# Causal Inference using Difference-in-Differences

## Lecture 13: Challenges with Treatments Turning On-and-Off

Pedro H. C. Sant'Anna
Emory University

January 2025

# Introduction

- We have covered a great variety of DiD designs with multiple periods and groups:
  - With and without covariates;
  - Different comparison groups;
  - Triple-Differences.
- All these embraced treatment effect heterogeneity and did not restrict treatment effect dynamics beyond the required PT.
- However, we exclusively focus in setups that treatment does not turn-off.

What if treatment can turn on and off?

What are the big challenges here?

What are the alternatives we have?

# Setup

# Setup

## Framework

# Framework

- There are $n$ units available, $i = 1, 2, \ldots, n$.
- There are several time periods available, $t = 1, 2, \ldots, T$.
- Units may be exposed to treatment at different points in time, and once treated, units can also turn off their treatment.
- Units are also allowed to "re-enroll" into the treatment later on.
- Let $D_{it}$ be a dummy variable that equals 1 if unit $i$ has "treatment turned on" at period $t$, and equals 0 otherwise.
- Let's assume that nobody is treated at time $t = 1$.
- There may be a set of units that have never experienced treatment until $t = T$.

# Setup

How do we define treatment groups in this context?

# Treatment sequences

- In contrast to the staggered setup, here we cannot fully characterize treatment groups (and potential outcomes) by the treatment adoption.
- We will talk about groups defined by **treatment sequences** as in Robins (1986).
- Let $\mathbf{d} = (d_1, d_2, d_3, \ldots, d_T)$ be a treatment sequence from $t = 1$ to $t = T$, where each $d_t \in \{0, 1\}$.
  - $d_t = 0$ meaning "is untreated" in period $t$; $d_t = 1$ meaning "is treated" in period $t$.
- Examples:
  - $\mathbf{d} = (0, 1, 0, 0)$ means untreated in period $t = 1$, then turn treatment on in period $t = 2$, but turn it off in period $t = 3$ and remain untreated in $t = 4$.
  - $\mathbf{d} = (0, 0, 0, 1)$ means untreated in periods $t = 1$, $t = 2$, and $t = 3$, then turn treatment on in period $t = 4$.

# Treatment sequences

- With $T$ time periods and nobody treated in $t = 1$, there are $2^{T-1}$ different treatment sequences that one can take.
- With $T = 4$, this is $2^{4-1} = 8$. With $T = 5$, $2^{5-1} = 16$.
- This is the number of potential treatment paths. Not all of them may be observed in the data.
- Denote the set of possible treatment sequences (those with strictly positive probability) as $\mathcal{D} \subseteq \{0, 1\}^T$ .
- Henceforth, let $D_{it} = 1$ if unit $i$ is treated in time $t$, and $D_{it} = 0$ otherwise.

# Treatment sequences and Treatment groups

- Let define groups depending on treatment sequences.
- Let $G_i \in \mathcal{D}$ denote the treatment group that unit $i$ belongs to.
  - We can "label/index" each value in $\mathcal{D}$ so $G_i$ take values on $\{1, 2, \ldots, |\mathcal{D}| - 1, \infty\}$, where $|\mathcal{D}|$ denotes the cardinality of $\mathcal{D}$ (its "size"). However, the "value" of these indexes would not have a clear interpretation in this general setup with treatment on and off.
  - It is arguably useful to denote the "never-treated" group $G_i = \infty = (0, 0, \ldots, 0)$.
- Let also define $G_i^{start}$ denote the first-period which unit $i$ is treated, i.e., $G_i^{start} = \min\{t \in \{2, 3, \ldots, T\} : D_{it} = 1, \text{ and } D_{it} = 0\}$. If unit $i$ is never-treated until $t = T$, we set $G_i^{start} = \infty$.
- Note that $G_i^{start} \in \mathcal{G}^{start} = \{2, 3, \ldots, T, \infty\}$, so the cardinality of $\mathcal{G}^{start}$ is substantially smaller than $\mathcal{D}$.

# Treatment sequences and Treatment groups

- **Example**: Consider a case with $T = 4$ and we have eight different treatment groups:
  1. $\mathbf{d} = (0, 0, 0, 0)$: units that remain untreated in all periods.
  2. $\mathbf{d} = (0, 1, 1, 1)$: units that are treated for the first time in $t = 2$ and stays treated.
  3. $\mathbf{d} = (0, 1, 1, 0)$: units that are treated for the first time in $t = 2$, remain treated in $t = 3$, but turn treatment off in $t = 4$.
  4. $\mathbf{d} = (0, 1, 0, 0)$: units that are treated for the first time in $t = 2$, turn off treatment in $t = 3$ and remain untreated until $t = 4$.
  5. $\mathbf{d} = (0, 1, 0, 1)$: units that are treated for the first time in $t = 2$, turn off treatment in $t = 3$, and turn it back on in $t = 4$.
  6. $\mathbf{d} = (0, 0, 1, 1)$: units that are treated for the first time in $t = 3$ and stays treated.
  7. $\mathbf{d} = (0, 0, 0, 1)$: units that are treated for the first time in $t = 4$.
  8. $\mathbf{d} = (0, 0, 1, 0)$: units that are treated for the first time in $t = 3$, but turn it off in $t = 4$.

# Treatment sequences and Treatment groups

- **Example**: Consider a case with $T = 4$ and we have seven different treatment groups:

1. $\mathbf{d} = (0, 0, 0, 0)$: $G_i^{start} = \infty$;
2. $\mathbf{d} = (0, 1, 1, 1)$: $G_i^{start} = 2$;
3. $\mathbf{d} = (0, 1, 1, 0)$: $G_i^{start} = 2$;
4. $\mathbf{d} = (0, 1, 0, 0)$: $G_i^{start} = 2$;
5. $\mathbf{d} = (0, 1, 0, 1)$: $G_i^{start} = 2$;
6. $\mathbf{d} = (0, 0, 1, 1)$: $G_i^{start} = 3$;
7. $\mathbf{d} = (0, 0, 0, 1)$: $G_i^{start} = 4$;
8. $\mathbf{d} = (0, 0, 1, 0)$: $G_i^{start} = 3$.

# Setup

Potential outcomes and parameters of interest

# Potential outcomes

- As you may have guessed, we need to talk about potential outcomes.
- Potential outcomes will be indexed by treatment sequences $\mathbf{d}$, as in Problem Set 1 (remember that?!)
- Let $Y_{i,t}(\mathbf{d})$ be the potential outcome for unit $i$, at time $t$, if this unit had a treatment sequence $\mathbf{d}$.
- We will denote $Y_{it}(\infty) = Y_{i,t}(0, 0, \ldots, 0)$ the "never-treated" potential outcomes.
- Observed outcomes at time $t$ are realized as

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

# Parameter of interest

- In this context, one is arguably interested in the average treatment effect of one treatment sequence $\mathbf{d}$ versus no-treatment, among units that actually received the treatment-sequence $\mathbf{d}$, at time $t$:

$$ATT(\mathbf{d}, t) = \mathbb{E}[Y_t(\mathbf{d}) - Y_t(\infty)|G = \mathbf{d}].$$

- This is very similar to the $ATT(g, t)$'s we have considered in staggered designs:
  - Group is now defined by treatment sequences instead of treatment starting date. Of course, these two coincide in staggered setups.
  - To be clear, there will be many more groups here than in the staggered setup.
- $ATT(\mathbf{d}, t)$ can be seen as the "building blocks" of the analysis: aggregating them into fewer parameters is definitely possible.

# How can we move forward?

# We need assumptions!

- Surprise, surprise: we need to make assumptions to move forward; see, e.g., de Chaisemartin and D'Haultfœuille (2024).

**Assumption (No-Anticipation)**

*For all units $i$, $Y_{i,t}(\mathbf{d}) = Y_{i,t}(\infty)$ for all groups in their pre-treatment periods, i.e., for all time periods $t$ before they are first-exposed to treatment, $t < g^{start}$.*

**Assumption ("Strong" Parallel Trends Assumption)**

*For all $\mathbf{d} \times \mathbf{d}' \in \mathcal{D}^2$ and all $t \in \{2, \ldots, T\}$,*

$$\mathbb{E}\left[Y_t(\infty)|G = \mathbf{d}\right] - \mathbb{E}\left[Y_{t-1}(\infty)|G = \mathbf{d}\right] = \mathbb{E}\left[Y_t(\infty)|G = \mathbf{d}'\right] - \mathbb{E}\left[Y_{t-1}(\infty)|G = \mathbf{d}'\right].$$

This version of the PTA rule out non-parallel pre-trends between any group, and also allows one to use "never-treated" and many other comparison groups (model is very over-identified).

# ATT(d,t) Estimands

- Identifications is very similar to the staggered treatment adoption setup of Callaway and Sant'Anna (2021) that we discussed in Lecture 12.
- In fact, when we have a never-treated group, we can show that for any $\mathbf{d} \in \mathcal{D}$ and any $t \geq g^{start}$, under the aforementioned assumptions,

$$ATT^{nev}_{unc}(\mathbf{d}, t) = \mathbb{E}[Y_t - Y_{g^{start}-1} | G = \mathbf{d}] - \mathbb{E}[Y_t - Y_{g^{start}-1} | G = \infty].$$

- We can also use the "not-yet-exposed" to treatment as a comparison group:

$$ATT^{ny}_{unc}(\mathbf{d}, t) = \mathbb{E}[Y_t - Y_{g^{start}-1} | G = \mathbf{d}] - \mathbb{E}[Y_t - Y_{g^{start}-1} | G^{start} > t, G \neq \mathbf{d}].$$

- This looks very similar to the staggered setup of Lecture 12.
- The difference is now we need to be careful with the definition of "treatment groups".

# Where is the catch?

# Complications

- In terms of identification, things are relatively "standard".
- However, with many more treatment groups than the staggered case, effectively summarizing the effect of treatment is much more challenging.
- Even when $T = 4$, as in our example before, we have 7 different treated groups in this non-staggered case, compared to 3 in the staggered setup.
- If you count the $(\mathbf{d}, t)$ pairs with $t \geq g^{start}$, there are 17 parameters to be identified (and estimated)!
- If this were a staggered setup, we would have only 6 (g,t) (post-treatment) pairs.

Estimating these many parameters with some degree of precision is hard!
Even when possible, effectively communicating the TE heterogeneity is possibly harder!

# What about aggregation?

# Aggregation step

- Just like in the staggered setup, we can follow Callaway and Sant'Anna (2021) and propose different aggregation procedures to try to bypass these complications.
- This is what de Chaisemartin and D'Haultfœuille (2024) proposes.
- Let us describe their "event-study-type" aggregation, but using our notation.
  - They have a somehow non-standard setup with non-iid observations, and that implicitly implies that there are "multiple populations", making the notation more cumbersome. They also assume that treatment status is fixed. Here, we follow the more standard "sampling-perspective" to foster intuition.

# Aggregated parameters

- We want to first aggregated $ATT(\mathbf{d}, t)$ according to the time a unit is first exposed to treatment, $g^{start}$.
- Let $Start(\mathbf{d}) = min\{s \geq 2 : d_s = 1, d_{s-1} = 0, \mathbf{d} = (d_1, d_2, \ldots, d_T) \in \mathcal{D}\}$.
- For a fixed $g^{start} \in \mathcal{G}^{start}$ and fixed $t \geq g^{start}$, let

$$\widetilde{ATT}(g^{start}, t) = \sum_{\mathbf{d} \in \mathcal{D}} 1\{Start(\mathbf{d}) = g^{start}, t \geq g^{start}\} \mathbb{P}\left(G = \mathbf{d} | Start(\mathbf{d}) = g^{start}\right) ATT(\mathbf{d}, t)$$

- $\widetilde{ATT}(g^{start}, t)$ is "similar" to the $ATT(g, t)$ in the staggered designs, except that it is harder to interpret under our assumptions.
- $\widetilde{ATT}(g^{start}, t)$ is a "population-weighted average" of the $ATT(\mathbf{d}, t)$'s across different treatment paths that share the same treatment starting date $g^{start}$.

# Aggregated parameters

- Mechanically speaking, when it comes to estimation, this is equivalent to treating $G^{start}$ as your treatment group variable, though you need to remember that the interpretation is much different.
- Example: Let $g^{start} = 2$ and $t = 4$. $\widetilde{ATT}(2, 4)$ would aggregate the following groups:
  1. $\mathbf{d} = (0, 1, 1, 1)$: $G_i^{start} = 2$;
  2. $\mathbf{d} = (0, 1, 1, 0)$: $G_i^{start} = 2$;
  3. $\mathbf{d} = (0, 1, 0, 0)$: $G_i^{start} = 2$;
  4. $\mathbf{d} = (0, 1, 0, 1)$: $G_i^{start} = 2$;
- Under our identification assumptions, how could we actually interpret such parameters?
- For time $t = g^{start}$, this would be fine, but for $t > g^{start}$, this is definitely more complicated.

# Aggregated parameters

- At best, $\widetilde{ATT}(g^{start}, t)$ can be interpreted as the Average treatment effect at time $t$ from switching from never being treated to being treated for the first time at $g^{start}$, among units that are first-treated in $g^{start}$.

- If you are not comfortable with this type of interpretation and are interested in comparing more general treatment sequences, it is worth being careful.

# Aggregated parameters

- In some cases, there may still be several "treatment-starting-date" groups, and one may desire to further aggregate the $\widetilde{ATT}(g^{start}, t)$ to get an "event-study"-type curve.
- Towards this end, we will follow the same steps as in Lecture 12 but using $\widetilde{ATT}(g^{start}, t)$ in place of $ATT(g, t)$.
- Consider the probability of a unit being in cohort $G^{start} = g^{start}$ given that it is among the units that at least $e = t - g^{start}$ periods have passed since they were 'first-treated":

$$w(g^{start}; e) \equiv P\left(G^{start} = g^{start} | \text{First-treated at least } e \text{ periods ago}\right)$$

- Now, we can define the "event-study"-type parameter

$$\widetilde{\theta}_D(e) = \sum_{g^{start}=2}^{T} \mathbf{1}\{g^{start} + e \leq T\} \cdot w(g^{start}; e) \cdot \widetilde{ATT}(g^{start}, g^{start} + e)$$

Unfortunately, $\widetilde{\theta}_D(e)$ does not represent an average effect with respect to the length of exposure, because we aggregate across different treatment paths with different exposure patterns.

# Aggregated parameters of interest: Questions that I still have

- What economic questions $\widetilde{\theta}_D(e)$ can answer, under our maintained assumptions?
- What if we strengthen the assumptions?
- Are there other feasible aggregation schemes that highlight other dimensions of treatment effect heterogeneity?
- Is difference-in-differences the right tool here? What about the setup Bojinov, Rambachan and Shephard (2021)? What about Robins, Hernán and Brumback (2000), Blackwell (2013), Imai and Ratkovic (2015), Blackwell and Glynn (2018)?

# Aggregated parameters of interest: Questions that I still have

- What economic questions $\widetilde{\theta}_D(e)$ can answer, under our maintained assumptions?
- What if we strengthen the assumptions?
- Are there other feasible aggregation schemes that highlight other dimensions of treatment effect heterogeneity?
- Is difference-in-differences the right tool here?
- What about the setup Bojinov et al. (2021)? What about Robins et al. (2000), Blackwell (2013), Imai and Ratkovic (2015), Blackwell and Glynn (2018)? Worth thinking!

# What about aggregation?

What if we normalize by the number of treated periods?

# Aggregated parameters: normalizing by the number of exposed periods

- In an attempt to get more interpretable parameters, de Chaisemartin and D'Haultfœuille (2024) proposes to normalize $\widetilde{\theta}_D(e)$ by by number of treated periods.
- The idea is to treat $\widetilde{\theta}_D(e)$ as an "intention to treat" parameter, and then, by diving it with a "first-stage", we could get something that is "more interpretable".
- To operationalize this idea, for $t \geq g^{start}$, let $FS(g^{start}, t) = \mathbb{E}[\sum_{s=1}^{t} D_s | G^{start} = g^{start}]$, where the expectation is taken with respect to the units.
  - This computes the average number of treated periods from time $s = 1$ until time $s = t$ among units who were first treated in period $g^{start}$.
- Likewise, let $FS(e) = \sum_{g^{start}=2}^{T} \mathbf{1}\{g^{start} + e \leq T\} \cdot w(g^{start}; e) \cdot FS(g^{start}, g^{start} + e)$

# Aggregated parameters: normalizing by the number of exposed periods

- In order to get an "event-study average total effect per unit of treatment", de Chaisemartin and D'Haultfœuille (2024) propose to focus on

$$\widetilde{\theta}_D^{per\ unit}(e) = \frac{\widetilde{\theta}_D(e)}{FS(e)}.$$

- It is worth noticing that de Chaisemartin and D'Haultfœuille (2024) assume that treatment sequences (and therefore treatment status) are fixed, i.e., deterministic, so there is no uncertainty coming from estimating the denominator.
- Another caveat is that interpreting this as "per unit of treatment" somehow implicitly relies on us ignoring the substitution/complementarity effects of sequential treatments.
- This may be reasonable under a potentially strong assumption of no-carryover, i.e., treatment today only affects outcomes today. This was assumed in de Chaisemartin and D'Haultfœuille (2020).

# Take-way messages

# DiD procedures with treatment on and off

- Once we establish the “building blocks”, we can leverage the principles discussed in Lecture 12 to identify the group and time-specific treatment effects.
- The main difference here is that we must be careful when defining treatment groups.
- Since there are potentially many groups, one may be interested in aggregation schemes, and this is where things get very complicated.
- Interpretation of these aggregated causal parameters is much harder and full of caveats.
- Unless we are comfortable with stronger assumptions, DiD tools may be less than ideal in this setup. Of course, this is my personal view only.

# References

**Blackwell, Matthew**, "A Framework for Dynamic Causal Inference in Political Science," *American Journal of Political Science*, apr 2013, *57* (2), 504–520.

_ **and Adam N. Glynn**, "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables," *American Political Science Review*, 2018, *112* (4), 1036–1049.

**Bojinov, Iavor, Ashesh Rambachan, and Neil Shephard**, "Panel experiments and dynamic causal effects: A finite population perspective," *Quantitative Economics*, 2021, *12* (4), 1171–1196.

**Callaway, Brantly and Pedro H. C. Sant'Anna**, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

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

__ and __ , “Difference-in-Differences Estimators of Intertemporal Treatment Effects,” *The Review of Economics and Statistics*, 2024, *Forthcoming*, 1–45.

**Imai, Kosuke and Marc Ratkovic**, “Robust Estimation of Inverse Probability Weights for Marginal Structural Models,” *Journal of the American Statistical Association*, 2015.

**Robins, James**, “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect,” *Mathematical Modelling*, January 1986, *7* (9), 1393–1512.

**Robins, James M, Miguel Ángel Hernán, and Babette Brumback**, “Marginal Structural Models and Causal Inference in Epidemiology,” *Epidemiology*, 2000, *11* (5), 550–560.