When Do Outcome-Driven Treatments Break Parallel Trends?

Zach Shahn

June 2022

1 Introduction

The parallel trends assumption required for validity of Difference in Differences (DiD) studies is strong and untestable. Sensitivity analyses for violations of parallel trends are available (Rambachan and Roth (2022); Shahn et al (2022)), and pre-treatment trend tests are flawed (Roth, 2020) yet useful empirical sanity checks. But still lacking is intuition for when parallel trends is likely to hold rooted in structural assumptions about the data generating process (DGP). In studies employing other identification strategies, such as backdoor confounding adjustment or instrumental variables, researchers and readers can reason about the plausibility of structural assumptions required for validity using tools such as Directed Acyclic Graphs (Greenland, Pearl, Robins, 1999) or Single World Intervention Graphs (Richardson and Robins, 2013). In this paper, we take an incremental step toward enabling similar reasoning about parallel trends assumptions in DiD studies.

In particular, a key question is under what circumstances it is a threat to parallel trends for treatment decisions to be based on past values of the outcome. Daw and Hatfield (2018) showed via simulation analyses that if the counterfactual untreated outcome follows a random walk autoregressive process, then parallel trends for a point exposure treatment is satisfied (though not after matching on pre-treatment outcome values). We expand on Daw and Hatfield (2018) to consider the impact of outcome-responsive treatment decisions in a wider range of data generating processes and study designs, including for time-varying treatments.

We consider three basic DGPs, which might be considered platonic ideals of DGPs conducive to parallel trends. In the first DGP, similar to Daw and Hatfield (2018), counterfactual untreated outcomes follow a random walk in which initial values depend on an unobserved baseline confounder but subsequent jumps are independent of the confounder conditional on the previous value of the counterfactual outcome. In the second DGP, counterfactual untreated outcomes are generated from a hidden Markov model (HMM) in which counterfactual untreated outcome means follow a random walk (with the initial value dependent on an unobserved baseline confounder) and the counterfactual
untreated outcomes are drawn from normal distributions about those means. In the third DGP, a baseline confounder has a constant direct additive effect on the counterfactual untreated outcomes at each time-step.

We also consider several study designs (i.e. approaches to ‘control group’ and cohort selection) under each of the above DGPs. In the time varying treatment setting, parallel trends posits that at each time $t$ the mean counterfactual untreated outcome trajectories are parallel from $t$ onward in those who started treatment at $t$ and a ‘control group’. Callaway and Sant’anna (2021) describe two candidate control groups—the ‘not yet treated’ and the ‘never treated’. The ‘not yet treated’ have yet to receive treatment at time $t$ but may or may not receive treatment in the future. The ‘never treated’ never start treatment for the duration of followup. The analyst will use a different estimator depending on whether they believe parallel trends holds relative to the ‘not yet treated’ or ‘never treated’ control group. Sometimes, analysts will only include in their cohort units that eventually start treatment, perhaps because it is too challenging to collect data on the full cohort. In these studies of the eventually treated, the control group is ‘later treated’.

We explore via simulation studies whether parallel trends holds under each combination of DGP (random walk, HMM, or constant direct additive confounding), study design (never treated, not yet treated, or later treated), and outcome responsiveness of treatment (yes or no). We might summarize these results as follows. Outcome-influenced treatments break parallel trends under all study designs under the constant additive confounding and HMM DGPs due to regression to the mean. Under the random walk DGP, there is no regression to the mean. However, under the random walk DGP, outcome-influenced treatments still break parallel trends under ‘never treated’ and ‘later treated’ control group designs due to selection on post-treatment variables. Under a random walk DGP with a not yet treated control group, parallel trends is preserved even if past outcomes influence treatment as there is no regression to the mean or selection on post-treatment variables (consistent with Daw and Hatfield (2018)).

By way of illustration, we apply the lessons learned from our simulations to consider a real application investigating the effects of electing reformer prosecutors on crime rates (Agan et al, 2022). We reason about the bias that might result if voters are more likely to elect reformer prosecutors when recent crime is low, otherwise viewing reform as too risky. Our discussion of this application is purely illustrative as we lack expertise in criminology and political science.

We interpret the upshot of our simulation results to be that parallel trends is typically not a credible assumption when treatments are influenced by past outcomes. Since timing of treatment initiation is frequently influenced by past outcomes when the treatment is targeted at the outcome, we believe DiD is perhaps generally better suited for studying unintended consequences of interventions (as a rule of thumb).
2 The Data Generating Processes

2.1 Random Walk With Starting Point Confounded by Baseline $U$

Consider a data generating process in which a baseline confounder $U$ has an additive effect on the initial outcome value, and the counterfactual untreated outcome follows a random walk thereafter. In the DAG for this DGP, there are no arrows from $U$ to outcomes at post-baseline timepoints. We consider two versions of this DGP—one in which treatments are impacted by past outcomes and one in which they are not.

**RW-DGP 1: Random Walk With Treatments Influenced By Past Outcomes**

\[
U \sim N(0, 1) \\
Y_0 \sim N(U, 1);
Y_1(0) \sim N(Y_0(0), 1); \ldots; Y_K(0) \sim N(Y_{K-1}(0), 1) \\
Pr(A_t = 1|A_{t-1} = 0) = \logit^{-1}(U + Y_{t-1}) \text{ for } t \in 1, \ldots, K \\
Pr(A_t = 1|A_{t-1} = 1) = 1
\]  

**DGP2: Random Walk With Treatments Uninfluenced by Past Outcomes**

\[
U \sim N(0, 1) \\
Y_0 \sim N(U, 1);
Y_1(0) \sim N(U, 1); \ldots; Y_K(0) \sim N(U, 1) \\
(A_t = 1|A_{t-1} = 0) = \logit^{-1}(U) \text{ for } t \in 1, \ldots, K \\
Pr(A_t = 1|A_{t-1} = 1) = 1
\]

2.2 Constant Direct Additive Effect of Baseline $U$

Consider a data generating process where at each timepoint the counterfactual untreated outcome is drawn from a normal distribution with mean a function of baseline confounder $U$ and time, with $U$ entering the function additively. The DAG for this DGP has arrows from $U$ to the outcome at each timestep. We again consider two versions of this DGP—one in which treatments are impacted by past outcomes and one in which they are not.

**CDA-DGP1: Constant Direct Additive Confounding With Treatments Influenced By Past Outcomes**

\[
U \sim N(0, 1) \\
Y_0 \sim N(U, 1);
Y_1(0) \sim N(Y_0(0), 1); \ldots; Y_K(0) \sim N(Y_{K-1}(0), 1) \\
(A_t = 1|A_{t-1} = 0) = \logit^{-1}(U + Y_{t-1}) \text{ for } t \in 1, \ldots, K \\
Pr(A_t = 1|A_{t-1} = 1) = 1
\]  

**CDA-DGP2: Constant Direct Additive Confounding With Treatments Uninfluenced By Past Outcomes**

\[
U \sim N(0, 1) \\
Y_0 \sim N(U, 1);
Y_1(0) \sim N(U, 1); \ldots; Y_K(0) \sim N(U, 1) \\
(A_t = 1|A_{t-1} = 0) = \logit^{-1}(U) \text{ for } t \in 1, \ldots, K \\
Pr(A_t = 1|A_{t-1} = 1) = 1
\]
Figure 1: DAG illustrating RW-DGP1 and RW-DGP1
Figure 2: DAG illustrating CDA-DGP1 and CDA-DGP1

CDA-DGP2: Constant Direct Additive Confounding With Treatments Uninfluenced By Past Outcomes

\[ U \sim N(0, 1) \]
\[ Y_0 \sim N(U, 1); Y_1(0) \sim N(U, 1); \ldots; Y_K(0) \sim N(U, 1) \]
\[(A_t = 1|A_{t-1} = 0) = \text{logit}^{-1}(U) \text{ for } t \in 1, \ldots, K \]
\[ Pr(A_t = 1|A_{t-1} = 1) = 1 \]

2.3 Hidden Markov Model With Initial Mean Confounded by Baseline \( U \)

Consider a data generating process in which a baseline confounder \( U \) has an additive effect on the mean initial outcome value, and the counterfactual untreated outcome means follow a random walk thereafter, with the counterfactual untreated outcomes drawn from normal distributions about these means. Such a process is a standard HMM. We consider two versions of this DGP—one in which treatments are impacted by past outcomes and one in which they are not.
HMM-DGP 1: HMM With Treatments Influenced By Past Outcomes

\[ U \sim N(0, 1) \]
\[ \mu_0 \sim N(U, 1); \mu_1(0) \sim N(\mu_0(0), 1); \ldots; \mu_K(0) \sim N(\mu_{K-1}(0), 1) \]
\[ Y_t(0) \sim N(\mu_t, 1) \]
\[ Pr(A_t = 1 | A_{t-1} = 0) = \text{logit}^{-1}(U + Y_{t-1}) \text{ for } t \in 1, \ldots, K \]
\[ Pr(A_t = 1 | A_{t-1} = 1) = 1 \]

HMM-DGP 2: HMM With Treatments Uninfluenced By Past Outcomes

\[ U \sim N(0, 1) \]
\[ \mu_0 \sim N(U, 1); \mu_1(0) \sim N(\mu_0(0), 1); \ldots; \mu_K(0) \sim N(\mu_{K-1}(0), 1) \]
\[ Y_t(0) \sim N(\mu_t, 1) \]
\[ Pr(A_t = 1 | A_{t-1} = 0) = \text{logit}^{-1}(U) \text{ for } t \in 1, \ldots, K \]
\[ Pr(A_t = 1 | A_{t-1} = 1) = 1 \]

3 Study Designs

We consider staggered adoption settings with binary treatment \( A_k \) at each time point and \( A_k = 1 \) for all \( k > t \) if \( A_t = 1 \). Let \( K \) denote the time of end of followup. We denote by \( T \) the time of treatment initiation with \( T = \infty \) if treatment is never initiated. Let \( Y_k(\infty) \) denote the counterfactual outcome under no treatment through \( k \), assuming that there is ‘no anticipation’, i.e. future treatments do not influence past outcomes.

The not yet treated control group design assumes, for each \( t, k \geq t \),
\[ E[Y_k(\infty) - Y_{k-1}(\infty)|T = t] = E[Y_k(\infty) - Y_{k-1}(\infty)|T > t]. \tag{7} \]

The never treated control group design assumes, for each \( t, k > t \),
\[ E[Y_k(\infty) - Y_{k-1}(\infty)|T = t] = E[Y_k(\infty) - Y_{k-1}(\infty)|T = \infty]. \tag{8} \]

The later treated control group design assumes, for each \( t, k > t \),
\[ E[Y_k(\infty) - Y_{k-1}(\infty)|T = t] = E[Y_k(\infty) - Y_{k-1}(\infty)|t < T \leq K]. \tag{9} \]

4 Simulation Study Results

Under DGPs (1)-(6) with \( K = 5 \), we generated treatments and counterfactual untreated outcomes for 10 million subjects. In Figures 4, we present plots
Figure 3: DAG illustrating HMM-DGP1 and HMM-DGP1
of expected counterfactual outcome increments in the treated (i.e. \(E[Y_k(\infty) - Y_{k-1}(\infty)|T = m]\) for \(k > m\)), not yet treated (i.e. \(E[Y_k(\infty) - Y_{k-1}(\infty)|T > m]\) for \(k > m\)), later treated (i.e. \(E[Y_k(\infty) - Y_{k-1}(\infty)|K > T > m]\) for \(k > m\)), and never treated (i.e. \(E[Y_k(\infty) - Y_{k-1}(\infty)|T = \infty]\) for \(k > m\)) for each DGP.

Under parallel trends between the treated and a given control group, expected untreated counterfactual increments among the treated would be equal to those among the control group at each time point after treatment, i.e. the expected counterfactual increment trajectories of the treated and the control group should be overlaid.

Expected increments were uniformly equal for all DGPs and study designs when treatments were not influenced by past outcomes, indicating that parallel trends held as expected. Thus we only display results under DGPs in which outcomes influence future treatments. Violations of parallel trends where they appeared were driven by two ‘forces’: regression to the mean and selection on future treatment. Below, we explain how these forces led to the results under each simulation setting. In these explanations, we always assume that higher values of prior outcomes make immediately subsequent treatment more likely, as in the DGPs we simulated, but the same general reasoning can straightforwardly be applied to other scenarios.

First, we consider CDA-DGP1 and explain the results in Figure 4. Treated units tend to have higher prior outcome values conditional on \(U\), and these outcome values regress toward mean \(U\) at the next time point, pushing initial increments in the treated downward. Conversely, the not yet treated have on average lower prior outcomes given \(U\) which regress upward toward mean \(U\) at the next time point, pushing initial increments in the not yet treated upward. The later treated are similar to the not yet treated, except they also tend to have higher outcomes at subsequent time steps (when they will be treated), amplifying the upward pressure on initial increments in this group compared to the not yet treated. The never treated do not exhibit regression to the mean in either direction, as their outcomes are on average lower at all time points that precede treatments. Thus, the never treated increments are also not equal to the treated (because the treated increments are impacted by regression to the mean) as required to satisfy parallel trends. At the final increment, the dynamics are a bit different because the final outcome follows the final treatment and there is therefore no selection pressure on it from future treatment values. In the never treated, the final increment is positive because the negative pressure on the second-to-last outcome from the requirement that the final treatment value be 0 is removed at the last time step. In the later treated, the final increment is negative because the positive pressure on the second-to-last outcome from the requirement that the final treatment be 1 is removed at the last time step. This combination of regression to the mean and selection on future treatment values fully explains the results in Figure 4.

The results under HMM-DGP1 displayed in Figure 4 are also driven by a combination of regression to the mean and selection on future treatment. Increments in the treated and not yet treated follow a similar pattern as under CDA-DGP1, with similar explanation. In units treated at \(t\), the prior \(\text{outcome
residual’ $Y_t(\infty) - \mu_t$ will be higher on average. Without any further selection pressure at time $t+1$, $\mu_{t+1}$ will on average be equal to $\mu_t$ and $Y_{t+1}(\infty) - \mu_{t+1}$ will on average regress toward 0, leading to a negative initial increment. Conversely, for the not yet treated, the prior ‘outcome residual’ $Y_t(\infty) - \mu_t$ will be lower on average. Without any further selection pressure at time $t+1$, $\mu_{t+1}$ will on average be equal to $\mu_t$ and $Y_{t+1}(\infty) - \mu_{t+1}$ will on average regress toward 0, leading to a positive initial increment. In the HMM DGP, selection on future treatment values plays more of a role in the never treated and later treated control groups than it did in the CDA DGP. In HMM-DGP1, for values $x' > x$, a subject not treated through $t$ is more likely to be never treated if $\mu_t = x$ and $\mu_{t-1} = x'$ than if $\mu_t = x'$ and $\mu_{t-1} = x$. This is because under HMM-DGP1 $\mu_t$ influences future counterfactual untreated outcomes (and therefore treatment initiation), while $\mu_{t-1}$ does not except through $\mu_t$. The asymmetry that a decline in $\mu$ is more likely than an increase of the same magnitude explains the negative increments among the never treated. (Note that if outcomes have delayed effects on treatment initiation this dynamic could reverse. In HMM-DGP1, treatment initiation is only influenced by the most recent outcome. But in a more general setting represented by the DAG in Figure 3 outcomes can influence treatments at all future times. If the effect of an outcome on treatment grows with time, the dynamic described above would reverse and we would see positive increments in the never treated.) The later treated are the converse of the never treated. For values $x' > x$, a subject not treated through $t$ is more likely to be later treated if $\mu_t = x'$ and $\mu_{t-1} = x$ than if $\mu_t = x$ and $\mu_{t-1} = x'$. This asymmetry leads to positive increments in the later treated. Similar dynamics apply in the final increment under HMM-DGP1 as under CDA-DGP1, where selection pressure from future treatments is removed due to the final outcome following the final treatment.

In the random walk DGP RW-DGP1, there is no regression to the mean at play because each counterfactual untreated outcome is drawn from a distribution centered at the counterfactual untreated outcome that preceded it. Therefore, of the two forces we have identified that break parallel trends, only selection on future treatment values is applicable to RW-DGP1. Since the not yet treated control group does not select on future treatment values, it does not violate parallel trends in this DGP, as is evident in Figure 4. Under a ‘later treated’ control group study design, units in the control group have systematically higher future counterfactual untreated outcome trajectories than those merely not yet treated, leading to positive increments. Conversely, never treated units have systematically lower future counterfactual untreated outcome trajectories than those merely not yet treated, leading to negative increments.
Expected Counterfactual Untreated Increments From Time 1

Expected Counterfactual Untreated Increments From Time 2

Expected Counterfactual Untreated Increments From Time 3

Figure 4: Counterfactual untreated outcome increments under DGP-RW1 in the treated, never treated, not yet treated, and later treated among those not treated before time $k=1$ (top), $k=2$ (middle), and $k=3$ (bottom)
Figure 5: Counterfactual untreated outcome increments under DGP-CDA1 in the treated, never treated, not yet treated, and later treated among those not treated before time $k=1$ (top), $k=2$ (middle), and $k=3$ (bottom)
Figure 6: Counterfactual untreated outcome increments under DGP-HMM1 in the treated, never treated, not yet treated, and later treated among those not treated before time $k=1$ (top), $k=2$ (middle), and $k=3$ (bottom)
5 An Illustrative Consideration of A Real Application

Agan et al. (2022) recently used a time-varying DiD design to investigate the effects on crime of electing progressive reformer District Attorneys. Due to challenges in collecting outcome data on all counties, they used a later treated control group design only including counties that eventually elect progressive reformer prosecutors. They found no significant effects on crime rates from electing progressive reformer prosecutors. We are not criminologists and do not pretend to offer any serious critiques of their paper. We just found it an interesting concrete example in which to consider the concepts we explored in simulations.

In particular, we would like to think through the potential implications for parallel trends if voters are more likely to elect reformer prosecutors if recent crime is low, otherwise viewing the proposition as too risky. Again, lacking expertise in criminology or political science, we do not claim that this hypothetical scenario reflects reality.

Under all DGPs we considered, counties that elect a reformer at a later time will systematically have lower future crime rates, which will push the trends in the later treated control group downward relative to the treated group. Further, under the constant direct additive confounding and HMM DGPs but not the random walk DGP, later treated counties that did not (yet) elect a reformer in a given year are more likely to have had abnormally high crime in that year, and expected regression to the mean in the following year will also push trends in the later treated control group downward relative to the treated group. Therefore, the estimated effect on crime of reformer prosecutors would be pushed upward. That is, we might expect this study design to make reformers appear more harmful than they are. Such considerations might inform sensitivity analyses.

6 Discussion

It seems significant that under most DGP/study design combinations we considered, influence of past outcomes on treatment decisions breaks parallel trends. The only exception was if the counterfactual untreated outcome follows a random walk and the study design employs a not yet treated control group. (While none of the DGPs considered is likely to be particularly realistic, the constant additive confounding and HMM DGPs are likely better approximations to reality than a random walk for most outcomes.) Usually, if an intervention is designed to address a particular problem, then timing of implementation of the intervention is responsive to the time-varying magnitude of the problem. This is suggestive that DiD designs are perhaps generally best suited to estimate unintended effects of interventions.

Of course, we did not consider every DGP that might result in parallel trends. However, we did include those with the (subjectively, to us) clearest structural justifications for parallel trends in the absence of outcome influenced
treatment decisions. We believe that to justify a parallel trends design a researcher should have in mind some approximate structural model (see Roth and Sant’Anna (2020) for some further discussion of this point), and given this model the researcher could apply reasoning similar to our own to assess the impact of outcome influenced treatments.

Perhaps future work can identify analytic approaches to remove or partially remove bias from violations of parallel trends due to outcome influenced treatment. Daw and Hatfield (2018) warned against adjustment for prior outcomes because they showed that it can induce bias where none previously existed in a random walk DGP. However, since this is actually the only DGP where there is no bias without adjustment and it might be considered the least realistic DGP, perhaps methods of adjustment for previous outcomes should be explored if they work well in other DGPs. In Shahn et al (2022), it is not possible to condition on the outcome immediately preceding treatment, but it is possible to condition on earlier outcomes. Also, perhaps reasoning about the likely impact of outcome influenced treatments can inform sensitivity analyses (Rambachan and Roth, 2022; Shahn et al, 2022).

7 References

Agan, Amanda, Jennifer L. Doleac, and Anna Harvey. "Prosecutorial Reform and Local Crime Rates." George Mason Law & Economics Research Paper 21-31 (2021).

Callaway, Brantly, and Pedro HC Sant’Anna. "Difference-in-differences with multiple time periods." Journal of Econometrics 225, no. 2 (2021): 200-230.

Daw, Jamie R., and Laura A. Hatfield. "Matching and regression to the mean in difference-in-differences analysis." Health services research 53, no. 6 (2018): 4138-4156.

Greenland, Sander, Judea Pearl, and James M. Robins. "Causal diagrams for epidemiologic research." Epidemiology (1999): 37-48.

Rambachan, Ashesh, and Jonathan Roth. A More Credible Approach to Parallel Trends. Working Paper, 2022.

Richardson, Thomas S., and James M. Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality." Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128, no. 30 (2013): 2013.

Roth, Jonathan. "Pre-test with caution: Event-study estimates after testing for parallel trends." Department of Economics, Harvard University, Unpublished manuscript (2019).
Roth, Jonathan, and Pedro HC Sant’Anna. "When is parallel trends sensitive to functional form?" arXiv preprint arXiv:2010.04814 (2020).

Shahn, Zach, Oliver Dukes, David Richardson, Eric Tchetgen Tchetgen, and James Robins. "Structural Nested Mean Models Under Parallel Trends Assumptions." arXiv preprint arXiv:2204.10291 (2022).