The Econometrics Journal

Econometrics Journal (2020), volume **00**, pp. 1–22.
doi: 10.1093/ectj/utaa025

Quantifying the impact of nonpharmaceutical interventions during the COVID-19 outbreak: The case of Sweden

SANG-WOOK (STANLEY) CHO

School of Economics, University of New South Wales, Sydney, 2052, Australia.
Email: s.cho@unsw.edu.au

First version received: 9 July 2020; final version accepted: 11 August 2020.

Summary: This paper estimates the effect of nonpharmaceutical intervention policies on public health during the COVID-19 outbreak by considering a counterfactual case for Sweden. Using a synthetic control approach, I find that strict initial lockdown measures play an important role in limiting the spread of the COVID-19 infection, as the infection cases in Sweden would have been reduced by almost 75 percent had its policymakers followed stricter containment policies. As people dynamically adjust their behaviour in response to information and policies, the impact of nonpharmaceutical interventions becomes visible, with a time lag of around 5 weeks. Supplementary robustness checks and an alternative difference-in-differences framework analysis do not fundamentally alter the main conclusions. Finally, extending the analysis to excess mortality, I find that the lockdown measures would have been associated with a lower excess mortality rate in Sweden by 25 percentage points, with a steep age gradient of 29 percentage points for the most vulnerable elderly cohort. The outcome of this study can assist policymakers in laying out future guidelines to further protect public health, as well as facilitate plans for economic recovery.

Keywords: *COVID-19*, *causal impact*, *synthetic control method*, *public health policies*, *Sweden*.

JEL Codes: *C23*, *I18*, *O57*.

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a viral respiratory illness caused by a new coronavirus, first reported in Wuhan, Hubei Province, China, in November 2019. Over the following few months, the illness rapidly spread to almost every country. In response, the World Health Organization (WHO) declared COVID-19 a pandemic on March 11, 2020. As vaccines or medicines for COVID-19 have yet to be available, most countries around the world have resorted to nonpharmaceutical interventions (NPIs), or community mitigation strategies, to help slow the spread of the illness. Some of the NPIs involve government measures to close schools and workplaces, cancel and restrict public events and gatherings, shut down public transport, and implement stay-at-home requirements, as well as restrictions on domestic and international travel, in addition to general public information campaigns. By late March, nearly all countries in Europe had implemented these policies by placing themselves into a nationwide lockdown. These government policies remained in place until May, with a gradual easing of some of the harshest measures. One country, however, stood out for its decision to remain open: Sweden. In fact, Swedish officials

chose not to implement a nationwide lockdown, trusting that people would voluntarily do their part to stay safe. For example, although high schools and universities have switched to distance learning, elementary and preschools have remained open. Besides, although the government recommended that people stay at home, many nonessential businesses, such as restaurants, gyms, and bars, remained open, and gatherings of up to 50 people were allowed. Given this divergence in the policy measures between Sweden and the rest of Europe, I study the public health impact of NPIs by comparing how the trajectories of the COVID-19 infection and mortality rates would have evolved had Sweden opted for more stringent lockdown measures.

To study this counterfactual scenario, I use the synthetic control (SC) method pioneered by Abadie and Gardeazabal (2003) and analyse how a parallel version of Sweden (or a "synthetic" Sweden) would have evolved had it enforced a mandatory lockdown policy. This parallel version of Sweden is first constructed through a data-driven process with weights assigned to all possible donor countries that would best approximate the pre-lockdown characteristics of Sweden (our "treatment" unit). Once the policy intervention takes place, I can trace its effect with the evolution of the untreated SC unit to assess the counterfactual situation corresponding to a regime where strict lockdown measures were in place. The causal effect of the lockdown is then measured by the post-intervention difference in infection rates between Sweden and its synthetic counterpart. It has been shown that the SC method offers several advantages over traditional difference-in-differences (DD) or fixed-effect models, as not only is the procedure a transparent data-driven one, but it also allows the effect of unobservable country heterogeneity to vary over time, as discussed by Abadie, Diamond, and Hainmueller (2010) and Imbens and Wooldridge (2009). I further quantify the causal effect of counter-COVID measures by using a DD research design that controls for additional variables regarding people's behaviour. This helps us to assess how much of the observed infection rate dynamics is attributed to the NPIs by themselves relative to voluntary changes in people's behaviour for fear of infection.

The key findings from the analysis are as follows. One hundred days after the infection case per one million population exceeds one, I find that the infection case in synthetic Sweden is around 75 percent lower than that of actual Sweden, which implies that stricter containment measures are associated with limiting the spread of the COVID-19 infection. I also find that as people dynamically adjust their behaviour in response to information and policies, the impact of NPIs does not manifest immediately but only with a time lag of approximately 5 weeks. Supplementing the main findings with several robustness checks and a complementary analysis using a DD approach does not fundamentally alter the main findings. Furthermore, profiling excess mortality trends between Sweden and its synthetic counterpart, I find that the excess mortality rate under a counterfactual lockdown would have been reduced by 25 percentage points. The effectiveness in death prevention follows a disproportionately steep age gradient that ranges from a 14–percentage point reduction in the excess mortality rate for the working-age cohort to a 29–percentage point reduction for the elderly cohort 85 years of age and above.

The present paper contributes to the ongoing discussion on the effectiveness of NPI policy response to the COVID-19 shock; see Chernozhukov, Kasahara, and Schrimpf (2020); Chen and Qiu (2020); Gonzalez-Eiras and Niepelt (2020); Ullah and Ajala (2020); Goodman-Bacon and Marcus (2020); Gupta et al. (2020); and the contributions in the volume by Baldwin and di Mauro (2020). Empirically, the present paper extends cross-country experiences in policy effectiveness. Castex, Dechter, and Lorca (2020) showed that the effectiveness of lockdown policies is declining with GDP per capita and population density but increasing with health expenditure and proportion of physicians in the population. Focusing on the policy choices in Sweden, Conyon, He, and Thomsen (2020) compared the COVID-19 deaths, while Andersen

et al. (2020) examined aggregate spending patterns in comparison to Sweden's neighbours. In terms of scope and methodology, the present paper is closest in spirit to Born, Dietrich, and Müller (2020) (hereinafter BDM), who conducted a similar counterfactual lockdown scenario for Sweden using the SC method. Documenting the infection dynamics of one month after lockdown, they found that counterfactual Sweden did not differ from the actual infection dynamics observed in Sweden. In their discussion, they attributed this outcome to the voluntary precautions taken by the general public that essentially had the same impact as a mandatory lockdown.

The present paper extends BDM in the following aspects. First, I consider the post-lockdown period as extending for around 75 days, which mostly covers the time horizon during which the initial lockdown measures were fully in place outside Sweden. Consistent with BDM, I also find that during the first 5 weeks, the infection rate in synthetic Sweden was not significantly lower than that in actual Sweden. However, over time, synthetic Sweden shows a significant slowdown in the infection rate, which demonstrates that the lockdown measures would eventually have had a containment effect in the longer horizon. Second, using a DD approach and Google Mobility Tracker, I formally control for the behavioural changes and show that the mandatory lockdown measures would have significantly reduced the infection rate in comparison with a voluntary social distancing scenario.

The rest of the paper is organised as follows. Section 2 describes the methodology and data for the SC approach. Section 3 presents the main estimation results and robustness checks. Section 4 extends the analysis to mortality rate and discusses the role of voluntary social distancing, followed by a DD estimation in Section 5. Finally, the conclusion provided in Section 6 discusses some limitations and caveats.

2. METHODOLOGY AND DATA

The main methodology of choice is the SC method proposed by Abadie and Gardeazabal (2003) and later developed in Abadie, Diamond, and Hainmueller (2010) and Abadie, Diamond, and Hainmueller (2015). The SC method has recently become a popular approach for comparative case studies and has also been used to quantify the economic effects of shocks or policy interventions.

Under a standard SC approach, one would compare a lockdown country to a synthetic unit composed of countries where no lockdown was imposed. In the context of this analysis, I swap the setting by assigning treatment to Sweden, where no lockdown was in place, and compare it with a counterfactual designed to capture how the infection rates would have evolved in Sweden had it followed a policy approach (or a mandatory lockdown) similar to that taken by other European countries. This counterfactual (or SC) unit tracks the actual path of infection rates in Sweden as closely as possible before the policy intervention but, at the same time, shares other similar characteristics with Sweden. After the policy intervention, the control unit follows a path of mandatory lockdown measures, whereas Sweden does not. As such, the notion of policy intervention in our setting refers to the *absence* of mandatory lockdown measures or no changes in government policy. Because of difficulties in picking individual countries that satisfy these criteria, I resort to a weighted average of potentially comparable countries that best resemble the characteristics of Sweden before the policy intervention. Any discrepancy in the infection dynamics between the two units during the post-intervention period can be interpreted as an outcome of the policy or the treatment effect. This is also the setup used in BDM and labelled as an "upside-down" approach.

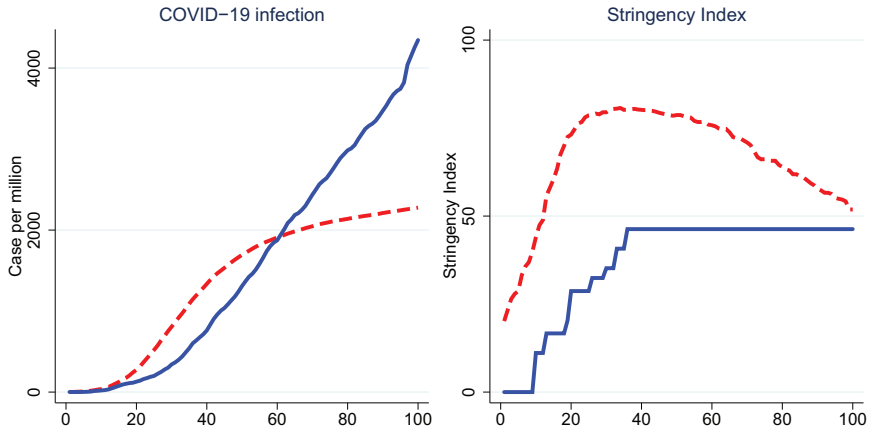

Figure 1. Infection profile and the Stringency Index.
Notes: Horizontal axis measures days since the cases per million exceed one. The profile for Sweden is shown in the solid blue line, whereas the average of all donors is in the red dashed line.

Because the SC method exploits the pre-intervention data to form better counterfactual values, it is often preferred over other program evaluation methods, such as DD, in comparative case studies. See Abadie (2020) for a broader overview of the methodology and its applicability.

2.1. Data

The outcome variable of interest is daily infection dynamics in Sweden, as measured by cumulative confirmed cases per million population. As for potential donors, I select 30 countries consisting of the other European Union members (excluding Malta because of lack of data), as well as Iceland, Israel, Norway, Switzerland, and the United Kingdom.[1] Because the infection dynamics vary across countries, I normalise the time unit such that "Day 1" refers to the day on which the infection cases per million exceed one. Next, the policy intervention term "lockdown" is a policy package consisting of various socioeconomic measures, including school and workplace closings, cancellation of public events, restrictions on gatherings, closing of public transport, stay-at-home requirements, and restrictions on domestic/international travel, as well as public information campaigns. Because a different sequence of measures took place over time with varying magnitudes, I resort to an all-inclusive index measure. The Oxford COVID-19 Government Response Tracker (OxCGRT) collects information on several different common government responses and provides a daily Government Response Stringency Index (Stringency Index; SI), which ranges from 0 to 100, with each additional policy response leading to a higher index value.[2] In Figure 1, I plot the series of infection cases per million as well as the SI for

[1] Full list of countries are as follows: Austria (AUT), Belgium (BEL), Bulgaria (BGR), Croatia (HRV), Cyprus (CYP), Czech Republic (CZE), Denmark (DNK), Estonia (EST), Finland (FIN), France (FRA), Germany (DEU), Greece (GRC), Hungary (HUN), Iceland (ISL), Ireland (IRL), Israel (ISR), Italy (ITA), Latvia (LVA), Lithuania (LTU), Luxembourg (LUX), Netherlands (NLD), Norway (NOR), Poland (POL), Portugal (PRT), Romania (ROU), Spain (ESP), Switzerland (CHE), Slovakia (SVK), Slovenia (SVN), and the UK (GBR).

[2] Hale et al. (2020) provide detailed information on the construction of the index.

Table 1. COVID-19 and demographic characteristics.

Variables	Sweden	All donors (N=30)
COVID-19 dynamics		
Day 1	February 29	March 4
Case per million on Day 1	1.188	1.445
Lockdown day		March 28 (Day 25)
Case per million on Day 25	199.618	519.577
SI on Day 25	28.7	78.6
Demographics		
Population density	24.981	142.974
Urban population fraction (%)	87.431	74.152
Average household size	2.2	2.453

Notes: Day 1 refers to the date on which the infection per million exceeds one. Lockdown day refers to the date on which the SI peaked in the donor countries.

Sweden and the average of all donors. The data are taken from the Coronavirus Pandemic section in Our World in Data (Roser et al., 2020).

It is worth noting that the average profile of the SI for the donors shows a rapid increase that peaks at around Day 25. Despite variations across countries, no country peaked below the index value of 50. The average index gradually decreases as many countries later eased some of the lockdown measures. On the other hand, for Sweden, the index rises slowly and always stays below the average profile of the donors, and at the same time, never exceeds 50 over the whole period of observation.[3] Because the SC approach requires a dichotomisation of policy intervention, I pick the average day on which the SI peaked in each donor country to pinpoint the timing of our lockdown treatment. This occurs on Day 25.

In Table 1, I summarise the COVID-19 dynamics for Sweden and the average of all donors, as well as some relevant country-specific demographic covariates. Rather than the overall population size, it is the immediate access to a population susceptible to infection that matters more in the transmission dynamics. As such, I compare population density, because early epidemiological studies such as Rocklöv and Sjödin (2020) found that high population density can catalyse the spread of COVID-19. In a similar manner, I also include the fraction of population living in urban areas. Finally, at a more disaggregate level, Sá (2020) found that regions with larger average household size experienced higher infection rates. The latest available figures for the country-specific demographic covariates were taken from the World Development Indicators and the United Nations report (United Nations, 2019).

Descriptive statistics in Table 1 indicate that the COVID-19 infection started a few days earlier in Sweden than in the average of donor countries, but the latter group experienced a faster rise in the infection cases, prompting a swift government response. On Day 25, the average SI of all donors reached its peak at 78.6, whereas in Sweden the SI remained much lower at 28.7. As for the demographic covariates, Sweden is characterised by a smaller population density and household size than the pool of donors but has one of the highest urbanisation rates in Europe.

Because of disparities in demographic characteristics, as well as pre-lockdown infection patterns in the two groups, I proceed with the SC approach and find a weighted average of the

[3] For Sweden, this period of 100 days ranges from February 29 to June 7.

Table 2. Profile of synthetic Sweden.

	Sweden	Synthetic Sweden
Chosen donors and weights		FIN (0.489), GRC (0.242), NOR (0.216), DNK (0.033), EST (0.020)
Demographic covariates		
Population density	24.981	37.327
Urbanisation (%)	87.431	82.927
Average household size	2.2	2.198
Pre-treatment outcomes		
Average death cases (Day 1–20)	0.158	0.159
Infection cases (Day 7)	6.040	8.196
Infection cases (Day 14)	61.391	61.382
Infection cases (Day 23)	172.884	176.858
RMSPE		4.585

Notes: Weights for chosen donors are shown in brackets. All others in the donor pool receive zero weight.

countries in the donor pool that generates the SC unit for Sweden. The weights are assigned by minimising the distance between Sweden and the SC unit along a set of predictors that balances aforementioned demographic covariates and pre-lockdown COVID-19 transmission profiles. For the latter, I choose the average deaths per million reported in the first 20 days, as well as three lagged values of infection cases per million on Days 7, 14, and 23. Including lagged terms of the outcome variable often helps mitigate the problem of omitting important predictor effects, as suggested by Athey and Imbens (2006). The benchmark set of predictors is subject to various robustness checks later in Section 3.3.

3. RESULTS

In this section, I present the results from the SC analysis by first showing the benchmark profile of synthetic Sweden, followed by an inference test and various robustness checks.

3.1. Synthetic Sweden

Table 2 summarises the seven predictors for Sweden and synthetic Sweden, where the latter is constructed with positive weights assigned to Finland, Greece, Norway, Denmark, and Estonia in descending order.[4] Compared with the simple average of all countries in the donor pool (as shown in Table 1), the SC unit provides a much better-matched profile of Sweden along the predictors. In other words, the weighted selection of countries is more appropriate as a control unit than taking a simple average of all countries in the donor pool. In the SC approach, the root mean square prediction error (RMSPE) measures the gap between the variable of interest for the treated unit and its synthetic counterpart for all pre-treatment periods, which is reported in the last row of Table 2.

[4] In the Appendix, Table A1 breaks down the predictors for each of the chosen donors, as well as other descriptive statistics. Simple and weighted averages of lockdown days were Day 23 and Day 24, respectively, which are close to the normalised lockdown treatment on Day 25.

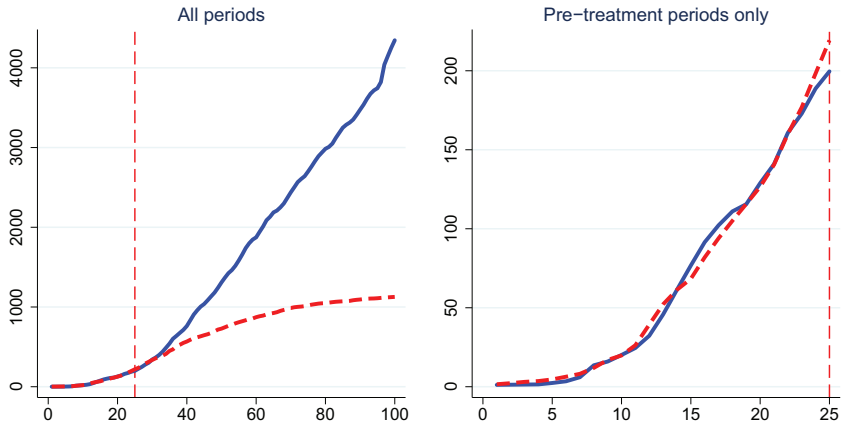

Figure 2. Infection profile of Sweden versus synthetic Sweden.
Notes: Horizontal axis measures days since the cases per million exceed one. The profile for Sweden is shown in the solid blue line, whereas that of synthetic Sweden is in the red dashed line. Vertical dashed line represents the lockdown date.

Next, Figure 2 compares the profile of infection dynamics for synthetic Sweden and actual Sweden. The left panel covers the entire 100 days since the cases per million exceed 1, which roughly includes the entire months of March to May and early June. The right panel zooms into the pre-treatment period to visually assess the quality of fit in the first 25 days. The policy intervention takes place on Day 25—as indicated by the dashed vertical line—which falls on the midpoint of actual lockdown dates of the five countries comprising the SC unit.

Although the cumulative infection cases in synthetic Sweden follow those of actual Sweden quite closely before the policy intervention, there is a visible divergence shortly after the intervention, from which Sweden follows a much steeper path than its synthetic counterpart.[5] By Day 100, or roughly 11 weeks after the lockdown intervention in synthetic Sweden, the infection cases in Sweden exceed 4,300 per million population. On the other hand, the corresponding figure for synthetic Sweden is around 1,100. In other words, with a gap of around 3,200 cases per million, the infection case in Sweden would have been lower by almost 75 percent had there been a similar policy implemented in the most comparable version of itself.

3.2. Inference

To evaluate the significance of the benchmark results, I conduct a test of inference for the SC framework based on permutation techniques, as suggested in Abadie, Diamond, and Hainmueller (2010). First, I run a cross-sectional placebo test (or "placebo in-space") by sequentially applying the SC algorithm to each country in the pool of all donors, which generates a distribution of placebo estimates across 30 countries.[6] We can then compare the benchmark estimates of

[5] The individual profiles for donors with nonzero weights are shown in Figure A1 in the Appendix.

[6] Because the permutation inference procedure relies on very strong random assignment assumptions, it will not likely hold in the current context in aggregate units. Although it is difficult to articulate the nature of a placebo intervention,

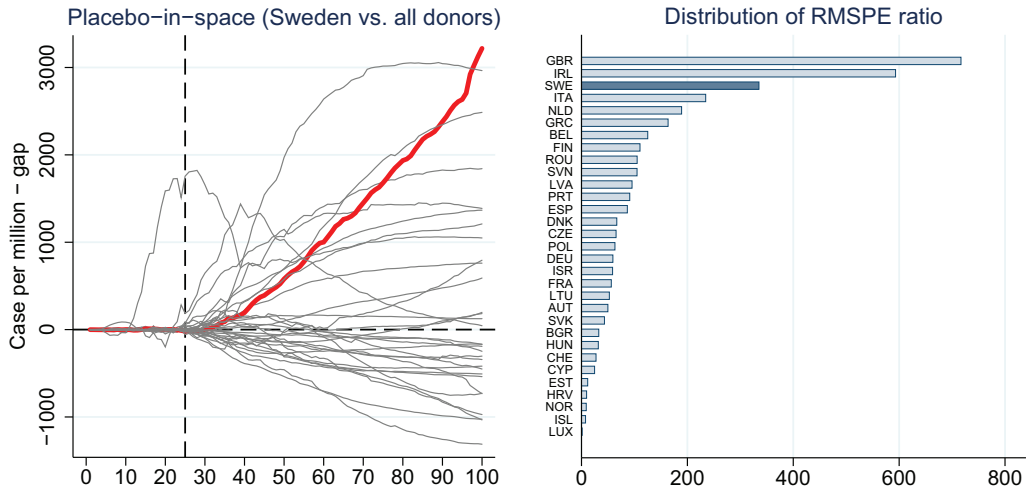

Figure 3. Inference test.

Notes: In the left panel, the gap between the treated unit and SC unit is plotted for Sweden (in red) and each of the donors (in grey). Vertical dashed line indicates the date of policy intervention. In the right panel, the RMSPE ratios are listed in descending order. The ratio for Sweden is shown in the darker shade.

the truly treated economy with this distribution to assess whether Sweden's treatment effect is extreme relative to the donor pool. This cross-sectional distribution of placebo tests is shown on the left side panel of Figure 3. The grey lines show the gap in the infection rates between each country in the donor pool and its respective synthetic version. The thick red line depicts the baseline results obtained for Sweden. A visual inspection suggests that Sweden joins in the list of countries with positive treatment effects, but not necessarily at the right tail of the distribution of treatment effects throughout the whole post-treatment period. However, toward the end of the sample period, the treatment effect for Sweden is distinctly higher than in most other countries.

Although the cross-sectional distribution of placebo tests offers visual evidence of the treatment effects over time, it does not provide a measurement that quantifies the overall significance of the results. To tackle this issue, I follow Abadie, Diamond, and Hainmueller (2010), who offered an alternative test statistic by constructing exact p-values based on Fisher (1935). Because the RMSPE measures the gap between the variable of interest for the treated country and its synthetic counterpart, we can calculate a set of RMSPE values for the pre- and post-treatment period for all units in the cross-sectional placebo test. I then compute the country-specific ratio of the post- to pre-treatment RMSPE to quantify the post-treatment divergence in the infection rate relative to the estimated pre-treatment gap. The distribution of the RMSPE ratios from highest to lowest is shown in the right-hand panel of Figure 3. For Sweden, the RMSPE ratio of around 335 is far higher than those obtained for most other countries in the control group, only following the United Kingdom and Ireland. The ranking, converted into fractions, provides the basis for a p-value for Sweden, which measures the probability of observing a ratio as

in the present comparative case study context, one could interpret this as a scenario in which an alternative containment strategy was randomly assigned to one of the countries in the donor pool.

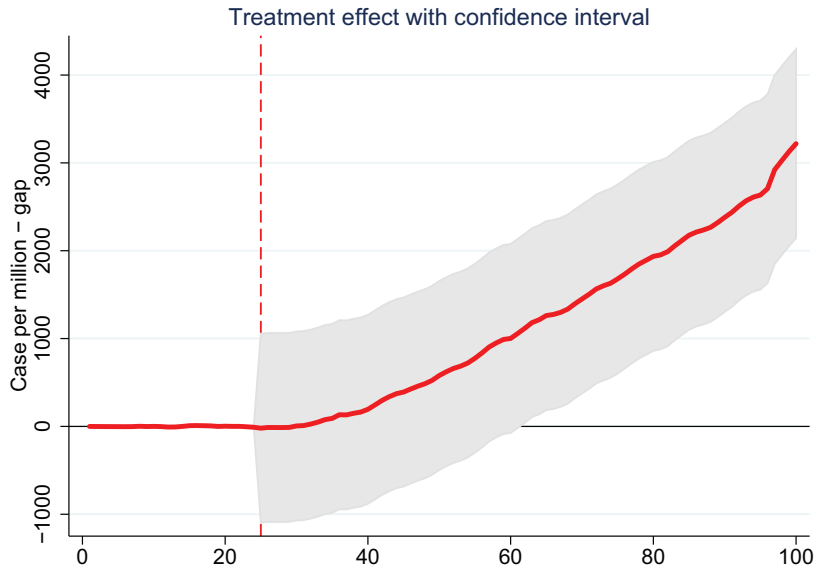

Figure 4. Ninety percent confidence interval.
Notes: Red line is the average treatment effect and the grey area is the confidence set. Vertical dashed line indicates the date of policy intervention.

high as the one obtained for Sweden if one were to pick a country at random from the list of potential controls. In this case, an exact *p*-value for Sweden is 0.097, as it ranks third out of 31 countries, which falls within the conventional range of statistical significance used in the SC literature.[7]

Inverting the permutation test statistics allows me to construct uniform confidence sets for the treatment effect function with equal weights on all observed units. Using the methodology first proposed in Firpo and Possebom (2018) and further elaborated in Ferman, Pinto, and Possebom (2020), I compute confidence sets for the treatment effect by considering functions that deviate from the estimated treatment effect by an additive and constant factor and are not rejected by the placebo test. As presented in Figure 4, although I cannot reject treatment effects that are initially positive, the treatment effect function in the confidence set turns significantly positive over time in the long run. A quick visual assessment shows that this significance occurs around Day 60 onward, or 5 weeks after the implementation of the lockdown measures. On one hand, this result is consistent with BDM, which looked at the first 5 weeks of the lockdown measures and concluded that the mandatory lockdown would not have made significant differences in the infection rate in Sweden. However, expanding the horizon over the entire lockdown period, the epidemiological impact of lockdown measures takes places with a time lag and eventually becomes more visible in the longer horizon.

[7] An underlying assumption of the inference test is that if the estimated effects for Sweden are vastly different, then I reject the null hypothesis of no effect whatsoever. As such, rejecting this exact null hypothesis does not rule out the possibility of countries with a sizeable treatment effect—either positive or negative—for parts of the post-treatment period.

Table 3. Leave-one-out test.

	(1) Benchmark	(2) No FIN	(3) No GRC	(4) No NOR	(5) No DNK	(6) No EST
Chosen donors	FIN (0.489), GRC (0.242), NOR (0.216), DNK (0.033), EST (0.020)	*HUN* (0.297), *BGR* (0.282), NOR (0.26), EST (0.09), *LTU* (0.065), *LVA* (0.005)	FIN (0.748), NOR (0.162), DNK (0.088), EST (0.002)	FIN (0.87), DNK (0.119), *LUX* (0.011)	FIN (0.51), GRC (0.224), NOR (0.204), EST (0.062)	FIN (0.426), NOR (0.262), *BGR* (0.242), DNK (0.039), *SVK* (0.031)
Population density	37.327	59.901	28.135	34.960	32.759	36.096
Urban fraction (%)	82.927	74.667	85.061	85.740	82.303	81.165
Household size	2.198	2.308	2.117	2.103	2.200	2.199
Average deaths	0.159	0.348	0.058	0.109	0.134	0.158
Cases (Day 7)	8.196	10.084	6.867	5.762	7.580	9.676
Cases (Day 14)	61.382	62.937	61.376	60.627	61.262	61.180
Cases (Day 23)	176.858	171.181	178.855	179.599	176.549	176.696
RMSPE	4.585	3.726	5.057	5.684	4.844	4.173
p-Value	0.097	0.1	0.1	0.1	0.1	0.1

Notes: New donors not in the benchmark specification are italicised.

3.3. Robustness checks

To evaluate the credibility of the benchmark settings and results, I follow the suggestions in Abadie (2020) and conduct various robustness checks.

3.3.1. Choice of donors. One way the design of the study may influence the result comes from the choice of countries in the donor pool with positive weights assigned. If dropping one country from the donor pool creates a large effect on the outcomes without a discernible change in pre-intervention fit, a re-examination might be necessary to assess whether the change in the magnitude of the estimate is caused by large idiosyncratic shocks on the outcome of the removed country. As such, I perform a leave-one-out analysis, where I exclude from the sample, one at a time, each country that contributes to the SC in the benchmark specification. For each case, the new list of donors with positive weights, as well as the values of the predictors and pre-treatment fit, are shown in Table 3. In the benchmark, Finland was the country assigned with the largest weight, followed by Greece and Norway. Dropping Finland from the donor pool generates a new set of donors, with large weights on Hungary and Bulgaria, followed by Norway and the Baltic countries. Although this synthetic unit does not match the demographic profile of Sweden as well, it does match pre-treatment outcomes better, with a lower RMSPE than the benchmark case.[8] On the other hand, dropping other countries from the benchmark donor list produced new donor lists with higher weights assigned to Finland or other Nordic countries. As shown in Figure A2 in the Appendix, the estimated gaps in the infection rate under the leave-one-out re-analysis all closely revolve around the benchmark estimate. Finally, with regard to the significance of the results under each analysis, permutation-based inference tests indicate that all the *p*-

[8] Recent work by Botosaru and Ferman (2019) offers new insights on the tradeoffs involved in the choice of covariates in the SC estimation.

Table 4. Different predictors.

	(1)	(2)	(3)	(4)	(5)
Chosen donors	FIN (0.489), GRC (0.242), NOR (0.216), DNK (0.033), EST (0.020)	FIN (0.489), GRC (0.24), NOR (0.204), DNK (0.04), EST (0.027)	FIN (0.65), *LTU* (0.211), EST (0.103), DNK (0.032), *LUX* (0.003), *HRV* (0.001)	FIN (0.748), NOR (0.163), DNK (0.089)	*HUN* (0.521), NOR (0.173), *BGR* (0.148), EST (0.096), DNK (0.029), FIN (0.024), *LUX* (0.009)
Predictors	Population density	–	–	–	
	Urban fraction	–	–	–	
	Household size	–	–	Population	
	Average death	–	–	–	Days 1,$\cdots$,24
	Day 7	Day 8	Days 1–7	–	
	Day 14	Day 15	Days 8–14	–	
	Day 23	Day 24	Days 15–23	–	
RMSPE	4.585	4.567	5.472	5.057	3.526
p-Value	0.097	0.097	0.097	0.129	0.129

Notes: Specification (1) refers to the benchmark. Specification (2) and (3) replace benchmark predictors with different dates and sub-period averages, respectively. Specification (4) replaces household size with population. Specification (5) matches pre-treatment cases only. Cells marked "–" indicate predictors that are identical to the benchmark. New donors not in the benchmark specification are italicised.

values are within the threshold of 10 percent as Sweden's ranking in the RMSPE ratios remain unchanged.

3.3.2. Selection of predictors. In another robustness check, I test the choice of predictors along both demographic covariates and pre-treatment infection profiles. The first test concerns the choice of three lagged infection outcomes. I change the respective dates one day forward to match infection cases on Days 8, 15, and 24. The second check replaces lagged dates with lagged period averages by splitting the pre-treatment period into the first two weekly averages and the last 9-day average. The third check concerns the choice of the demographic covariates. Because BDM included population size as a predictor, I replace average household size variable with the total population in the list of predictors. Finally, Doudchenko and Imbens (2016) suggested that lagged outcomes tend to be substantially more important than other covariates in terms of predictive power. Labelled as constrained regression, their approach has become popular recently, as it also mitigates concerns of specification search. Following this advice, I estimate donor weights on the basis of pre-treatment outcomes only.

The complete list of new predictors is shown in Table 4. Specification (1) refers to the benchmark case. Specifications (2) through (4) keep the set of predictors with a different balance on demographic covariates and pre-treatment outcomes. As such, the estimated weights and the chosen control units do not change much, with Finland still receiving the largest weight. On the other hand, when I match only pre-treatment outcomes in specification (5), the chosen donors change quite drastically, with Finland being replaced by Hungary as the country with the largest weight. It is not surprising, however, that the pre-treatment fit, as shown in the RMSPE, is the lowest

Table 5. More robustness checks.

	(1)	(2)	(3)	(4)
Chosen donors	FIN (0.489), GRC (0.242), NOR (0.216), DNK (0.033), EST (0.020)	FIN (0.593), *LTU* (0.208), NOR (0.089), EST (0.078), *LVA* (0.02), *HRV* (0.006), *ISL* (0.005), *ESP* (0.001)	FIN (0.422), *BGR* (0.255), NOR (0.229), EST (0.06), *LTU* (0.024), *HRV* (0.01)	FIN (0.513), DNK (0.306), *ISR* (0.155), *LUX* (0.025)
RMSPE	4.585	4.324	4.315	0.136
p-Value	0.097	0.129	0.065	0.323

Notes: Specification (1) refers to the benchmark. Specification (2) backtracks treatment to Day 22 while specification (3) applies SC to demeaned series. Specification (4) transforms infection cases into logs. New donors not in the benchmark specification are italicised.

in this case. The treatment estimates, as depicted in Figure A3 in the Appendix, show that the estimates for different specifications mostly revolve around the benchmark estimates. Inference tests using the permutation technique in Section 3.2 show that the *p*-values in specifications (2) and (3) remain below the 10 percent threshold, resulting in a rejection of the exact null hypothesis. For specifications (4) and (5), on the other hand, the treatment effect estimates are not significant at the 10 percent level with a *p*-value of 0.129.[9]

3.3.3. More robustness checks. In the last round of robustness checks, I consider three more variations to the original benchmark specification. First, I apply an alternative treatment date that is 3 days earlier than the baseline specification. Lagged outcomes in the set of predictors were accordingly modified to infection cases on Days 7, 14, and 20. Next, I calculate pre-treatment averages and apply SCs on demeaned outcomes. According to Ferman and Pinto (2019), this specification is equivalent to adding an intercept and can be directly comparable to a DD approach. Finally, I convert the infection cases in logs to check whether the outcome is invariant to functional form transformations.

The outcomes for the last set of robustness check are summarised in Table 5, while Figures A4–A6 in the Appendix show the profiles of infection cases for the alternative specifications. With an earlier assignment of lockdown treatment, as shown in specification (2), the estimated weights do not change much, and Finland is still the donor closest to Sweden. In addition, more weights are assigned to other nearby countries, such as Lithuania and Estonia. Although the pre-treatment fit improves with a lower RMSPE, the *p*-value in the permutation-based inference test is higher than the threshold of 0.1, and thus I am unable to reject the null hypothesis.[10] On the other hand, with demeaned series in specification (3), there is an improvement in the pre-treatment fit and a lower *p*-value of 0.065 in the inference test to reject the null hypothesis. Finally, in specification (4), scaling of infection cases into logs significantly reduces the magnitude of post-intervention

[9] This reflects Sweden's ranking in the RMSPE ratios falling from third to fourth out of 31 observations. In specification (4), Sweden and the placebo Italy swap their ranks, whereas in specification (5), Sweden is right behind the placebo Czech Republic, whose treatment effect is predominantly negative.

[10] A closer look at the distribution of RMSPE ratios shows that the fall in Sweden's ranking from third to fourth is caused by the placebo The Netherlands, whose treatment effect is predominantly negative.

divergence, as shown in Figure A6. As such, this leads to a high *p*-value that prevents me from rejecting the exact null hypothesis.

4. DISCUSSION

4.1. Infection to mortality

So far, the focus of the analysis has been the rate of infection. One caveat of the analysis using infection cases to assess the impact on the spread of COVID-19 is that testing policies were quite idiosyncratic across countries in terms of eligibility and accessibility. Because of these endogenous differences in testing rates, there are limitations in using infection cases to assess true epidemiological effects. Although this issue is difficult to resolve under current data availability, one could investigate death counts and compare how the NPIs impacted mortality rates during the COVID-19 crisis.

Although national health protection agencies report daily death counts, some jurisdictions include both confirmed and probable cases and deaths, while others only report confirmed cases. As such, daily reported figures for deaths are equally difficult to compare and qualify across countries, which is further exacerbated by differences in health care and aged care systems. Instead, I use excess mortality rate—the ratio of numbers of deaths over and above the historical average between 2015 and 2019—as a more reliable source of information for comparison. The Short-Term Mortality Fluctuation (STMF) data series from Human Mortality Database offers weekly death counts by age groups and sex for 22 countries, including Sweden as well as most—but not all—countries assigned with positive weights in the construction of synthetic Sweden in Section 3.1.[11] This allows me to generate weekly excess mortality for synthetic Sweden and compare that with the profile of actual Sweden, which is shown in Figure 5. As a reference, the left panel shows cumulative death counts per million population taken from Our World in Data on a weekly basis from Week 9 (starting February 23), with Week 13 (starting March 22) corresponding to the week of policy intervention. The right panel shows excess mortality rates for Sweden and synthetic Sweden. Before Week 13, there is no discernible difference in the excess mortality rates between the two groups. Close to Week 13, however, the excess mortality rate rises more steeply in Sweden and remains consistently higher than its synthetic counterpart. At its peak, the mortality rate in Sweden is more than 40 percent above its historic average, while the corresponding peak for the synthetic unit is around 10 percent above the historic average. On average, as summarised in the first row of Table 6, the excess death rate over the twelve weeks of the post-intervention period in Sweden is 23.4 percent higher than its historic average. In contrast, the corresponding gap in synthetic Sweden is 3.1 percent. In other words, assuming synthetic Sweden is a comparable counterfactual, the excess death rate would have been more than 20 percentage points lower had Sweden followed policies similar to those adopted by its parallel counterpart. A simple back-of-the-envelope calculation of the DD shown in the last column of the table suggests a gap of 25 percentage points.

Because the database provides mortality information by age, I extend the same analysis across different age groups as shown in Figure 6. A visual inspection shows that the gap in excess mortality between the two units after the lockdown becomes significantly more pronounced for

[11] Among the countries with positive weights assigned in the benchmark synthetic unit, there is no excess mortality data for Greece. As such, I refer to synthetic Sweden constructed from the donor pool that excludes Greece. This corresponds to specification (3) in Table 3.

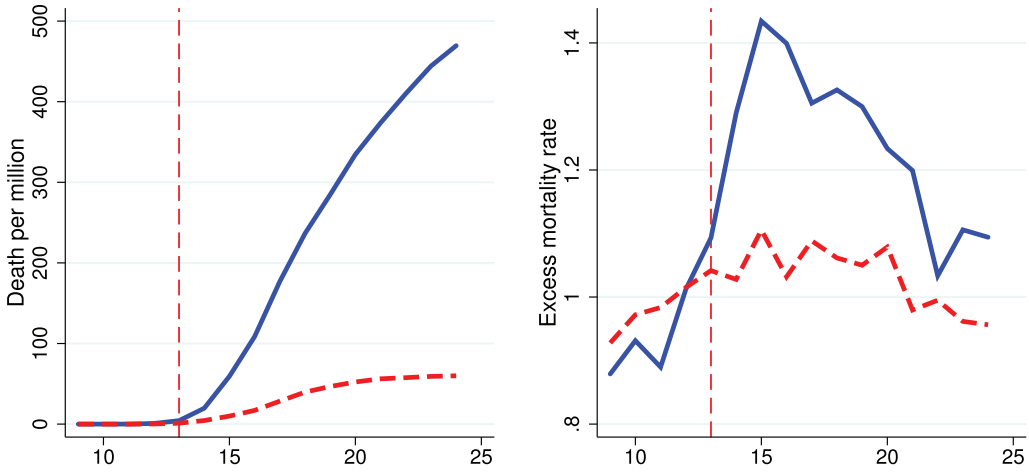

Figure 5. Mortality statistics.

Notes: Horizontal axis measures calendar weeks of 2020 starting from Week 9 (starting February 23) until Week 24 (ending June 13). Vertical dashed line in red denotes the week of policy intervention in Week 13 (March 22–28). Sweden is shown in the solid blue line, whereas synthetic Sweden is shown in the red dashed line.

Table 6. Excess mortality before versus after lockdown.

	Before lockdown (Weeks 9–12)			After lockdown (Weeks 13–24)			
	Sweden	Synthetic Sweden	(1)	Sweden	Synthetic Sweden	(2)	(2)–(1)
Total population	0.928	0.975	−0.047	1.234	1.031	0.203	0.250
Age 15-64	0.908	0.941	−0.033	1.050	0.941	0.110	0.143
Age 65-74	0.925	0.988	−0.064	1.150	1.030	0.120	0.184
Age 75-84	0.884	0.897	−0.013	1.201	0.969	0.232	0.245
Age 85 plus	0.917	0.926	−0.010	1.271	0.994	0.277	0.287

Notes: Columns labelled (1) and (2) measure the difference in excess mortality rates between Sweden and synthetic Sweden for each sub-period. Column labelled "(2)–(1)" measures the DD.

older age cohorts. As presented in Table 6, this gap grows from around 11 percentage points among the working-age cohort to approximately 28 percentage points for the elderly cohort aged 85 years and above. Applying the simple DD calculation, the corresponding gaps range from 14 to 29 percentage points, respectively.

4.2. Voluntary social distancing or involuntary lockdown?

Naturally, infection dynamics are dependent not only on lockdown measures. There were indeed signs that the public was already taking precautionary actions before various lockdown measures. For example, as infection cases grew, people made more trips to grocery stores and pharmacies to stock up on essential items. On the other hand, although the Swedish government allowed many

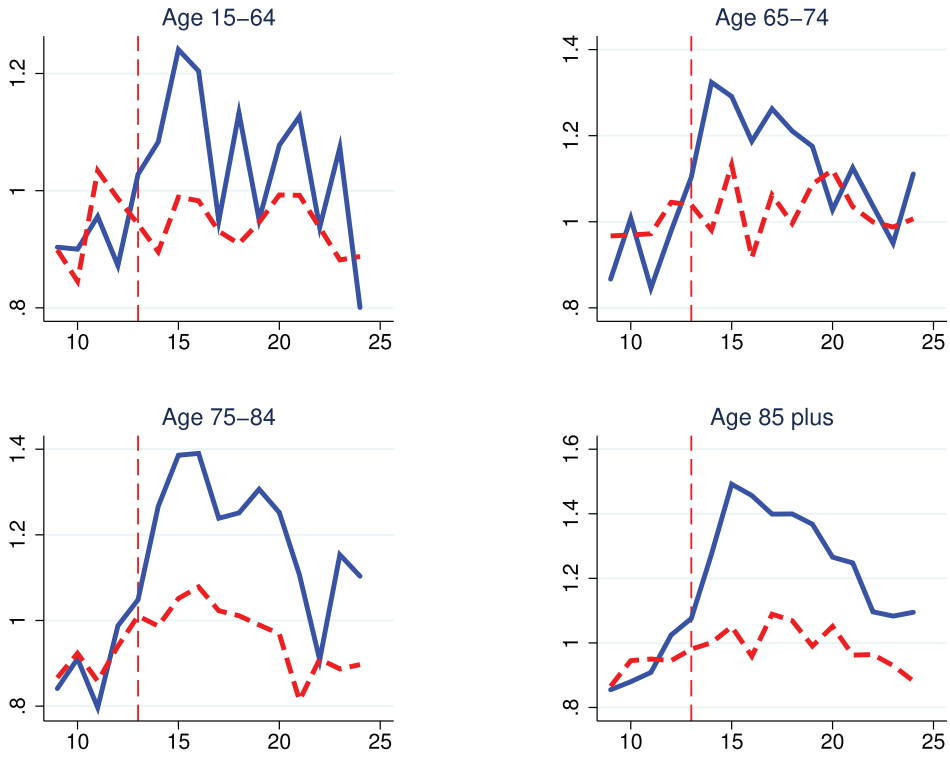

Figure 6. Excess mortality by age.
Notes: Horizontal axis shows calendar weeks of 2020 from Week 9 until Week 24. Vertical dashed line in red denotes the week of policy intervention in Week 13. Sweden is shown in the solid blue line, whereas synthetic Sweden is shown in the red dashed line.

businesses to remain open, most people stayed at home or followed social distancing protocols. Born, Dietrich, and Müller (2020) speculated that voluntary social distancing essentially had the same impact as a mandatory lockdown in the initial weeks, while Gupta et al. (2020) found that a large share of the fall in mobility was not induced by strong stay-at-home mandates. In fact, using its location services, Google provides mobility trends by geography across different categories of places, such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential areas.[12] The mobility trends for Sweden and the benchmark synthetic Sweden are shown in Figure 7, where the baseline—shown as zero in the vertical axis—is the median value for the corresponding day of the week during the 5-week period between January 3 and February 6, 2020.[13]

A first glance at mobility patterns shows a similar trend between Sweden and synthetic Sweden. Except for park visits, there is a significant drop in visits to groceries, transit stations, workplaces,

[12] https://www.google.com/covid19/mobility/

[13] One should bear in mind that the Google data rely on a group of the population who enables Google location services on their mobile devices and that this database has not been widely used in social science research, as mentioned in Gupta et al. (2020).

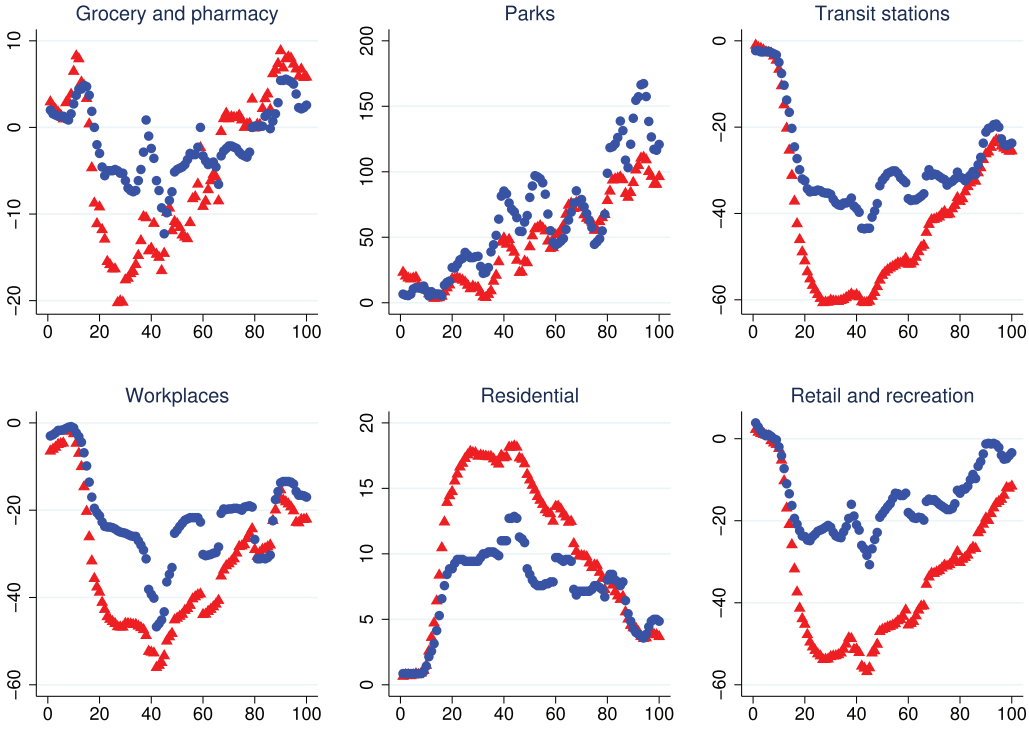

Figure 7. Mobility trends.
Notes: Horizontal axis measures days since February 29, 2020. Seven-day moving averages for Sweden and synthetic Sweden are shown in blue and red dots, respectively.

and outdoor retail in the month of March. A noteworthy uptick is shown in trips to groceries and pharmacies in both units shortly before they fall. On the contrary, a steep increase in movements within residential areas reflects more people staying at home. Comparing the two units, the magnitude of drop in mobility in March is more pronounced in synthetic Sweden, and this may reflect different behavioural patterns due to voluntary and involuntary lockdown measures. From April onward, the trend reverses, and by the end of May, mobility in both units tends to converge and revert to pre-COVID levels.

Given that mobility levels fell in Sweden as a result of voluntary precaution, one could impute this to the delay in the divergence of infection rate in Sweden relative to its synthetic counterpart. However, in the longer horizon, infection rates diverge significantly despite voluntary social distancing protocols. To better control for this behavioural change, I now consider a DD approach to complement earlier findings in the SC analysis.

5. EMPIRICAL APPROACH: DIFFERENCE-IN-DIFFERENCES (DD)

A more standard setting is applied for the DD analysis, where treatment is assigned to multiple countries that opted for a hard lockdown, and Sweden is assigned as the control unit. This

framework also allows different treatment times specific to each lockdown country. Although there are several benefits to this arrangement, one should bear in mind that there may also be issues of misspecification with the possibility of a heterogeneous treatment effect.[14] For treatment units, I choose the five donor countries with positive weights in the benchmark synthetic Sweden for a direct comparison between the DD analysis and the benchmark SC results. Although most DD applications customarily cluster standard errors at the aggregate level, there are too few units of observation to follow clustering methods. As such, I report heteroscedasticity robust standard errors.

Taking a standard DD approach pioneered by Card and Krueger (1994) to quantify the effects of lockdown measures, I consider the following two-way fixed-effects specification:

$$Y_{it} = \theta L_{it} + X'_{it}\beta + \nu_t + u_i + \epsilon_{it}, \tag{5.1}$$

where Y_{it} denotes infection cases in country i at time t, and L_{it} is the binary lockdown treatment index.[15] X captures additional control variables that include various mobility trends, as shown in Section 4.2, as well as the SI. ν_t is a time fixed effect, and u_i controls for country fixed effects, while ϵ_{it} denotes the error term. Our focus is on the coefficient θ, which essentially captures the causal impact of the lockdown policy on the infection rate dynamics. A negative estimate would imply a lower post-lockdown infection rate in the treatment countries in comparison with Sweden.

5.1. Results

The regression DD estimates for different specifications of the model in equation (5.1) are presented in Table 7.[16] The first row shows average treatment effects for all specifications. Specification (1) considers the baseline effect of lockdown without any controls. The estimated coefficient shows that, on average, post-intervention infection cases in the lockdown countries were lower than in Sweden by a magnitude of 750 infection cases per million population. Specification (2) replaces time fixed effect with a common time trend, and though the average treatment effect is still significant, its magnitude drops as infection cases grow over time. Specification (3) includes the continuous stringency index into regressors to control for different intensities in government responses, while specifications (4) to (9) individually control for changes in the Google mobility index. The main findings on the average treatment effects remain robust and even stronger in magnitude than the baseline result when additional country-specific mobility trends and government responses are controlled for.[17] Finally, allowing for all covariates, as shown in specification (10), shows that post-intervention infection cases in the lockdown countries were lower than in Sweden by around 470 cases per million population. These results confirm earlier findings that stricter containment measures are associated with limiting the spread of the COVID-19 infection.

[14] Although this paper does not explore such issues in depth, recent papers such as de Chaisemartin and D'Haultfoeuille (2020) and Goodman-Bacon (2018) address models with heterogeneous treatment effects.

[15] Time horizon in the DD setup is slightly different from the SC framework, as I use daily observations since February 29.

[16] As a robustness check, I convert variables expressed in first differences and report the outcomes in Table A2 in the Appendix.

[17] As for individual covariates, I refrain from making causal interpretation on the basis of contemporaneous correlation. For example, initial drop and subsequent recovery in mobility was also associated with a large spike and gradual slowdown in infection cases.

Table 7. DD estimates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Lockdown	− 751.10	− 291.52	− 808.38	− 791.43	− 720.86	− 876.02	− 869.76	− 851.07	− 859.91	− 468.80
	(90.75)	(60.60)	(81.25)	(84.66)	(87.50)	(83.59)	(85.67)	(83.88)	(96.57)	(68.20)
Time		23.65								25.30
		(1.10)								(1.34)
SI			28.27							9.86
			(2.78)							(2.09)
Grocery				− 22.11						− 15.53
				(3.00)						(2.82)
Park					1.53					2.84
					(0.83)					(0.73)
Transit						− 27.81				16.59
						(3.40)				(5.06)
Workplace							− 29.03			19.10
							(3.42)			(5.65)
Residential								87.66		175.83
								(10.46)		(28.41)
Retail									− 12.70	32.56
									(3.01)	(3.54)
N	600	600	600	600	600	600	600	600	600	600
Adj. R^2	0.773	0.780	0.835	0.800	0.774	0.818	0.814	0.826	0.786	0.845

Notes: Robust standard errors in parentheses.

5.2. Anticipatory and time-varying effects

The key identifying assumption of DD regression design is the common trend assumption, meaning that, in the absence of treatment, the average change for the treated group would have been identical to the observed average change for the control group. In our setup, this implies that infection trends would have been the same had the lockdown countries followed the same strategy as Sweden did. One strategy to deal with this issue—referred to by Autor (2003) as a "placebo" test—is to examine the possibility that exposures to future treatment are anticipated by current outcomes by including lead terms in the baseline regression. If the coefficients for the leads, shown as γ in equation (5.2) below, are very close to zero, we can expect that future policy changes are unlikely to be associated with current outcomes. Here, I include three weekly leads right until the day of lockdown in each treatment country.

For post-treatment effects, on the other hand, I have implicitly assumed that the coefficient θ in equation (5.1) is constant over time. However, the impact of lockdowns could be immediate or lagged over time and may possibly vary with different intensities. To explore the time-varying effects of the lockdown measures, I also allow for lagged treatment variables in the regression specification, as suggested by Autor (2003) and Wing, Simon, and Bello-Gomez (2018). More specifically, I add a dummy variable for each week up to the seventh week after the lockdown, as well as a dummy that captures all the periods after the eighth week since lockdown. The modified specification examining both anticipatory and time-varying effects combined is:

$$Y_{it} = \sum_{s=1}^{S} \gamma_s L_{i,t+(7\times s)} + \sum_{m=0}^{M} \theta_m L_{i,t-(7\times m)} + \nu_t + u_i + \epsilon_{it}. \tag{5.2}$$

Here, the first sum on the right-hand side captures weekly leads, while the latter sum captures weekly lags, with the first term ($m = 0$) representing the immediate treatment effect in the first 7 days of lockdown. If the initial effect of the policy is negative, then negative values of subsequent θ_m ($\forall m > 0$) imply that the initial effect amplifies over time, whereas positive values would imply that the initial impact fades with time.

Figure 8 plots the estimated leads and lags, running from 3 weeks ahead to 8 weeks behind with 95 percent confidence intervals. In the week leading to the day before the policy implementation, the coefficient is in fact significantly different from zero and negative. Given that our lockdown treatment is defined as the maximum accumulation of containment measures, this notion of anticipatory effect in the preceding week is not surprising, as earlier policy measures may have taken effect and additional measures could be associated with current outcomes. On the other hand, the coefficients for 2- and 3-week leads are not significantly different from zero and provide some confidence that the policy intervention occurs before its effect. For post-treatment periods, the coefficients are significantly negative from the onset of lockdown implementation and monotonically decrease over time.

6. CONCLUSION

Policymakers have implemented a wide range of NPIs to fight the spread of COVID-19. Using variation in policies across countries and over time, I use a flexible statistical methodology for data-driven case studies—the SC method—to investigate the causal effects of counter-COVID policies. I find that the lockdown measures played an important role in limiting the spread of the COVID-19 infection and that Swedish policymakers would have reduced the infection cases

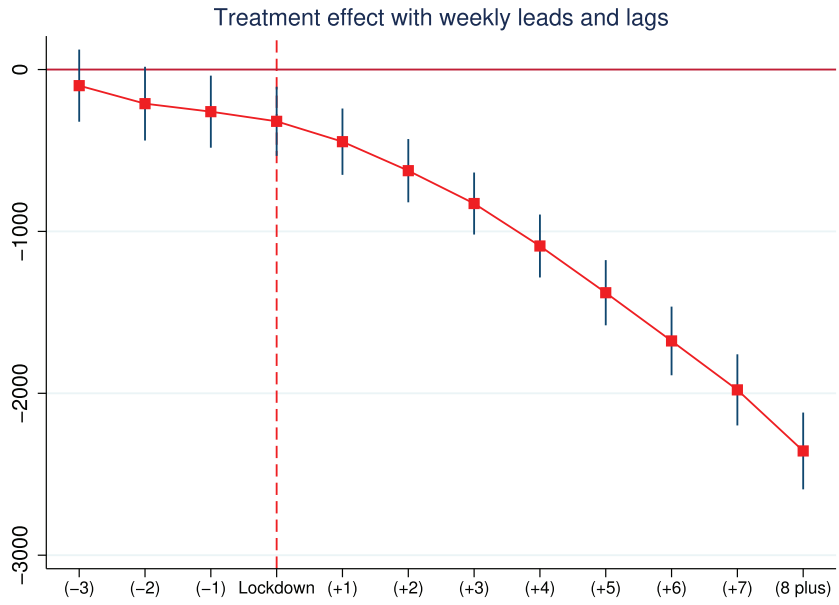

Figure 8. Anticipatory and time-varying effects.
Notes: Horizontal axis shows weekly leads and lags from the day of lockdown.

and mortality rates in the long run had they followed policies similar to those implemented elsewhere in Europe. Supplementing the main analysis with several robustness checks, as well as one that uses an alternative DD research design, does not fundamentally alter the main findings.

My analysis has some limitations and caveats. First, the COVID-19 policies are not randomly assigned, and, in many cases, government policies were in direct response to past and current epidemiological conditions but also to contain future spread, which complicates causal identification. As such, caution is needed to claim that the decision to go lockdown-free was completely independent from other considerations. Next, various policies were introduced over a short timeline across countries, making it difficult to compare and assess the intensity of treatment. In particular, this can become a source of bias when using an all-inclusive response index in the current context. For example, mobility trends in visits to grocery stores and pharmacies indicate possible behavioural changes in response to stay-at-home announcements, whereas issuance of travel restrictions ahead of time may induce possible spillover effects. Discussions by Goodman-Bacon and Marcus (2020) and Gupta et al. (2020) provide several challenges and suggestions in the empirical research design of COVID-19 policies. Finally, the present study abstracts from exploring individual policies and how their timing can have different epidemiological impacts. A worthwhile project to pursue would be one similar to Chernozhukov, Kasahara, and Schrimpf (2020), who investigated the impact of individual measures along both epidemiological and economic aspects. Such explorations would better inform policymakers seeking to protect public health and facilitate an eventual economic recovery.

ACKNOWLEDGEMENTS

I thank the Editor and two anonymous referees for their insightful comments. I am also indebted to Seojeong (Jay) Lee, Nicola Aravecchia, Hansoo Choi, and Julián P. Díaz for their constructive feedback and comments. I also thank Charles Wyplosz for disseminating an earlier version of the paper at Covid Economics, Vetted and Real-Time Papers, Issue 35. All remaining mistakes are my own.

REFERENCES

Abadie, A. (2020). Using synthetic controls: feasibility, data requirements, and methodological aspects. *Journal of Economic Literature* (forthcoming).

Abadie, A., A. Diamond and J. Hainmueller (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association 105*, 493–505.

Abadie, A., A. Diamond and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*, 495–510.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: a case study of the Basque country. *American Economic Review 93*(1), 113–32.

Andersen, A. L., E. T. Hansen, N. Johannesen and A. Sheridan (2020). Pandemic, shutdown and consumer spending: lessons from Scandinavian policy responses to COVID-19. Papers 2005.04630, (arXiv.org).

Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*, 431–97.

Autor, D. H. (2003). Outsourcing at will: the contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of Labor Economics 21*, 1–42.

Baldwin, R. and B. W. di Mauro (2020). *Economics in the Time of COVID-19*. CEPR Press.

Born, B., A. M. Dietrich and G. J. Müller (2020). Do lockdowns work? A counterfactual for Sweden. *Covid Economics 16*, 1–22.

Botosaru, I. and B. Ferman (2019). On the role of covariates in the synthetic control method. *Econometrics Journal 22*, 117–30.

Card, D. and A. B. Krueger (1994). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review 84*(4), 772–93.

Castex, G., E. Dechter and M. Lorca (2020). COVID-19: cross-country heterogeneity in effectiveness of non-pharmaceutical interventions. *Covid Economics 14*, 175–99.

Chen, X. and Z. Qiu (2020). Scenario analysis of nonpharmaceutical interventions on global Covid-19 transmissions. *Covid Economics 7*, 46–67.

Chernozhukov, V., H. Kasahara and P. Schrimpf (2020). Causal impact of masks, policies, behavior on early Covid-19 pandemic in the US. *Covid Economics 35*, 116–76.

Conyon, M. J., L. He and S. Thomsen (2020). Lockdowns and COVID-19 deaths in Scandinavia. *Covid Economics 26*, 17–42.

de Chaisemartin, C. and X. D'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Doudchenko, N. and G. W. Imbens (2016). Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. Working Paper 22791, National Bureau of Economic Research.

Ferman, B. and C. Pinto (2019). Synthetic controls with imperfect pre-treatment fit. Papers 1911.08521, (arXiv.org).

Ferman, B., C. Pinto and V. Possebom (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management 39*, 510–32.

Firpo, S. and V. Possebom (2018). Synthetic control method: inference, sensitivity analysis and confidence sets. *Journal of Causal Inference 6*(2), 1–26.

Fisher, R. A. (1935). *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Gonzalez-Eiras, M. and D. Niepelt (2020). On the optimal 'lockdown' during an epidemic. *Covid Economics 7*, 68–87.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Working Paper 25018, National Bureau of Economic Research.

Goodman-Bacon, A. and J. Marcus (2020). Using difference-in-differences to identify causal effects of COVID-19 policies. *Survey Research Methods 14*, 153–58.

Gupta, S., T. D. Nguyen, F. L. Rojas, S. Raman, B. Lee, A. Bento, K. I. Simon and C. Wing (2020). Tracking public and private responses to the COVID-19 epidemic: evidence from state and local government actions. Working Paper 27027, National Bureau of Economic Research.

Hale, T., S. Webster, A. Petherick, T. Phillips and B. Kira (2020). Variation in government responses to COVID-19. BSG Working Paper Series 2020/032.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*, 5–86.

Rocklöv, J. and H. Sjödin (2020). High population densities catalyse the spread of COVID-19. *Journal of Travel Medicine 27*, 1–2.

Roser, M., H. Ritchie, E. Ortiz-Ospina and J. Hasell (2020). Coronavirus pandemic (COVID-19). *Our World in Data*, https://ourworldindata.org/coronavirus.

Sá, F. (2020). Socioeconomic determinants of Covid-19 infections and mortality: evidence from England and Wales. *Covid Economics 22*, 47–58.

Ullah, A. and O. A. Ajala (2020). Do lockdown and testing help in curbing COVID-19 transmission?. *Covid Economics 13*, 138–56.

United Nations, Department of Economic and Social Affairs, Population Division(2019). Patterns and trends in household size and composition: evidence from a United Nations dataset. (ST/ESA/SER.A/433).

Wing, C., K. Simon and R. A. Bello-Gomez (2018). Designing difference in difference studies: best practices for public health policy research. *Annual Review of Public Health 39*, 453–69.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Victor Chernozhukov handled this manuscript.