Survey Research Methods (2020)
Vol. 14, No. 2, pp. 153-158
doi:10.18148/srm/2020.v14i2.7723

ISSN 1864-3361
http://www.surveymethods.org

Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies

Andrew Goodman-Bacon
Vanderbilt University, and
NBER

Jan Marcus
University of Hamburg, and
DIW Berlin

Policymakers have implemented a wide range of non-pharmaceutical interventions to fight the spread of COVID-19. Variation in policies across jurisdictions and over time strongly suggests a difference-in-differences (DD) research design to estimate causal effects of counter-COVID measures. We discuss threats to the validity of these DD designs and make recommendations about how researchers can avoid bias, interpret results accurately, and provide sound guidance to policymakers seeking to protect public health and facilitate an eventual economic recovery.

Keywords: COVID-19; difference-in-differences; non-pharmaceutical interventions; causal inference

1 Introduction

To mitigate the spread of the novel coronavirus SARS-CoV-2 and the disease it causes, COVID-19, policymakers around the world have placed severe restrictions on their citizens. Schools and businesses have been shuttered, gatherings banned, and more than half the world's population lived under a shelter-in-place order at some time (Sandford, 2020). Limiting interactions stems person-to-person contagion, but at a high cost. Business closures in Europe, for example, are estimated to reduce GDP by three percent per month (Thomsen, 2020).

Understanding which non-pharmaceutical interventions actually contain the pandemic is therefore crucial for balancing public health and economic and social costs. The fact that governments enact their own policies differently across place and time strongly suggests a difference-in-differences (DD) design for estimating causal effects in the COVID-19 context.[1] A DD design compares changes in COVID-related outcomes before and after a given policy takes effect in one area, to changes in the same outcomes in another area that did not introduce the policy. At least five recent papers use DD methods to show that non-pharmaceutical interventions reduce interactions, infections, or deaths (Dave, Friedson, Matsuzawa, & Sabia, 2020; Fang, Wang, & Yang, 2020; Friedson, McNichols, Sabia, & Dave, 2020; Gupta et al., 2020; Hsiang et al., 2020).

The validity of DD relies on assumptions about the comparability of treatment and control areas. The dynamics of COVID-19, the way people respond to it, and the flood of policy responses all make it difficult to develop credible DD research designs. Careful DD analyses, however, can be transparent, convincing, timely and policy-relevant. This article discusses challenges to using DD to evaluate counter-COVID measures as well as possible strategies to tackle these challenges.

2 Madrid and Liverpool: A Running Example

We center our discussion around a hypothetical analysis of the effect of lockdown policies on COVID-19 cases (y) in Madrid and Liverpool. Lockdown refers to a legally enforceable order that residents remain in their homes except for essential trips. Spain imposed a lockdown on March 14th and the UK did so on the 24th (Flaxman, Mishra, Gandy, et al., 2020).

While modern DD analyses apply new econometric techniques to large datasets with many groups and time periods, they all build on basic comparisons between changes in outcomes in a "treated" jurisdiction before and after it implemented a specific policy (first difference) with changes in outcomes over the same period in a "control" jurisdiction that did not implement the policy (second difference). To begin, consider a simple DD estimator that uses Madrid as the treatment group, Liverpool as the control group, March 1st–March 13th as the pre-treatment period (Pre), and March 14th–March 23rd as the post-treatment period (Post). (We discuss the role of the UK's lockdown below.) The DD estimate

Contact information: Jan Marcus, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany (Email: jan.marcus@uni-hamburg.de)

[1] While the term DD is widely used in economics, political science, and sociology, other disciplines sometimes refer to it as "controlled before-and-after study" or "(untreated) control group design with pretest and posttest". Angrist and Pischke (2009) and Lechner (2011) provide general discussions of DD.

equals:

$$\widehat{\beta}^{DD}_{\text{Madrid}} \equiv \overbrace{(\bar{y}^{\text{Post}}_{\text{Madrid}} - \bar{y}^{\text{Pre}}_{\text{Madrid}})}^{\Delta\bar{y}_{\text{Madrid}}} - \overbrace{(\bar{y}^{\text{Post}}_{\text{Liverpool}} - \bar{y}^{\text{Pre}}_{\text{Liverpool}})}^{\Delta\bar{y}_{\text{Liverpool}}} \quad (1)$$

Subscripts denote the city and superscripts denote the periods over which averages are taken. Differences over time *within* each city eliminate constant underlying differences in the level of y between Madrid and Liverpool that do not vary over time. Comparing these changes *across* cities eliminates factors that drive infection rates equally in both places. To interpret $\widehat{\beta}^{DD}_{\text{Madrid}}$ as the causal effect of the lockdown, we must assume that COVID-19 infections in Liverpool reflect how infections *would have changed* in Madrid had they not enacted a lockdown—the common trends assumption. If common trends fails, then Madrid's infections would have changed differently even without a lockdown and $\widehat{\beta}^{DD}_{\text{Madrid}}$ cannot be interpreted as a lockdown effect. If common trends holds, then $\widehat{\beta}^{DD}_{\text{Madrid}}$ gives the causal effect of treatment (lockdown) on the treated (Madrid in the third week of March).

3 DD Challenges

All DD analyses must carefully consider potential violations of the common trends assumption, many of which appear likely in the COVID-19 context. Moreover, the dynamics of COVID-19—lags between exposure and recorded infections, nonlinearities that arise form person-to-person transmission, and the likelihood that policies have different effects over time—complicate potential threats to the DD research design.

Packaged Policies

Governments typically implement several policies to reduce COVID-19 infections. For instance, Madrid implemented an early lockdown, but both Madrid and Liverpool took other steps to address COVID-19. Spanish officials encouraged social distancing on March 9th, closed schools on the 13th, and banned public events on the 14th, all of which could have reduced infections in Madrid ($\Delta\bar{y}_{\text{Madrid}}$). Disentangling the lockdown effect from the collective effect of Madrid's policies requires a control group that reflects the effects of the other policies that Madrid enacted. While social distancing was also encouraged in Liverpool, Liverpool may have issued different advice than Madrid, they did so later in the sample (March 16th), and Liverpool did not close schools. Liverpool is probably not a good control for Madrid.

Reverse Causality

Governments have enacted restrictions because of worsening outbreaks, so variation in policies to fight COVID-19 may be functions of past changes in COVID-19 itself (see Gupta et al., 2020). Small differences at the infection's outset (pre-period) that trigger local interventions can imply large differences in the infection's subsequent development (post-period). In early March, Madrid had much higher growth in per-capita cases than Liverpool and a higher but earlier peak in infections. Falling infection rates in the post-lockdown period may therefore reflect the natural dynamics of the outbreak, not the effect of the lockdown. Areas with high enough infection rates to trigger strong policy responses probably just have worse outbreaks than areas that did not impose restrictions. The COVID-19 outbreaks in Madrid and Liverpool may simply never have been comparable to each other.

Voluntary Precautions

As outbreaks grow and are publicized, evidence suggests that people take precautions before any official restrictions go into place. Data from Google users, for example, show that public transport use in both Liverpool and Madrid started to fall days before their lockdowns began (Google, 2020). DD estimates will be biased toward finding a spurious negative effect of lockdowns on infection rates because people and policymakers respond to the same information. Moreover, if worse infections trigger earlier and more restrictive policies, they may also generate stronger precautionary behavior and larger bias.

Anticipation

When governments announce policy ahead of time, however, behavior may change in response to information about the policy itself. For example, just before lockdowns took effect, people in Madrid and Liverpool made more trips to grocery markets and pharmacies, stocking up on toilet paper and pasta (Google, 2020; Oakley, 2020).[2] If resulting infections were reported before the 14th, cases may fall in Madrid after lockdown ($\Delta\bar{y}_{\text{Madrid}} < 0$) but only because anticipation raised the pre-lockdown infection rate.[3]

Spillovers

Infectious diseases do not stop at regional borders, so the timing and effectiveness of Spain's lockdown may affect COVID-19 cases in Liverpool. For example, 3,000 football fans traveled from Spain to Liverpool on March 11th for a Champions League match between Atletico Madrid and Liverpool FC (ESPN, 2020), but a March 17th match between Real Madrid and Manchester City FC was cancelled. Infection trends in Liverpool cannot represent what would have

[2] Fang et al. (2020) estimate large anticipation effects on travel out of Wuhan in a matter of hours.

[3] A pre-lockdown spike in outcomes makes the most sense for outcome variables that can respond to information and policies immediately, such as population movements.

happened in Madrid absent the lockdown if they are a function of Spain's lockdown.[4] In the context of infectious diseases, spillovers from travel will typically bias DD estimates towards zero. Lockdowns that help the treatment group also help the control group.

Variation in Policy Timing

If the observation period is extended beyond March 23rd, an analysis of the two cities would have to incorporate Liverpool's lockdown. The most common way to do this is to estimate a regression with unit and time fixed effects, and a dummy variable that equals one when unit i has the policy in time t. The DD estimate then equals an average of $\widehat{\beta}^{DD}_{\text{Madrid}}$ and a similar DD term for Liverpool's lockdown effect that uses Liverpool as the treatment group, Madrid as the control group (after its lockdown is already in effect), after March 24th as the post-period, and March 14th through March 23rd as the pre-period (Goodman-Bacon, 2019).

Unfortunately, when treatment is staggered and treatment effects vary over time—both true in the COVID-19 context—two-way fixed effects estimates are typically biased away from the sign of the true treatment effect (for example, see Goodman-Bacon, 2019). The bias comes from the fact that the DD term for Liverpool uses Madrid after March 14th as a control group. If Madrid's lockdown effect were constant and immediate, it would difference out when calculating outcome trends in Madrid after the 14th. If the lockdown effect grows over time, however, then the trend in Madrid's infections after the 14th *includes* an evolving treatment effect that Liverpool would never have experienced. Madrid does not reflect what would have happened Liverpool without lockdown, and the regression DD estimate is positively biased.

Measurement and Scaling of the Dependent Variable

COVID-19 outcomes are inherently hard to measure, which compounds the previous challenges. The virus' incubation period, for example, means that reported infections lag true infections by several days.[5] Therefore, policies that limit exposure (as well as confounding factors like voluntary precautions or anticipation) will not affect recorded infection rates immediately. For instance, the DD estimate in equation (1) has a post-lockdown period of 9 days. Madrid's lockdown is unlikely to have had any effect on reported cases on many of those days. Finally, authorities have legitimate trouble counting the number of COVID-19 infections and deaths. DD estimates will be biased if testing improves differently in the treatment and control groups.

Further, decisions about how to measure outcomes (counts, rates, or logs) can strongly affect analyses of COVID-19 policies because COVID-19 outbreaks vary widely across areas, and their trajectories vary widely over time. The common trend assumption, for example, depends on scaling. If it holds in outcome levels, it cannot generally hold for the log of the outcome, so only one transformation can yield valid DD estimates.[6]

4 Recommendations for COVID-19 DD Designs

While we use a specific case to make our points, they apply to policies other than lockdowns, outcomes other than infections, and datasets with many more than two jurisdictions. Here we recommend steps to avoid these biases and accurately interpret DD estimates.[7]

Recommendation 1: Estimate Dynamics

Researchers should present "event-study" estimates that trace a policy's effect on individual days before and after it takes effect.[8] This generalization of the canonical DD model in equation (1) can reveal many of the biases outlined above. Reverse causality, for example, implies that COVID-19 outcomes should be getting worse in the days leading to the policy, which will show up as increasing pre-treatment event-study estimates. Voluntary precautions, on the other hand, would improve outcomes in treated areas prior to the policy. The virus' incubation period also means that the shape of post-policy event-study estimates can act as a check on the research design. Changes in reported COVID-19 cases and fatalities immediately after treatment, for example, are unlikely to be caused by the treatment.

Recommendation 2: Choose the Control Group Wisely

Good control groups will have to match treatment groups on many dimensions. Smart research designs will try to focus on situations where treatment and control groups differ only by the introduction of a single COVID-19 policy (or, at least, only few policies). DD analyses at the sub-national level, where all areas share national counter-COVID measures can

[4] This is a violation of the stable unit treatment value assumption (SUTVA), which states that each unit's outcomes are independent of other units' treatment status.

[5] Lauer et al. (2020) report a median incubation period of 5.1 days, with 97.5 percent of cases developing symptoms within 11.5 days.

[6] Measurement error interacts with outcome scaling. For example, if areas with different infection rates only record 80 percent of cases, their infection rates (in levels) will diverge over time but log infection rates will not.

[7] Data quality and availability underlie all of our recommendations. Governments, firms, and researchers collect different information on outcomes like mobility, symptoms, vital records, employment, and behaviors. This information should continue to be publicly available. High-quality data make credible comparisons possible and without them, even the most sophisticated techniques will fail to uncover causal effects.

[8] For examples see Dave et al. (2020), Fang et al. (2020), Friedson et al. (2020), Gupta et al. (2020).

help (although they may have worse spillovers from travel). Researchers should focus especially on techniques, such as propensity score reweighting or synthetic control (Abadie, Diamond, & Hainmueller, 2010), that impose balance in pre-policy infection levels and trends, characteristics that affect transmission (ie. population density), virulence (ie. comorbidities or age structure), or the ability to comply with movement restrictions (ie. occupation and income).[9] Note that controlling directly for confounders in a regression may not be adequate, because it "removes" an average partial correlation (while the true bias likely varies strongly across units and over time), and can severely restrict the independent variation in the policy of interest (Goodman-Bacon, 2019).

Recommendation 3: Be Careful of Regression DD

Even when common trends holds, regression DD can be biased when treatment effects vary over time.[10] Fortunately, a range of alternative estimators have been developed that avoid these biases. Callaway and Sant'Anna (2018), for example, propose averaging together a series of (propensity score reweighted) DD models like (1) that use later treated units as a control group for earlier treated ones. COVID-19 policies are well-suited to these alternative approaches and applying them may be crucial to avoiding biases inherent in regression approaches.

Recommendation 4: Sign the Bias

Sometimes it will not be possible to eliminate all sources of bias, but it may be possible to report the direction of the bias.[11] For instance, the UK's social distancing guidelines presumably reduce only $\bar{y}^{\text{Post}}_{\text{Liverpool}}$ and bias the estimated effect of Madrid's lockdown toward zero. One may still have confidence, then, in a DD result showing that Madrid's lockdown *did* reduce infections. Alternatively, Madrid's school closure likely reduced infections at almost the same time as its lockdown, making it less credible to attribute the DD estimate in (1) entirely to the lockdown. Researchers should clearly state potential sources of bias, their sign, and their likely magnitudes when interpreting DD estimates.

Recommendation 5: Be Clear about What Is Knowable

COVID-19 policies may have very different effects in the areas that actually implement them than in areas that chose not to do so, but DD identifies effects in treated areas. DD models based on policy timing necessarily focus on small windows of time after one area has a policy and before another area does. As we discuss, policies almost certainly do not have immediate effects on reported outcomes, though. Null short-run effects do not mean a policy has no effect. Researchers should be clear about these interpretation issues, while also trying to understand the sources of treatment effect heterogeneity across space and time. In fact, many policymakers may be most interested in heterogeneity in the effect of public health measures.

5 Conclusion

Because non-pharmaceutical COVID-19 interventions have not been randomized, researchers have to rely on quasi-experimental strategies to identify causal effects. By including control groups, DD provides important advantages over methods like before-and-after comparisons and interrupted time-series designs (e.g., Tobías, 2020). Moreover, the graphical and parametric tools developed for DD in recent years make it possible to assess the plausibility of the common trends assumption.

Causal estimates are important inputs into COVID-19 policymaking, and bias can have serious consequences. If policymakers mistakenly believe that restrictions have little effect, they may relax policy too soon and infections will spike again. If, on the other hand, they mistakenly believe that policies have large effects they may maintain restrictions for too long, hampering economic and social recovery. Unbiased and transparent DD evidence, however, can provide timely and accurate guidance required to navigate the COVID-19 pandemic.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. doi:10.1198/jasa.2009.ap08746.http

Angrist, J. D., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Callaway, B., & Sant'Anna, P. H. C. (2018). Difference-in-differences with multiple time periods and an application on the minimum wage and employment. eprint: 1803.09015. Retrieved from http://arxiv.org/abs/1803.09015

Dave, D., Friedson, A. I., Matsuzawa, K., & Sabia, J. J. (2020). When do shelter-in-place orders fight COVID-19 best? Policy heterogeneity across states and adoption time. *IZA Discussion Paper*, *13190*.

[9] See Friedson et al. (2020) for a synthetic control analysis of counter-COVID policies.

[10] Using a flexible event-study specification does not automatically solve these problems (Sun & Abraham, 2018).

[11] Rambachan and Roth (2019) develop methods to bound DD estimates when common trends may not hold.

ESPN. (2020). Coronavirus: Liverpool vs. Atletico Madrid to be investigated by council. Retrieved from https://www.espn.com/soccer/uefa-champions-league/story/4088954/coronavirus-liverpool-vs-atletico-madrid-to-be-investigated-by-council

Fang, H., Wang, L., & Yang, Y. (2020). Human mobility restrictions and the spread of the novel Coronavirus (2019-nCoV) in China. *NBER Working Paper*, *26906*.

Flaxman, S., Mishra, S., Gandy, A., et al. (2020). Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. *Imperial College London*, (March), 1–35. doi:10.25561/77731

Friedson, A. I., McNichols, D., Sabia, J. J., & Dave, D. (2020). Did California's shelter-in-place order work? Early Coronavirus-related public health effects. *NBER Working Paper*, *26992*.

Goodman-Bacon, A. (2019). Difference-in-differences with variation in treatment timing. Retrieved from https://cdn.vanderbilt.edu/vu-my/wp-content/uploads/sites/2318/2019/07/29170757/ddtiming_7_29_2019.pdf

Google. (2020). Covid-19 community mobility reports. Retrieved from https://www.google.com/covid19/mobility/

Gupta, S., Nguyen, T. D., Rojas, F. L., Raman, S., Lee, B., Bento, A., … Wing, C. (2020). Tracking public and private response to the COVID-19 epidemic: Evidence from state and local government actions. *NBER Working Paper*, *27027*.

Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., … Krasovich, E., et al. (2020). The effect of large-scale anti-contagion policies on the Coronavirus (COVID-19) pandemic. Cold Spring Harbor Laboratory Press. doi:10.1101/2020.03.22.20040642

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., … Lessler, J. (2020). The incubation period of Coronavirus Disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, *172*(9), 577–582. doi:10.7326/M20-0504

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*(3), 165–224. doi:10.1561/0800000014

Oakley, J. (2020). Coronavirus: Unravelling the mystery of the vanishing toilet rolls in Spain's andalucia. Retrieved from https://www.theolivepress.es/spain-news/2020/03/12/coronavirus-unravelling-the-mystery-of-the-vanishing-toilet-rolls-in-spains-andalucia/

Rambachan, A., & Roth, J. (2019). An honest approach to parallel trends. Retrieved from https://scholar.harvard.edu/jroth/publications/Roth_JMP_Honest_Parallel_Trends

Sandford, A. (2020). Coronavirus: Half of humanity now on lockdown as 90 countries call for confinement. Retrieved from https://www.euronews.com/2020/04/02/coronavirus-in-europe-spain-s-death-toll-hits-10-000-after-record-950-new-deaths-in-24-hou

Sun, L., & Abraham, S. (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. eprint: 1804.05785. Retrieved from http://arxiv.org/abs/1804.05785

Thomsen, P. (2020). Transcript of April 2020 European Department Press Briefing. Retrieved from https://www.imf.org/en/News/Articles/2020/04/15/tr041520-transcript-of-april-2020-european-department-press-briefing

Tobías, A. (2020). Evaluation of the lockdowns for the SARS-CoV-2 epidemic in Italy and Spain after one month follow up. *Science of the Total Environment*, *725*, 138539. doi:10.1016/j.scitotenv.2020.138539

Commentary

The authors provide an interesting paper that thoughtfully lays out some possible issues with using the Difference-in-Differences (DD) analysis method and walks through the best ways to tackle these to avoid pitfalls. The authors do a very nice job explaining the DD analysis method, as well as the potential challenges when using this method. The writing is clear and provides food for thought, but a few modifications could provide further context and insight.

In the introduction, the authors set the stage for using DD in the context of COVID-19 analyses and cite several papers that have used this method to show that non-pharmaceutical interventions reduce some outcomes. We would have expected to see commentary in this introduction or throughout the paper detailing these examples and whether they properly used DD in their analyses. This might have provided more context for the challenges of this method in COVID-19 analyses.

The paper explains the DD method; however, since many readers may not be familiar with DD, perhaps a more in-depth explanation of the method and how it works would be warranted. In addition, Equation 1 provides a helpful visual of DD, but, perhaps in addition, a figure would be useful in further explaining the method.

The authors highlight seven potential challenges with using the DD analysis method. These involve concepts of human behavior adoption, government policy implementation, and measurement and reporting of infection rates. While all provoked thought, raised good points, and are important to the paper, the most interesting were the arguments about the human behavior adoption, including the adoption of not-yet-mandatory behaviors ahead of government mandates, discussed as Voluntary Precautions and Anticipation, as well as the policy implications that affect both the control and the

comparison areas—discussed as Spillovers.

When examining the effect of policies, it is important to understand how those policies affect behavior. As the authors point out in the discussions of Voluntary Precautions and Anticipation, the humans involved in these analyses do not behave in exact accordance with the policies implemented by the governments or on the exact days. The behaviors associated with Voluntary Precautions are those that people start implementing before the policy is enacted. The authors use the example of decreased use of public transportation in the days leading up to lockdowns. One might also assume people were already limiting other behaviors such as non-essential medical appointments and social gatherings as well in anticipation of the lockdown. The authors make a valid argument that these behaviors may bias the analysis by skewing the data. Might using auxiliary data such as the Google data cited allow analysts to alter their analyses either by altering the timeframes they use, anticipating adoption of voluntary precautions, or adjusting for such behaviors in other ways?

Spillovers are presented as affects felt in one of the areas examined in the analysis due to changes in the other area. This is a valid point that disease does not know borders and application of certain policies in one area can affect the analysis by biasing the estimate downward in another area—resulting in a lower perceived effect of the lockdown policies.

In regards to the discussion of policy timing, it might be useful to further explore here by examining different pre-post periods. For a disease outbreak that can exist without being detected, because of the incubation time period and the high prevalence of asymptomatic cases coupled with the very high infection/transmission rate, applying a DD analysis to a policy is problematic because the comparison of the outcome is dependent on an unknown factor from the start. It would be useful to see a bit more of a justification for this type of analysis with a disease outbreak, as well as to discuss that within Recommendation 5 (it is touched on somewhat in Recommendation 1). In addition, a figure in Section F would be useful to further explain the challenges highlighted.

For Section G, the discussion of differential measurement error at the top of Page 4 is very highly relevant and warrants its own section. The authors supply five strong recommendations for avoiding bias in DD analyses. These recommendations are well thought out and supply sound reasoning. Overall, the authors point to DD as a valid method to compare causal effects of the policies put into place, as long as the analyses are handled carefully and use an appropriate control group. The importance of the interpretation of these analyses is reflected in future policy choices by lawmakers and public health officials. Supplying accurate analyses of data for these decision makers is crucial to effectively mitigating disease outbreaks.

Alicia Frasier, Stephanie Eckman, and Heidi Guyer

RTI International