=-T}^{T-1}\ddot{D}_{i,t}^{l}=0$ and we need to drop one $\ddot{D}_{i,t}^{l}$ due to multicollinearity. Another multicollinearity is $\sum_{l=-T}^{T-1}(l-l')\ddot{D}_{i,t}^{l}=0$ for l' the excluded relative time indicator variable. For simplicity, suppose that the excluded $\ddot{D}_{i,t}^{l}$ are two lead indicators i.e. $l<0$.

$$
\begin{aligned}
&\sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t} Y_{i,t}\right] = \sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t}\left(Y_{i,t} - Y_{i,0}\right)\right] = \sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t} E\left[Y_{i,t} - Y_{i,0} | E_i\right]\right] \\
=& \sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t} E\left[Y_{i,t} - Y_{i,0} | E_i\right]\right] - \sum_{t=0}^{T} \underbrace{E\left[\ddot{\mathbf{D}}_{i,t}\right] E\left[Y_{i,t}^{\infty} - Y_{i,0}\right]}_{=\mathbf{0}} \\
=& \sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t}\left(E\left[Y_{i,t} - Y_{i,0} | E_i\right] - E\left[Y_{i,t}^{\infty} - Y_{i,0}\right]\right)\right]
\end{aligned}
$$

Under the parallel trends and no anticipation assumptions, we have $E\left[Y_{i,t}^{\infty} - Y_{i,0} | E_i\right] = E\left[Y_{i,t}^{\infty} - Y_{i,0}\right]$. Thus the expression for projection coefficients μ_l's now simplifies to

$$
\left(\sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t} \mathbf{D}_{i,t}^{\intercal}\right]\right)^{-1} \left(\sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t} E\left[Y_{i,t} - Y_{i,t}^{\infty} | E_i\right]\right]\right)
$$

Furthermore, we can write

$$
\begin{aligned}
E\left[Y_{i,t} - Y_{i,t}^{\infty} | E_i\right] &= \begin{pmatrix} D_{i,t}^{-T} \\ \vdots \\ D_{i,t}^{0} \\ \vdots \\ D_{i,t}^{T-1} \end{pmatrix}^{\intercal} \begin{pmatrix} E\left[Y_{i,E_i-T} - Y_{i,E_i-T}^{\infty} | E_i\right] \\ \vdots \\ E\left[Y_{i,E_i} - Y_{i,E_i}^{\infty} | E_i\right] \\ \vdots \\ E\left[Y_{i,E_i+T-1} - Y_{i,E_i+T-1}^{\infty} | E_i\right] \end{pmatrix} \\
&= \underbrace{\begin{pmatrix} D_{i,t}^{-T} \\ \vdots \\ D_{i,t}^{0} \\ \vdots \\ D_{i,t}^{T-1} \end{pmatrix}^{\intercal}}_{\mathbf{D}_{i,t}} \underbrace{\begin{pmatrix} 0 \\ \vdots \\ CATT_{E_i,0} \\ \vdots \\ CATT_{E_i,T-1} \end{pmatrix}}_{f(E_i)}
\end{aligned}
$$

We can set $E\left[Y_{i,E_i-T} - Y_{i,E_i-T}^{\infty} | E_i\right]$ through $E\left[Y_{i,E_i-1} - Y_{i,E_i-1}^{\infty} | E_i\right]$ to zero by the no anticipation assumption. For index that is out of range e.g. $e + l < 0$ or $e + l > T$, even though $E\left[Y_{i,e+l} - Y_{i,e+l}^{\infty} | E_i = e\right]$ is undefined, the above expression is well-defined because the corresponding $D_{i,t}^{l} = 0$ for all t. Using this expression, the expression for projection coefficients μ_l's simplifies to

$$\left(\sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t}\mathbf{D}_{i,t}^{\intercal}\right]\right)^{-1}\left(\sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t}\mathbf{D}_{i,t}^{\intercal} f\left(E_i\right)\right]\right)$$

While $\mathbf{D}_{i,t}\mathbf{D}_{i,t}^{\intercal}$ is a diagonal matrix with diagonal entries equal to one for $t = E_i + l$ and zero otherwise, $\ddot{\mathbf{D}}_{i,t}\mathbf{D}_{i,t}^{\intercal}$ is not a diagonal matrix. This means that treatment effects at lags not equal to l might affect the coefficient on $D_{i,t}^{l}$. This proves that the projection coefficient on $D_{i,t}^{l}$ is a weighted average of $CATT_{e,l}$ for $1 \leq e \leq T$ and all $l' \neq l$.

Specifically, we can write μ_l as an affine average

$$\sum_{1\leq e\leq T-l} \omega_{e,l}^{l} CATT_{e,l} + \sum_{l'\neq l}\sum_{1\leq e\leq T-l'} \omega_{e,l'}^{l} CATT_{e,l'},$$

where $\sum_{-l_0\leq e\leq T-l_0} \omega_{e,l}^{l} = 1$ and for each $l' \neq l$, $\sum_{-l'\leq e\leq T-l'} \omega_{u,l'}^{l} = 0$. To see that $\sum_{-l\leq e\leq T-l} \omega_{e,l}^{l} = 1$, note that the sum of weights for $CATT_{e,l}$ across $1 \leq e \leq T-l$ in μ_l is

$$\begin{aligned}
&\left(\sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t}\mathbf{D}_{i,t}^{\intercal}\right]\right)^{-1}\left(\sum_{1\leq e\leq T-l} E\left[\ddot{\mathbf{D}}_{i,t} D_{i,e+l}^{l}\right]\right)\mathbf{e}_l^{\intercal}\\
=&\left(\sum_{t=0}^{T} E\left[\ddot{\mathbf{D}}_{i,t}\mathbf{D}_{i,t}^{\intercal}\right]\right)^{-1}\left(\sum_{0\leq t\leq T} E\left[\ddot{\mathbf{D}}_{i,t} D_{i,t}^{l}\right]\right)\mathbf{e}_l^{\intercal}
\end{aligned}$$

where $\mathbf{e}_l$ is a column vector with 1 in the entry corresponding to the entry of $D_{i,t}^{l}$ in $\mathbf{D}_{i,t}$, and 0 otherwise. That is, the above expression is the projection coefficient on $D_{i,t}^{l}$ from regressing $D_{i,t}^{l}$ on $\mathbf{D}_{i,t}$ and the unit and time fixed effects, which is just one. Similarly, for each $l' \neq l$, the sum of weights for $CATT_{e,l'}$ across $1 \leq e \leq T-l'$ in μ_l is projection coefficient on $D_{i,t}^{l'}$ from regressing $D_{i,t}^{l}$ on $\mathbf{D}_{i,t}$ and the unit and time fixed effects, which is zero. □

B Simulation design

We generate 1000 simulated datasets with $N = 1000$ and $T = 3$ according to the following DGP

$$Y_{i,t} = i + t + \sum_{e=1}^{3}\sum_{l=1-e}^{T-e} \delta_{e,l}\mathbf{1}\{E_i = e\}\cdot D_{i,t}^{l} + \epsilon_{i,t}$$

For each simulation, we draw E_i uniformly from $\{1, 2, 3\}$. We analyze the case where the DGP is a model of dynamic and non-stationary treatment effects. In particular, we set $\delta_{1,0} = 2$, $\delta_{1,1} = 18$, $\delta_{1,2} = 19$, $\delta_{2,0} = 3$,

$\delta_{2,1} = 4$, $\delta_{3,0} = 4$, $\delta_{e,l<0} = 0\ \forall e$ and $\epsilon_{i,t} \sim N(0,1)$. For each simulation, we estimate

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{l=-1}^{2} \mu_l D_{i,t}^l + \epsilon_{i,t}$$

While $\delta_{u,-1} = 0\ \forall u$ and hence $CATT_{e,-1} = 0$ for all e, the estimates $\hat{\mu}_{-1}$'s are usually negative as shown in Figure 1, which would suggest a negative pre-trend.

C Including covariates and doubly robust scores for event studies

When there are covariate-specific time trends, we need to modify the unconditional parallel trends assumption to the following one. Denote the confounding covariates by $\mathbf{X}_{i,t}$.

Assumption 5. Conditional parallel trends in baseline outcome. $E\left[Y_{i,t}^{\infty} - Y_{i,s}^{\infty} | E_i = e, \mathbf{X}_{i,t}\right]$ *is the same for all e, i.e., equal to* $E\left[Y_{i,t}^{\infty} - Y_{i,s}^{\infty} | \mathbf{X}_{i,t}\right]$.

We maintain the no anticipation assumption, but again allow the treatment effects to be dynamic for all cohorts and heterogenous i.e. non-stationary.

The causal parameter of interest is still

$$\theta_0 \coloneqq CATT_{e,l}$$

Suppose we know the covariate-specific time trends $g_0^{\infty}(\mathbf{X}_{i,t}) \coloneqq E\left[Y_{i,t}^{\infty} - Y_{i,0}^{\infty} \mid \mathbf{X}_{i,t}\right]$. Then we can form a consistent estimator for θ_0 by $E_N\left[Y_{i,t} - Y_{i,0} - g^{\infty}(\mathbf{X}_{i,t}) \mid E_i = t - l\right]$. This estimator is an extension to Heckman et al. (1997).

Suppose we know the probability of being in cohort $e = t - l$ conditional on $\mathbf{X}_{i,t}$, denoted by $m_0(\mathbf{X}_{i,t}) \coloneqq Pr\left\{E_i = t - l | \mathbf{X}_{i,t}\right\}$, and the probability of not having received treatment by time t conditional on $\mathbf{X}_{i,t}$, denoted by $n_0(\mathbf{X}_{i,t}) \coloneqq Pr\left\{E_i > t | \mathbf{X}_{i,t}\right\}$, as well as the share of cohort e, $m_0 \coloneqq Pr\left\{E_i = t - l\right\}$. Then we can form a consistent estimator for θ_0 by

$$\mathbb{E}_N\left[\frac{\mathbf{1}\left\{E_i = t - l\right\}}{m_0}(Y_{i,t} - Y_{i,0}) - \frac{\mathbf{1}\left\{E_i > t\right\} m_0(\mathbf{X}_{i,t})}{m_0 \cdot n_0(\mathbf{X}_{i,t})}(Y_{i,t} - Y_{i,0})\right]$$

This inverse propensity score reweighted estimator is an extension to Abadie (2005). A similar version has been proposed by Callaway and Sant'Anna (2018).

Collect

$$\eta_0(\mathbf{X}_{i,t}) \coloneqq \left(E\left[Y_{i,t}^{\infty} - Y_{i,0}^{\infty} \mid \mathbf{X}_{i,t}\right], Pr\left\{E_i = t-l|\mathbf{X}_{i,t}\right\}, Pr\left\{E_i > t|\mathbf{X}_{i,t}\right\}, Pr\left\{E_i = t-l\right\}\right)^{\intercal}$$

Let $\eta(\mathbf{X}_{i,t}) \coloneqq (g^{\infty}(\mathbf{X}_{i,t}), m(\mathbf{X}_{i,t}), n(\mathbf{X}_{i,t}), m)^{\intercal}$ denote the nuisance parameter with true value $\eta_0(\mathbf{X}_{i,t})$. In practice, it is likely that we make mistakes in estimating $\eta(\mathbf{X}_{i,t})$. We may employ the following doubly robust score as an extension to the doubly robust score derived for cross-sectional models. Define

$$\psi(W_i, \theta, \eta) \coloneqq \frac{\mathbf{1}\{E_i = t-l\}}{m}(Y_{i,t} - Y_{i,0} - g^{\infty}(\mathbf{X}_{i,t})) - \frac{\mathbf{1}\{E_i > t\}\, m(\mathbf{X}_{i,t})}{m \cdot n(\mathbf{X}_{i,t})}(Y_{i,t} - Y_{i,0} - g^{\infty}(\mathbf{X}_{i,t})) - \theta\frac{\mathbf{1}\{E_i = t-l\}}{m}$$

where $W_i \coloneqq (Y_{i,t}, Y_{i,0}, \mathbf{X}_{i,t})$.

We can show that the score satisfies the identification condition $E[\psi(W_i, \theta_0, \eta_0)] = 0$ and the Neyman orthogonality condition $\partial_\eta E[\psi(W_i, \theta_0, \eta)]\Big|_{\eta=\eta_0} = 0$. This score is robust to small mistakes in $\eta(\mathbf{X}_{i,t})$. The estimation can be done by K-fold cross-fitting as described in Chernozhukov et al. (2017).

For $\mathbf{X}_{i,t} = 1$, we are back to the unconditional parallel trends assumption. The DID estimator with $s = 0$ and $C = \{c : t < c\}$ is the root θ of the above doubly robust score in the sample, $\mathbb{E}_N[\psi(W_i, \theta, \widehat{\eta})] = 0$, where $\widehat{\eta}$ is the sample analog of η_0 .