COVID-19 counterfactual evidence.
Estimating the effects of school closures

Marco Giuliani

Orcid**:** 0000-0002-6927-7177

Department of Social and Political Science, Università degli studi di Milano, Via Conservatorio 7, 20122 Milano (Italy); +39.02.50321210

Marco.giuliani@unimi.it

Abstract

Scholars have started to estimate the effects of non-pharmaceutical interventions to reduce the health impact of COVID-19. However, the empirical evidence is highly contested, and since it is not known exactly what would have happened without those measures, political élites are left free to give credit to the voices that they prefer the most. We argue that any sensible assessment of the effectiveness of anti-COVID policies requires methodological reflection on what is actually comparable, and how to approximate the ideal "method of difference" theorized by John Stuart Mill. By evaluating the effectiveness of school closures as anti-COVID policy, we provide two examples in which appropriate counterfactuals are inductively discovered, rather than selected *a priori*. In the first one we use Coarsened Exact Matching in a cross-country setting, while in the second one we implement the Synthetic Control Method in a within-country analysis. The article highlights the methodological advantages of including this type of techniques in the toolbox of policy scholars, while both examples confirm the effectiveness of school closures.

Keywords:
Evaluation; Comparative analysis; Counterfactual; Covid-19; School closures

Funding:
This work was supported by the Università degli studi di Milano [PSR20120]

1 Introduction

One of the main divides in the fight against the spread of Coronavirus, especially in the first phases of the pandemic, has been the one between countries choosing a more voluntary approach, based on recommendations, and those countries that have opted for stricter regulations, including school closures, travel bans and lockdowns (Capano et al., 2020; Ceron et al., 2021). In Europe, the benchmark for the first type of approach has been the Swedish model devised by the state epidemiologist Anders Tegnell, who explicitly avoided imposing severe restrictions, especially regarding children's school attendance, mask-wearing, and lockdowns (Andersson and Aylott, 2020; Pierre, 2020; Wenander, 2021).

The debate on the preferability of this liberal approach has assumed multiple nuances, and it is also for this reason that the model has attracted so much political interest. Assessment of its effectiveness in containing infections and deaths, at least in the long run, has obviously been at the centre of the arguments, producing a polarized public discussion. For many, the most natural comparison for what the media have dubbed the "Swedish experiment" is with the other Nordic countries, which have adopted a more restrictive precautionary approach (Gordon et al., 2021). Given the better epidemiological records of these reference countries, restrictive policies can be considered to have reduced the incidence of the virus and the number of people who have died because of COVID. But others, including Tegnell himself, have suggested that Sweden should be matched with countries with similar populations, since that variable is relevant *per se* for different spreading dynamics. Sweden has approximately the same number of inhabitants as Hungary, Portugal and Greece, but its cumulated excess mortality rate is lower than theirs, and it is even lower than those of smaller well-governed countries like Austria and Switzerland.

Clearly, "numbers are not facts [and] don't speak by themselves" (Stone, 2020), and to attribute a causal effect to any policy we need more than some arbitrary paired comparisons. The reason derives from the famous second canon of John Stuart Mill, which governs the so-called “method of difference”:

> *If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon.* (Mill, 1843: 455)

In the absence of a perfect counterfactual, social scientists should try to approximate that ideal comparison given the observational data available. But once again the question is this: which countries are better suited to acting as counterfactuals for assessing the Swedish COVID policy? Are they Nordic countries, similar for geography and socio-economic characteristics, or some other West or South-European countries, similar for size of the population? In fact, a major problem in assessing the impact of any policy, and tracing the causal relationship between a set of government decisions and their outcomes, is that of identifying appropriate counterfactuals and avoiding the risks of selection bias (Costalli and Negri, 2021). Selection biases are common to both qualitative and quantitative research designs, since they simply reflect the potentially relevant differences between the compared units receiving and not receiving the treatment, such as implementing or not implementing a certain policy (Angrist and Pischke, 2009; King et al., 1994).

Furthermore, the decision whether or not to introduce certain policy solutions also depends on the extent and intensity of the problem itself. For example, employing restrictive policies to limit the diffusion of COVID-19 is not independent from the spread itself of the virus – that is, from the outcome that one may want to explain. To use the experimental jargon, the treatment is not randomly assigned but it is endogenous to the causation process. This greatly complicates the search for appropriate observational counterfactuals, which may sometimes not even exist.

Using the case of anti-COVID policies, and especially the contested decision to close schools, this paper explores two different econometric options to overcome this limit, identifying or constructing appropriate counterfactuals, and eventually testing causal hypotheses. The paper is organized as follows. The next section reviews the different arguments in favour and against closing schools as a non-pharmaceutical intervention to restrict the diffusion of COVID-19. In section 3, we use a matching strategy to test the effectiveness of this policy by comparing 31 European countries during the first six months of 2021. In section 4, we implement the synthetic research method in a within-country comparison of the effects of school openings in Italy, one of the countries that have recorded the highest coronavirus infection tolls. Section 5 discusses the implications of these results for the community of policy scholars and concludes. An online appendix with supplementary material complements the text with descriptive statistics, alternative specifications, and robustness tests.

2 The problem: to distance or not to distance (teaching)?

When, at the beginning of 2020, COVID-19 arrived in the Western world, one of the first policy decisions taken was to shut down schools and universities. Limiting our attention to the geographical area comprising the 27 EU member states, together with the United Kingdom, Switzerland, Norway and Iceland, no country, not even Sweden (Pashakhanlou, 2021), kept all schools entirely open during the first months of the pandemic.

If we use the first component of the stringency index developed by Hale et al. (2021), specifically dedicated to the closing of schools and universities, it is possible to trace the adoption of this measure throughout the first year and a half of the pandemic. This component is an ordinal scale in which the value of zero is given to countries without restrictions; one is attributed if closures are simply recommended or if the usual organization of teaching necessitates major alterations; two when closures are decided only for certain levels or types of schools; and three is given in the case of generalized closures. Moreover, in the case of restrictions applied only to particular parts of the country, the value on the scale is reduced by half a point. Figure 1 reports the evolution of this scale, translated into a 0-100 index, for the 31 countries detailed above for the period from January 2020 to June 2021.

The first mover was Italy, in which some regional governments chose to close schools at every level already on 24 February, 2020. The decision was shortly followed by other regions until eventually, on March 4, a national decree suspended all in-person teaching. Italy acted first because it was the first European country to be heavily hit by the virus, but the others soon took similar decisions. Twenty-six countries out of thirty-one enacted a general closure before the end of March 2020, although some of them kept distance teaching and learning until the end of the school and academic year, while others gradually relaxed the restrictions, and one country, Switzerland, entirely cancelled them before June 2020.

Figure 1 The 0-100 index of school closing in 31 European countries in the period between January 2020 and June 2021

In autumn 2020, with the opening of the new school and academic year, many countries again had to resort to closures in concomitance with the upswing in infections of the second wave of the pandemic, but this time the picture was much more differentiated and nuanced, as highlighted by the large cross-country and longitudinal variation of the index in Figure 1. Governments were once again faced with the difficult choice between delaying the return to normality for millions of children and adolescents, and risking a further upsurge in the number of new infections. Some countries, like Latvia and Finland, chose a rather stable policy halfway between full closure and full liberalization. Others, like Portugal and Greece, constantly adapted their restrictions in response to the signals coming from the continuous monitoring of the health situation.

Political, administrative and institutional factors account for the differentiated strategies (Capano et al., 2020; McConnell and Stark, 2021; Toshkov et al., 2021), but their public justification has been facilitated by the uncertain epidemiological evidence. Indeed, the first studies partially disagreed on the effectiveness of these measures, to say nothing of the feared psychological and educational consequences of protracted school closures, especially for young children.

Particularly influential, at least in supporting the strategy to keep schools open for adolescents, was the report published in July 2020 by the Swedish public health agency, which concluded that "closing of schools had no measurable effect on the number of cases of COVID-19 among children" (Folkhälsomyndigheten, 2020: 14). The analysis was based on a comparison of confirmed cases in the 1-19 year old segment of the population in Finland and Sweden, with the former country implementing a generalized school closure between March and May 2020, while the latter only recommended distance-learning for higher education institutions and upper secondary schools. The lack of any effect of school openings/closings was attributed to the fact that children had comparatively lower viral loads, which reduced their risks and also their transmission capacity (Ludvigsson, 2020). Whilst the report was criticized due to the limited testing implemented at that time by Sweden, similar results were found by research based on contact tracing amongst Norwegian primary school pupils (Brandal et al., 2021).

Several other studies conducted to evaluate the effectiveness of school closures using a range of diverse methodologies have confirmed the scepticism about this kind of restrictive policy, not least because children at home may interact with each other in any case (Park et al., 2020). Simulated policy combinations on the demand for hospital services have been tested in the UK, with the conclusion that each intervention on its own is insufficient to control the circulation of the virus (Davies et al., 2020). The non-systematic contribution of school re-openings to the second wave of the pandemic after the summer break has been confirmed for Italy and Germany using large official datasets (Gandini et al., 2021; Isphording et al., 2020). Time-series Bayesian models have shown that school closures in Japan did not have any mitigating effect on the transmission of the infection (Iwata et al., 2020). Finally, a meta-analysis performed by Busa et al. (2021) summarizes the differences between the present COVID-19 pandemic and the better-known transmission of influenza, raising doubts as to the effectiveness of the traditional school closures usually applied to combat the latter, while underlining the resulting severe psychological costs for children and adolescents.

On the other hand, studies focused on different parts of the world and adopting a broader perspective, less centred on children and teachers, have reported more promising effects of closures. Profiting from the diverse timing of school closures in spring 2020 in different US states, Auger et al. (2020) were able to attribute to that policy the variation in COVID incidence and deaths. Rauscher and Burns (2021) ran a more fine-grained analysis in which they matched similar and neighbour counties which differed mostly for the period in which school closures were implemented, and they were also able to highlight some further conditional factors for the effectiveness of the restrictive measure A similar research design was adopted to analyse the epidemiological dynamics in Italy in autumn 2020, this time exploiting the diverse school openings in three Italian regions after the vacation period and confirming their role in spreading the virus (Tosi and Campi, 2021).

A potentially confounding element for the causal attribution of the outcomes to that specific policy is the presence of other measures or recommendations. In order to assess the effect of school policies separately from the impact of other measures, Stage et al. (2021) comparatively analysed school closures and openings in Norway, Sweden, Denmark and Germany by "generating a counterfactual projection of daily cases or hospital admissions, which accounts as much as possible for events prior to, but excluding, school closure"(3). Even when controlling for other measures, the number of new cases typically started to decrease one week after the enforcement of school restrictions, confirming the centrality of this type of policy in tackling the spread of the virus.

The studies reported above are just a sample of the abundant scholarly literature on the issue; a sample selected amongst the policy-relevant articles on the topic included in the Web of Science platform. With few exceptions, there is no specific attention to case selection, or to the meaningfulness of the implicit or explicit comparisons, because of the rush to provide an empirical foundation for evidence-based policy decisions. This lack of consideration may bias the conclusions of these quantitative analyses in the same way as the use of inappropriate reference countries for the evaluation of the Swedish anti-COVID policies discussed in the introduction.

In what follows, we introduce and exemplify two different techniques – matching and synthetic controls – that put the appropriate identification strategy at the centre of their methodological attention. We simply test one expectation against the null hypothesis, supposing that the closing of schools – once a plausible set of covariates, including other restraining policies, has been controlled for – has a negative effect on the diffusion of the virus. The research focuses on the second and third waves of the virus in Europe so as to avoid any confounding factor due to the early surprise and absence of any previous experience in coping with the emergency. The hypothesis is first assessed in a cross-country comparison, and then from a within-country perspective in Italy, one of the countries most affected by the virus. The various techniques used to identify the appropriate counterfactuals are introduced together with the specific research design used in each empirical test. We provide further information on the data and variables used in the analysis in the online appendix.

3 Matching cases in a cross-country comparison

As we have seen in Figure 1, some countries chose tougher restrictions on school attendance to counter the spread of the virus, while others did not. Were these measures effective? To answer this basic question, we retrieved weekly data on the measures taken in each of the previously listed 31 European countries, together with the incidence of new certified COVID cases, and a set of other control variables. We start by proposing in Table 1 a naïve comparison of the effects of closure policies on the incidence of new COVID cases by running two simple regression models focusing on the first six months of 2021, when the different governments already had the opportunity to define and test their strategies after the new wave of autumn 2020. Our treatment variable is the presence, the week before, of highly restrictive regulations on in-person teaching, defined as more than 75% of the scale of the school closing index presented in the previous section (Hale et al., 2021).

Table 1. Estimating the effect of closing schools on the incidence of new cases

	(1)	(2)
L. New cases (1000000)	0.92***	0.89***
	(0.03)	(0.03)
L. School closing dummy	-90.70	-94.66
	(59.60)	(60.20)
L. Other policies		-101.87
		(77.24)
L. Fully vaccinated (100)		-10.49**
		(4.07)
Density		0.10
		(0.08)
Tests (1000)		0.15
		(0.28)
Constant	67.64	186.39***
	(63.76)	(67.20)
N	775	775
Countries	31	31
R-squared	0.85	0.86

Note: Panel corrected standard errors in parentheses.
*p < 0.10 **p < 0.05 ***p < 0.01

In the first model, together with the policy for school closures, we include in the right-hand side of the equation only the lagged incidence of new cases to account for the path-dependent diffusion of the virus.[1] In the second model, we add as further controls: a) the presence (the week before) of other rigid containment policies, computed on the remaining seven subcomponents of the stringency index, in order not to attribute to the school closure what was actually achieved by other regulations; b) the lagged percentage of fully vaccinated population, which is expected to reduce the spread of the virus; c) the demographic density of the population, which, conversely, is presumed to favour its diffusion; d) the number of tests (per thousand persons), which is directly associated with the discovery of new cases. Both models are time-series cross-section linear regressions with panel corrected standard errors (Beck and Katz, 1995).

In spite of their simplicity, and thanks to the inclusion of the highly significant lagged dependent variable, the two models have a large explained variance. All variables reflect the original expectations, but only the share of fully vaccinated population adds a systematic negative effect to the second model. Demographic density and the amount of tests, with the expected positive coefficients, fail to reach standard levels of statistical significance. But what is more important for our analysis, the school closure policy does not seem to systematically reduce the diffusion of the virus, and neither do the other containment policies. This empirical result would be a powerful argument against restrictive policies, and it would certainly bolster the case of those opposing the closure of schools because of its alleged ineffectiveness.

However, the models potentially suffer from a series of shortcomings, since the decision to close schools was not randomly assigned. Simply regressing our observations assumes that they are similar in all relevant respects except for the treatment variable, whereas they are obviously not. The

[1] The "European Centre for Disease Prevention and Control' of the European Union estimates that "the incubation period for COVID-19 (i.e. the time between exposure to the virus and the onset of symptoms) is currently estimated to be between one and 14 days" (https://www.ecdc.europa.eu/en/covid-19/questions-answers/questions-answers-basic-facts). Since our dataset is composed of weekly averages, we expect the incidence of new cases to be primarily a function of the discovered cases in the week before, i. e. of its lagged value.

inclusion of control variables is a partial solution, but a more direct search for counterfactuals is needed to avoid biased interpretations of causal effects. In fact, "even when all the confounding covariates are measured […] it can be difficult to properly control for them if the distributions of the predictors are not similar across groups, [that is, if there is a] lack of *balance*" between those observations that received the treatment (the policy) and those that did not (Gelman and Hill, 2007: 200).

Let us consider the example of the impact of the incidence of new COVID cases in a certain week on that same incidence the week after. That quantity is one of the factors that policy-makers may take into account in order to decide for or against closures: the greater the incidence of new cases, the more likely governments are to decide to close schools in order to reduce the further spread of the virus. This is clearly a situation of "lack of balance" among groups on some relevant factor affecting both the treatment and the outcome. If the lagged incidence of new COVID cases perfectly and completely explained the introduction of restriction policies, we would either have high incidence and consequent school closures, or low incidence with open schools. In other words, we would be without counterfactuals, having two perfectly separated groups.[2]

The $\mathcal{L}_1$ statistic introduced by Iacus et al. (2011) provides a useful measure of the level of imbalance between the treated and the untreated group. It ranges from 0, in the case of perfect balance, to 1 in the case of perfect separation, and it can be intuitively conceived as the complement to 1 of the degree of overlap between the two distributions on a single variable or on multiple ones. The top part of Table 2 provides measures of this overall measure of imbalance, together with some further

[2] In reality, other variables contribute to the adoption of such policies. They include some political factors affecting the decision-making process – number and range of veto-players, reliance or otherwise on a parliamentary majority, and even caretaker cabinets and electoral year (results of preliminary analyses available on request). These political variables usually do not affect the outcome, once the treatment is controlled for, allowing the identification of sufficiently similar counterfactuals.

details on the distributional imbalances, for our original sample consisting of 163 country-week observations with school closures and another 612 without such a policy.[3]

Table 2. Imbalances in the raw and in the matched sample

Raw sample	$\mathcal{L}_1$	Avg	Min	Med	Max
L. New cases (1000000)	0.28	1038.90	84.98	794.7	788.16
L. Other policies	0.34	0.34	0.00	0.00	0.00
L. Fully vaccinated (100)	0.27	-5.50	0.00	-3.20	-39.35
Density	0.27	17.86	0.00	21.41	0.00
Multivariate	0.72				
Balanced sample	$\mathcal{L}_1$	Avg	Min	Med	Max
L. New cases (1000000)	0.14	64.56	84.98	86.25	-4.21
L. Other policies	0.00	0.00	0.00	0.00	0.00
L. Fully vaccinated (100)	0.08	-0.11	0.00	-0.41	0.00
Density	0.06	5.37	0.00	5.08	0.00
Multivariate	0.55				

The overall imbalance on our pre-treatment covariates is rather high, with $\mathcal{L}_1$ equal to 0.72. All factors present some degree of imbalance, which manifest themselves in the rather large difference in average between the two parallel distributions, but also in the median values and sometimes even in the minimum and maximum ones. For example, the average incidence of new COVID cases in the instances that introduced school closures was almost double – i.e. 1039 more cases per million inhabitants – than in the observations that chose not to apply restrictions. Moreover, at that time, countries which adopted that type of restrictive policy were lagging behind in terms of vaccinations, having an average 5.5% fewer fully-vaccinated inhabitants than did those that opted for a more liberal approach.

[3] Following the good practices in these circumstances, we have included in the procedure the variables that are supposed to "affect both the treatment assignment Di and, controlling for it, the dependent variable Y"; i.e., all our variables with the exclusion of testing, which is associated with the outcome but not with the policy.

Matching – and more specifically Coarsened Exact Matching (CEM) (Blackwell et al., 2009; Iacus et al., 2012) – is a procedure with which to reduce the level of imbalance in a sample in order to allow a more robust test of the causal effect of the treatment, in our case the school closure policy. By identifying sufficiently similar observations, and pruning those without appropriate counterfactuals, CEM "preprocess(es) the raw data so that the treatment group becomes as similar as possible to the control group on a set of covariates chosen by the researcher" (Negri, 2022).

We applied CEM to our original sample using the covariates listed above, obtaining a more balanced sample consisting of 530 country-week cases, 150 experiencing the school closure policy and the other 380 not. As detailed in the lower part of Table 2, the multivariate $\mathcal{L}_1$ index in the balanced sample is now equal to 0.55, 25% smaller than before.

We could have included more covariates in the process, and tweaked their thresholds looking for a better match, but we refrained from doing so for two main reasons. Firstly, that strategy would probably have produced a much larger reduction in the number of observations, with the risk of entirely losing some of the countries covered by the analysis, whereas we managed to keep all the original 31 European nations.[4] Secondly, and most importantly, the univariate values reported in Table 2 show that the matching procedure has produced a substantial reduction in imbalances on all covariates, not only in their averages but also in the whole distributions of the data, making the treated and non-treated groups much more similar in many respects.[5] For example, the average gap in the incidence of new cases has diminished from 1039 to just 65, and all the other average distances have disappeared or have been greatly reduced. The same applies to the median and maximum discrepancies.

---

[4] Sample size and balance are in fact the two criteria with which the soundness of matching should be evaluated (King et al., 2017; Nielsen, 2020). With a smaller sample we would have probably also needed a different model, like a simple OLS, while we managed to keep the cross-country time-series structure and panel corrected standard errors, allowing a better comparison between the original and the matched regression results.

[5] As stated by Blackwell et al. (2009: 531), "the absolute values [of $\mathcal{L}_1$] mean less than comparisons between matching solutions", and our reduction in imbalance is larger than the one they report in their example.

The last stage of the matching process is to estimate again the effect of the policy only on the matched cases and with the appropriate weights defined by CEM. We tackle the remaining imperfect balance of the sample by further including the original covariates as control variables, Table 3 reports the results of this final step of our estimation strategy, mirroring what we did in Table 2.

The new results show some similarities, but also some important differences compared to the previous regression. The lagged dependent variable confirms the self-sustaining character of the epidemic in both models. In model 4, the only control variable that was significant before the matching – the number of vaccinations – keeps its systematic negative association with the outcome, and is now complemented also by the expected positive association with the number of tests per thousand persons. This should not be taken for granted, since a perfectly balanced sample would do without any controls, whereas our imperfect match still needs them.

Table 3. Estimating the effect of closing schools on the incidence of new cases using the balanced sample

	(3)	(4)
L. New cases (1000000)	1.01***	0.99***
	(0.03)	(0.03)
L. School closing dummy	-208.85***	-218.13***
	(57.89)	(58.94)
L. Other policies		-65.76
		(54.94)
L. Fully vaccinated (100)		-17.00**
		(8.13)
Density		-0.01
		(0.12)
Tests (1000)		0.73**
		(0.30)
Constant	36.04	171.81
	(64.77)	(122.96)
N	529	529
Countries	31	31
R-squared	0.86	0.87

Note: Panel corrected standard errors in parentheses.
*p < 0.10 **p < 0.05 ***p < 0.01

However, what is most interesting is that the reductive effect of closing schools and universities on the incidence of new COVID cases stands out in both models as highly statistically significant and large in magnitude. All other things being equal, introducing the closure policy produces a decrease in the number of new weekly cases equivalent to 0.02% of the population, which is approximately equivalent to 3700 fewer weekly infections for a country of average size like the Netherlands. Given the average fatality ratio in the matched sample, it would also mean approximately 70 fewer weekly deaths from COVID in that same benchmark country.

Considering that fewer contagions also means a decreasing probability of others being infected week after week, it is easy to understand why this policy is so central to the debate on so-called “nonpharmaceutical anti-COVID interventions”. The empirical evidence provided by a matched sample, approximating the requirements of John Stuart Mill’s method of difference, reverses the initial naïve findings, and supports the choices taken by the most prudent policy-makers.

4 Building counterfactuals in a within-country comparison

Coarsened exact matching is but one specific method, particularly suited to continuous variables, within the family of matching techniques (Iacus et al., 2011; Nielsen, 2020). Matching helps identify the appropriate counterfactuals on which to run comparisons, but it is certainly not the only approach useful for improving the capacity to infer the causal effects of a policy. Sometimes, counterfactuals are not discovered, or identified. Instead, they are (in a sense) built by the technique itself, as in the “synthetic control method” (Abadie et al., 2015). In this approach, the counterfactual against which to compare the trajectory of the case of interest is assembled by an appropriate weighted linear combination of the other cases. It is synthetically manufactured in order to maximize its resemblance to the object of the analysis before the treatment (Abadie et al., 2010; Abadie and Gardeazabal, 2003).

The synthetic control method is ideal for analysing situations in which a certain policy has been formulated and implemented in one specific unit, and longitudinal data before and after that intervention are available for that treated unit and multiple others without that intervention. As such, and in more or less sophisticated versions, it has been used to investigate the effects of mandatory face masks in Germany (Mitze et al., 2020), and of lockdown measures in the United States (Friedson et al., 2021), in Chile (Herrera and Godoy-Faúndez, 2021), as well as in Wuhan (Yang, 2021), Wenzhou and Shanghai in China (Tian et al., 2021). The method can also be applied to multiple treated units, comparing the average trajectory followed by those that have adopted a certain policy, or experienced certain events, compared to their average synthetic counterfactual (Cavallo et al., 2013).

This latter possibility is used in our within-country comparison of Italian provinces to further investigate the effect of school closures and openings. In Italy, legislative and administrative powers on several issues related to health and education are decentralized to the regional level, though some (e.g. the definition of the school calendar) may be further delegated to an even more local level. The pandemic has fostered a sudden, and often disordered, recentralization of many decisions. For instance, as stated in Section 2, the first school closures were decided in February 2020 by some regional administrations, but then a national decree extended that decision to the entire country. The re-openings of schools for the new 2020-21 year were again decided locally, although mixed teaching methods were defined centrally for the universities, and after a few weeks a new national decree again ordered distance teaching for all high schools, and restricted the options for primary and lower-secondary schools according to the severity of the local epidemiological situation (the so-called "colour-coding system"). Thereafter, the colour differentiation was extended to upper secondary schools, fixing different percentages of class filling according to the severity of the health conditions, and eventually all schools and universities were ordered to close in the so-called "red regions".

The overall process was often confused, if not contradictory, and offered the chance for opportunistic politicized behaviour in a centre-periphery game that certainly did not help combat the pandemic (Capano, 2020). Paradoxically, if fine-tuning the strictness of the regulations according to the epidemiological conditions seems a sensible decision, it was a uniform national decision that somehow delayed the arrival of the second wave in Italy (Coppola and Ryan, 2020; Manica et al., 2021).

Whatever the best overall strategy should have been, we profit from the different timings of some decisions regarding the opening of schools after the summer break to test again the contested effects of school attendance on the spread of the virus. Firstly, different Italian regions decided to start the new school year on different dates. Students living in the autonomous province of Bolzano were those who returned to school earliest, on September 7; several regions opted to open the schools on September 14; and another group started almost two weeks later.[6] If the return to school, with all its consequent mobility issues and multiple contacts, had an effect on the spread of the virus, we should observe an earlier increasing trend of COVID infections in the provinces that started earlier compared to those that started later.

This approach is similar to the one followed by Isphording et al. (2020), who exploited the large differences in the timings of the return to school among German *Länder* – from the beginning of August to the middle of September. They compared county cases two weeks before and three after the end of the vacation period, using as counterfactuals for the *Länder* whose students were going back in class, those who were still on vacation together with those who had re-opened at an earlier date. Isphording and colleagues counterintuitively found that "the end of summer breaks is associated with a distinct decrease [and not increase] in the number of SARS-CoV-2 cases" (14). The new organization of the school environment is prudently cited amongst the reasons explaining their unexpected findings, although they also admit that the results are mostly driven by "states with early summer breaks". This means that the negative effect of the return to school was mostly determined by a trend experienced in August, well before the onset of the second wave in the country.

[6] Having weekly data, we approximate the exact dates to the closest Monday in order to reflect the actual presence of students in class. In some regions, the original dates were postponed by the governors a few weeks before the start of classes, whilst in Sicily the regional government first allowed the deferral from 14 to 24 September only for schools used as polling stations for a national referendum, and then extended the possibility to all kinds of schools, producing, as a result, a *de facto* general rescheduling.

The different timings of school openings have been used in the Italian setting also by Gandini et al. (2021), who employed as dependent variable the effective reproduction number R_t at the regional level. Amongst other analyses, the authors provide a series of longitudinal graphs paralleling pairs of supposedly comparable cases – Trento and Bolzano, Abruzzo and Marche, Veneto and Apulia, Calabria and Sicily – whose main difference was thought to be exactly the diverse return to school by students. After a mostly visual inspection of seven-day moving averages, Gandini et al. (2021: 7) conclude that they “did not find an unequivocally constant delay between school opening and R_t rise”.

We believe that their identification strategy is questionable: Sicily *de facto* opened its schools on the same day as Calabria; Bolzano and Trento had large pre-treatment differences in their trajectories, so that they are not an ideal comparison; Veneto and Apulia are an odd pair of regions to compare; while Marche and Abruzzo, which actually show similar trends towards the end of September, have large confidence intervals before that period that make it difficult to evaluate the soundness of the comparison. We propose to improve Gandini et al.’s identification strategy by extending it to all Italian provinces, and with a more systematic definition of the counterfactuals.

Our outcome of interest is again the incidence of new weekly cases, and we define as treatment, the (relatively) early re-opening of schools. We use a series of COVID-related variables to construct an appropriate synthetic counterfactual through a weighted linear combination of the provinces that opened schools later, located in the regions of Abruzzo, Apulia, Basilicata, Calabria, Campania, Sicily and Sardinia. We preliminarily tested a difference-in-difference model and verified a systematic average positive effect on the treated group (ATET), meaning that provinces that opened their schools earlier had on average a larger number of COVID cases (results in the online appendix). However, this kind of model relies on a set of assumptions (Cunningham, 2021) that the evident North-South divide in the location of the provinces belonging to the two groups suggests may have been violated. The identification of a synthetic counterfactual as a linear combination of non-treated cases is defensible only if it can demonstrate an ability to sufficiently well approximate our treated cases in the pre-treatment period, i. e. before the opening of the schools.

After a series of multivariate panel tests, available in the online appendix, we decided to use two groups of variables as predictors associated with the outcome for our identification strategy. The first one has to do with other indices of the epidemiological situation: the incidence of overall and active cases, the number of COVID tests performed per thousand persons, the positivity rate, and the reproduction number R_t. The second one is demographic: the population of the province, its density per square kilometre, to which we added also the share of students' population (from preschool to high school) to reflect its potential multiplicative factor. Using these predictive variables, and merging with the appropriate weights the provinces in which children went back to school later, the synthetic control method is able to simulate the average trend of infections in a counterfactual province as similar as possible to those that actually opened schools earlier, except for the fact that it did not. In Figure 2 we compare the actual average trajectory with that of this synthetic counterfactual.

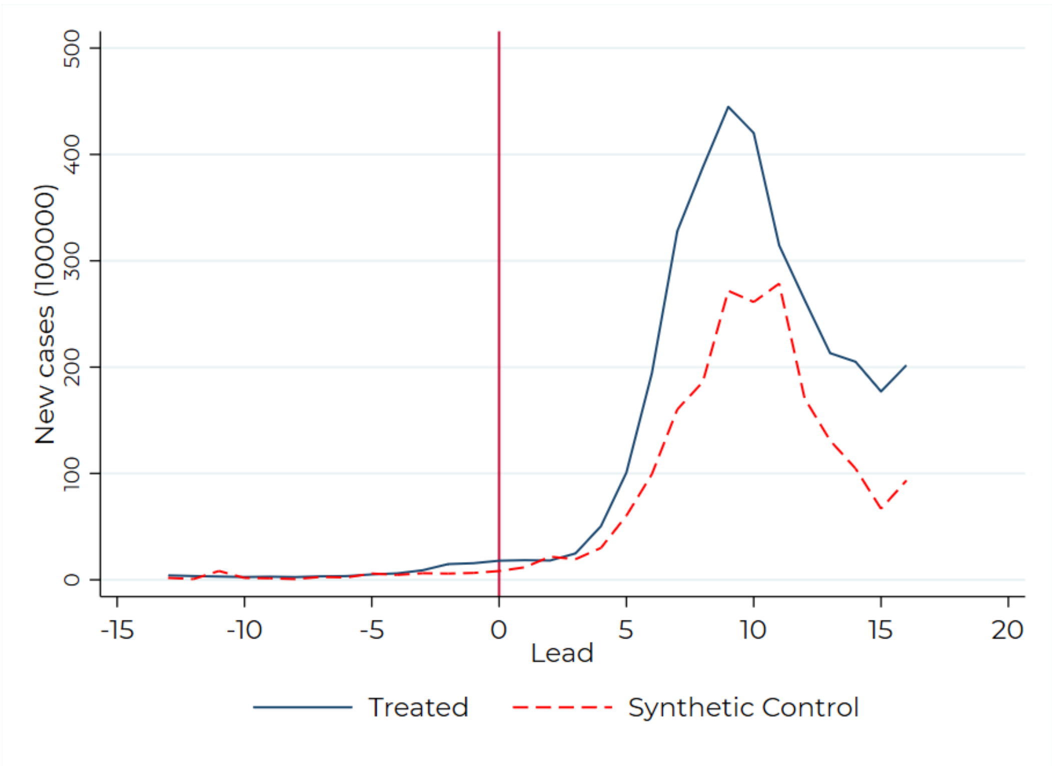


Figure 2 Incidence of new cases for school early opening provinces, and their counterfactual

We start the comparison in early June, and in the plot it is possible to see how close the two trends are until the week before the school openings, which are marked with the solid vertical line in the model. Our model has in fact a low pre-treatment Root Mean Squared Prediction Error (RMSPE=4.77), the statistic that evaluates the soundness of the counterfactual before the policy or the event. The trajectory of the provinces with schools that opened earlier (solid blue line), started to diverge from the synthetic counterfactual (dashed red line) a couple of weeks after the early openings, anticipating the surge in new cases that involved all provinces soon after. Thereafter, the gap between the two lines increases so much that the post-treatment RMSPE is more than twenty times larger than before, this being usually considered a useful indicator of the non-random character of the deviation (Abadie et al., 2010).

Using a sample of 1000000 placebo tests, it is possible to assess the statistical significance of the gap between the trajectory of the provinces that opened schools in early September and that of their synthetic counterfactual: the difference between the two trends became systematic three weeks after the first openings. The divergence remained statistically significant until the eleventh week, i.e. fourteen days after the implementation of the new government decree of 3 November 2020, which closed all high schools in the country, and the introduction of which contributed to the sharp common decrease in contagions clearly seen in the graph.

This analysis mirrors the previous one of school closures. Exploiting the quasi-experimental setting produced by the diverse timing of children's return to school, and manufacturing a synthetic trajectory that emulates the average trend of an early-opening province that counterfactually did not open its classrooms, we were able to assess the multiplicative effect on infections directly and indirectly exercised by class attendance.

To complement this analysis, we replicate the same approach selecting as the only "treated" unit the province of Bolzano, which was the only one that opened its schools on September 7, that is, at the beginning of the 37th week, while we include in the donor pool only those provinces that opened their schools more than two weeks later.[7] Running the synthetic control method using the same

[7] We lately found a work by Alfano et al. (2020), who similarly applied the synthetic research method on the early opening of schools in Bolzano. However, they defined the donor pool differently, focused on a shorter period, used daily

predictors included in the previous model, we obtain the actual and counterfactual trajectory of the infections for Bolzano plotted in the left panel of Figure 3. The pre-opening prediction error is still very low (RMSPE=4.58), while the post-treatment RMSPE is more than 60 times larger, confirming the visual departure of the actual Bolzano from its synthetic counterfactual.

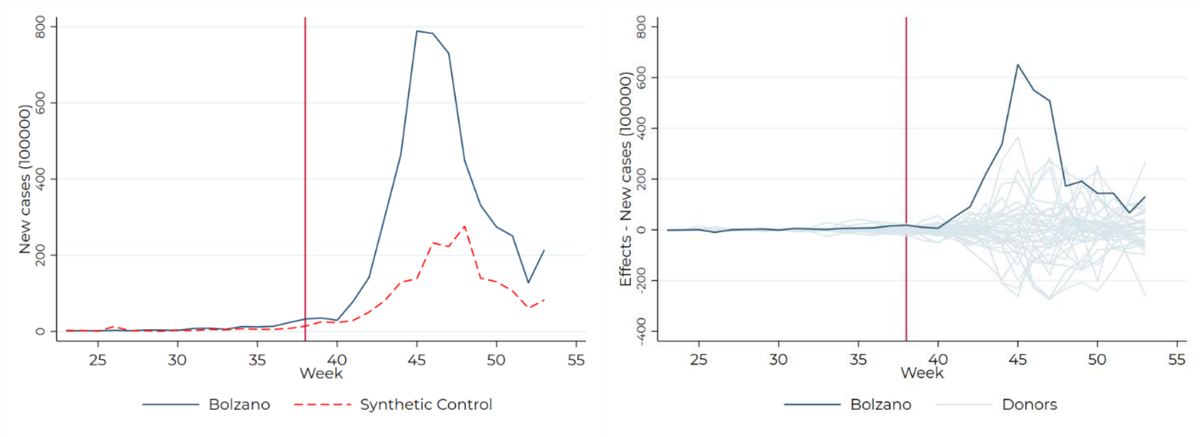


Figure 3 The incidence of new COVID cases in Bolzano and its synthetic counterfactual (left panel), and the effect of the early opening, together with the placebo effects (right panel)

The two trajectories start to diverge consistently from the 40th week, that is, three weeks after the opening of schools in the Bolzano province. By running a complete set of placebo tests, whose effects are plotted in the right panel of Figure 3, it is possible to test the statistical significance of that separation, which becomes systematic only after the fourth week following the end of the summer break. Due to the rapidly deteriorating situation, the national government later imposed a common national framework with a long list of measures to curb the spread of the virus, including curfews, school and business closures, mobility restrictions, etc. The implementation of those measures contributed to the sharp decline in new COVID cases after the 45th week, and also favoured the

instead of weekly data, and opted for a different set of predictive covariates and a different outcome – log total cases since August 1. Interestingly, they obtained results that are coherent with ours.

convergence of the incidence of infections between Bolzano and the provinces that opened their schools much later.

The synthetic research method uses a linear combination of observations from a donor pool to build a counterfactual that minimizes the error of prediction before the treatment. In our case, the best synthetic Bolzano was a weighted mix among the provinces of Pescara, Cosenza and Reggio Calabria. Recently, Cerulli (2020: 845) has suggested that “relaxing the linearity assumption by providing a nonparametric estimation of the weights may somehow improve their estimation”, and also better distribute them across a larger number of non-treated units. As a further robustness test, we ran a non-parametric synthetic control analysis whose results are reported in the online appendix. Using this approach, it was possible to slightly improve the pre-treatment prediction error (RMSPE=4.40) while distributing the weights across a much larger number of provinces, which makes the counterfactual less dependent on some local trend. However, the alternative specification does not modify, but instead reinforces, our conclusions regarding the systematic increase of infections in correspondence with the earlier opening of schools.

5 Conclusion: what can we learn?

This article is a methodological exercise with some substantive results. While summarizing the latter, we want to reflect on the advantages of the former and on its far-reaching significance.

We first used matching to test the effect of school closures in a comparative setting. Using a dichotomous measure constructed on the specific school subcomponent of the stringency index covering 31 European countries for the first six months of 2021, we showed how closure measures produced the effects expected by policy-makers even when controlling for other covariates, including different containment and closure policies. We then used the synthetic research method with Italian sub-national data. The results confirm the plausible expectation that school attendance, with all its indirect spillover effects, is a systematic channel for the spread of the virus.

Many studies point to school opening/closing, as one of those non-pharmaceutical interventions (apart from lockdowns) with the largest direct and indirect effects (Li et al., 2021; Liu et al., 2021). There has been much debate on this issue worldwide, with already systematic reviews and meta-analyses trying to summarize the findings of multiple empirical research, and making explicit the economic, psychological and educational downside of this kind of policy (Busa et al., 2021; Krishnaratne et al., 2020). Several scholars in this research field explicitly design, look for, model, match or reconstruct some counterfactual against which to compare the cases of interest (e. g. Auger et al., 2020; Cunha et al., 2021; Gordon et al., 2021; Rauscher and Burns, 2021). Notably, with a few exceptions, these comparisons are not performed by political scientists and, apart from some economists, not even by the more comprehensive category of social scientists.

It is probably true that there is already a large and interesting agenda for policy scholars (Dunlop et al., 2020; Weible et al., 2020), and many social scientists are more interested in looking at the opposite side of the COVID crisis, at the social and political consequences of the pandemic (Vezzoni et al., 2020), at the institutional factors triggering certain types of policy responses (Bandelow et al., 2021; Zhao et al., 2020), and at their variegated level of support (Altiparmakis et al., 2021; Bol et al., 2021; Jørgensen et al., 2021).

We understand this intellectual division of labour. However, for several reasons, we believe that policy scholars, and social scientists at large, should not refrain from assessing the outcome of (COVID) policies, and doing so in a manner which is methodologically informed. Firstly, because we should be concerned to check whether a problem has been solved, moderated, left untouched, or even exacerbated by the government’s responses, not least because evaluation remains one of the traditional phases of the policy cycle. Secondly, because ethical and normative issues are also connected to that assessment (Silverman et al., 2020), making that final stage of the cycle intrinsically political (Bovens et al., 2006). Thirdly, because this is not a time for separate working, since “facing up to these challenges will be complex, requiring integrated and interconnected responses that draw on diverse expertise, a range of actors and various disciplines”; “it is at the intersections, not in silos, that we are likely to move forward intellectually and practically” (O'Flynn, 2020). This is not necessarily a plea for interdisciplinarity, but it is certainly a call for mutual methodological understanding. Finally, in this latter regard, political (and policy) science was born comparative, and the issue of what is appropriate to compare has always been one of its methodological concerns (Sartori, 1991).

When it comes to policies, understanding the properties that makes something comparable may even be more complicated than in politics. We probably agree that judging the effectiveness of a country’s anti-COVID strategy by comparing a nation in the middle of Europe to some isolated island at its antipodes – say New Zealand and Japan – is not the best research approach. But neither is simply running models on any data available. Matching is a way to identify bottom-up sensible comparisons, especially when the researcher is confronted with new and complex issues that need to be tackled, but that are outside his/her comfort zone. And there is no doubt that COVID policies are one of those cases. Ultimately, it is not even an issue of the Qual/Quant divide (Plumper et al., 2019), since “what makes a statistical treatment theoretically significant has nothing to do with statistics” (Sartori, 1970: 1037). Counterfactuals are the essence of causal attribution (Brady, 2008; Paul, 2009), and the search for the most similar world should matter for any kind of comparison.

ONLINE APPENDIX

Datasets and code are available on Harvard Dataverse at xxx.yyy

Contents

1 Codebook 26

1.1 Comparative data 26

1.2 Italian subnational data 27

2 Descriptive statistics 28

3 Some preliminary models 31

4 Difference in Difference 32

5 Non-parametric synthetic control 34

1 Codebook

1.1 Comparative data

Variable	Measurement	Source	Link
Incidence of new cases	Weekly number of new COVID-19 cases per million persons	Our World in Data on COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)	https://github.com/owid/covid-19-data
School closing scale	Weekly average of the School closing 0-3 scale (weighted for possible subnational differentiation)	Oxford COVID-19 Government Response Tracker (OxCGRT)	https://github.com/OxCGRT/covid-policy-tracker
School closing index	Weekly average of the index measuring the 0-100 percentage of strictness of regulation relative to its theoretical maximum (0-100)	Oxford COVID-19 Government Response Tracker (OxCGRT)	https://github.com/OxCGRT/covid-policy-tracker
School closing 75 dummy	Dummy variable for the school closing policy (=1 if the school closing index is higher than 75%, and =0 otherwise)	Oxford COVID-19 Government Response Tracker (OxCGRT)	https://github.com/OxCGRT/covid-policy-tracker
Other containment and closure policies	Weekly average of the index computed on the other 7 containment and closure policies (weighted for possible subnational differentiation)	Oxford COVID-19 Government Response Tracker (OxCGRT)	https://github.com/OxCGRT/covid-policy-tracker
Other containment 75 policy dummy	Dummy variable for the other containment and closure policies (=1 if the other containment index is higher than 75%, and =0 otherwise)	Oxford COVID-19 Government Response Tracker (OxCGRT)	https://github.com/OxCGRT/covid-policy-tracker
Fully vaccinated	Cumulated percentage of fully vaccinated population	Data on COVID-19 by Our World in Data	https://github.com/owid/covid-19-data
Density	Population per squared kilometre	World Bank World Development Indicators	https://github.com/owid/covid-19-data

1.2 Italian subnational data

Variable	Measurement	Source	Link
Incidence of new/total cases	Weekly number of new COVID-19 cases per 100000 persons	Dipartimento Protezione Civile - Dati COVID-19 Italia	https://github.com/pcm-dpc/COVID-19
Incidence active cases	Weekly average of the active COVID-19 cases per 100000 persons	Dipartimento Protezione Civile - Dati COVID-19 Italia	https://github.com/pcm-dpc/COVID-19
Tests	Weekly number of COVID-19 tests per 1000 persons	Dipartimento Protezione Civile - Dati COVID-19 Italia	https://github.com/pcm-dpc/COVID-19
Positivity rate	Weekly ratio between new cases and tests	Dipartimento Protezione Civile - Dati COVID-19 Italia	https://github.com/pcm-dpc/COVID-19
Reproduction rate R_t	Regional reproduction rate	Sole 24 Ore	https://lab24.ilsole24ore.com/coronavirus/#
Population	Provincial and regional population	Dipartimento Protezione Civile - Dati COVID-19 Italia	https://github.com/pcm-dpc/COVID-19
Student population	Percentage of student population (from pre-school to high-school)	ISTAT	http://dati.istat.it/ Index.aspx?DataSetCode=DCIS_SCUOLE#
Density	Population per squared kilometre	Dati ISTAT 1 Gennaio 2021	https://www.tuttitalia.it/

2 Descriptive statistics

Figure A.1 Incidence of new and total cases per million population – 31 European countries (January-June 2021)

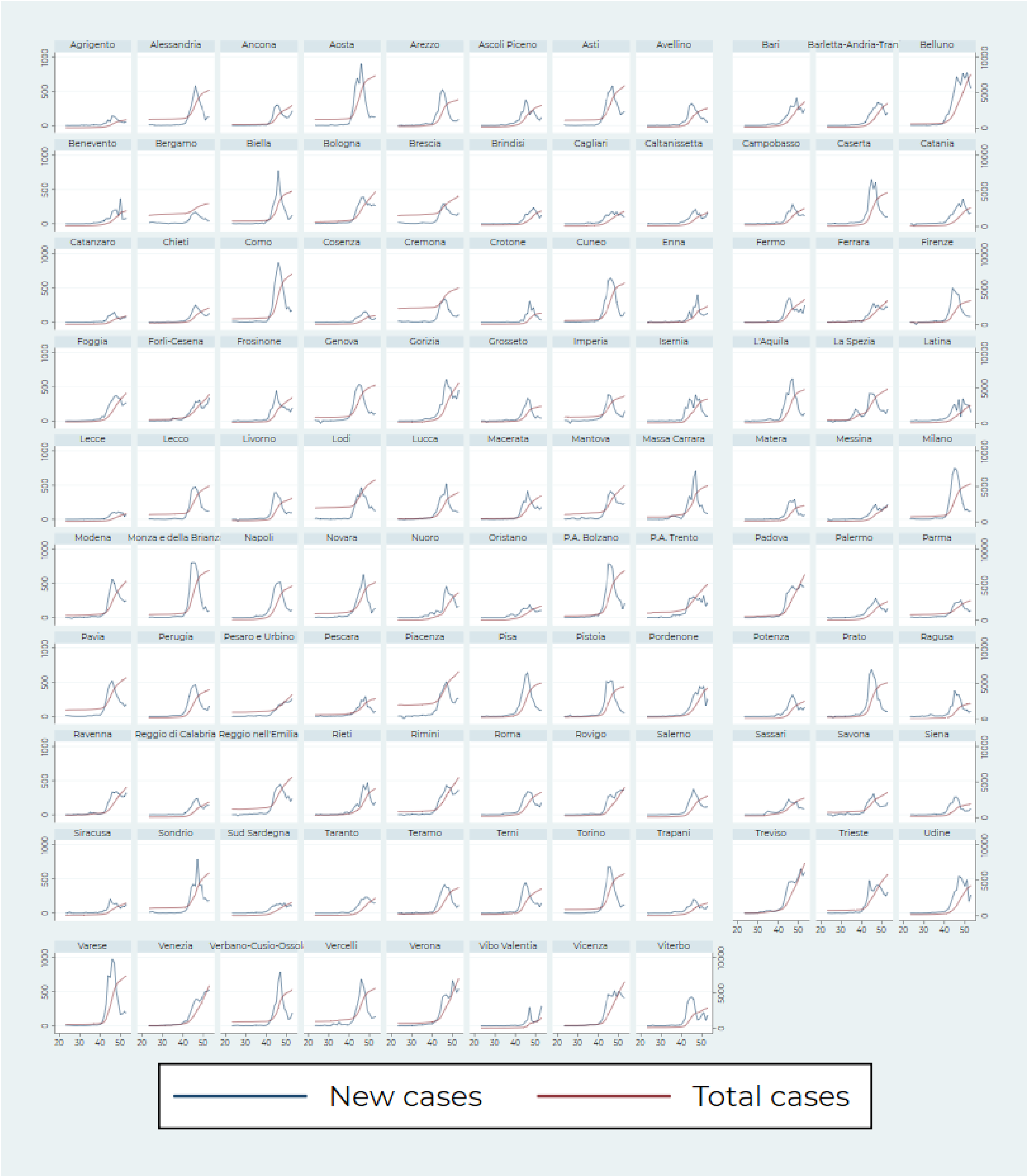


Figure A.2 Incidence of new and total cases per 100000 population – 31 European countries (June-December 2020)

Table A.1 Descriptive statistics of the major variables used in the comparative analysis

Variable	Mean	Std. dev.	Min	Max
New cases (1000000)	1586.21	1497.52	11.72	9260.41
Total cases (1000000)	61552.34	28313.84	6964.80	155653.50
Total Fully vaccinated (avg pct week)	9.41	10.94	0.00	70.18
Population	1.71e+07	2.28e+07	341250.00	8.38e+07
Density	171.91	257.70	3.40	1454.04
Tests (weekly 1000)	54.54	88.12	0.00	592.16
School index (avg week)	58.14	24.16	0.00	100.00
Avg other constraints indices (avg week)	59.08	15.45	0.00	88.10

Table A.2 Descriptive statistics of the major variables used in the within-Italy analysis

Variable	Mean	Std. dev.	Min	Max
New cases (100000)	96.12	148.11	0.00	973.15
Total cases (100000)	1090.28	1216.82	27.62	7586.21
Active cases (100000)	371.40	516.42	0.37	2177.94
Population	553809.00	602707.80	81918.00	4227588.00
Density	265.31	375.95	36.00	2560.00
Student population	14.04	1.17	11.54	17.58
Tests (weekly 1000)	12.44	8.71	0.00	45.96
Positivity rate	0.052	0.06	0.00	0.23
Reproduction rate R_t	.9510582	.3696576	0.00	2.37

3 Some preliminary models

Table A.4 Panel regressions on the weekly incidence of new cases in Italy (June-December 2020)

	(1)	(2)	(3)
Lag Total cases (100000)	0.005		0.003
	(0.007)		(0.007)
Lag Active cases (100000)	0.016		0.018
	(0.015)		(0.015)
Tests (1000)	5.835 ***		5.778 ***
	(0.547)		(0.538)
Positivity rate	1457.391 ***		1466.838 ***
	(91.390)		(89.187)
Reproduction rate R_t	-12.996		-13.997
	(7.476)		(7.450)
Log Population		6.134 **	-4.214 *
		(2.261)	(1.657)
Density		0.031 **	0.023 **
		(0.012)	(0.008)
Student population (100)		-9.404 ***	-0.470
		(2.243)	(1.166)
Constant	-50.738 ***	140.635 **	6.244
	(10.653)	(48.631)	(25.0533)
Provinces	107	107	107
Observations	3210	3317	3210

Note: Panel corrected standard errors in parentheses.
*p < 0.05 **p < 0.01 ***p < 0.001

4 Difference in Difference

We fitted two difference-in-difference models for cross-sectional data with panel and time fixed effects, to explain the impact of school openings on the incidence of new COVID cases. The first one uses only the covariates of interests, whereas the second one introduces also the set of control variables used also in the article: the number of total and active cases per 100000 inhabitants, the number of tests per thousand persons, the positivity rate and the reproduction rate R_t.

Table A.3 Average effects of early school-openings in Italy (June-December 2020)

Difference-in-differences regression   Number of obs = **3,317**
Data type: Longitudinal

(Std. err. adjusted for **107** clusters in **province**)

newnew	Coefficient	Robust std. err.	t	P>|t|	[95% conf. interval]	
ATET						
school (1 vs 0)	**83.08866**	**11.0505**	**7.52**	**0.000**	**61.17997**	**104.9973**

Note: ATET estimate adjusted for panel effects and time effects.

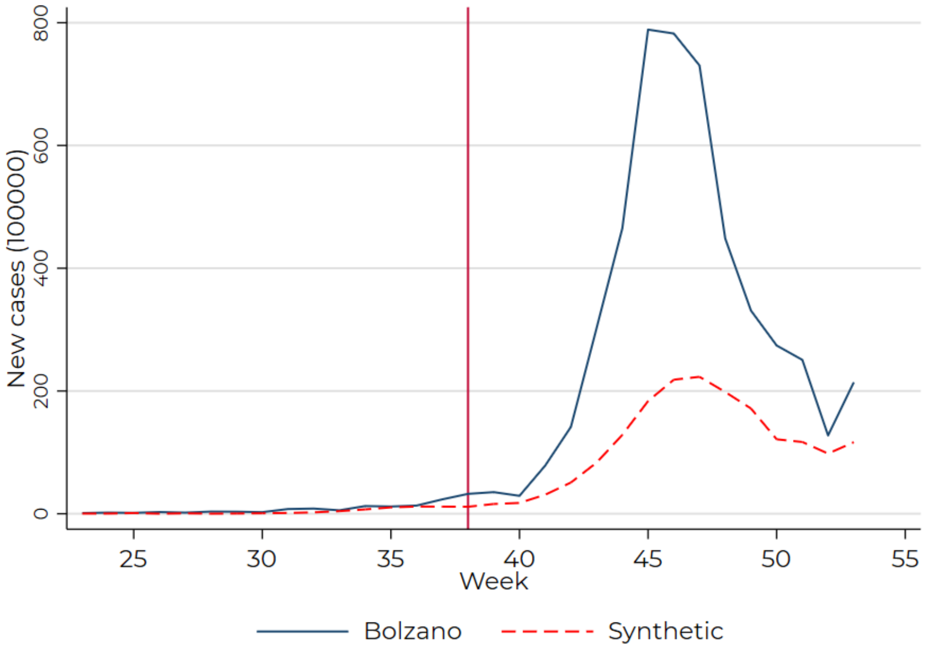
```
Difference-in-differences regression                    Number of obs = 3,317
Data type: Longitudinal

                               (Std. err. adjusted for 107 clusters in province)
```

newnew	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
school						
(1 vs 0)	**16.71952**	**6.884668**	**2.43**	**0.017**	**3.069993**	**30.36904**

Note: ATET estimate adjusted for covariates, panel effects, and time effects.

In both cases the average treatment effect for those provinces that opened the schools earlier is positive and highly significant. Adjusting the estimate for all control variables is almost 17 new weekly cases per 100000 inhabitants.

5 Non-parametric synthetic control



6 References

Abadie, A., Diamond, A. and 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.

Abadie, A., Diamond, A. and Hainmueller, J. (2015) 'Comparative Politics and the Synthetic Control Method', *American Journal of Political Science* 59(2): 495-510.

Abadie, A. and Gardeazabal, J. (2003) 'The Economic Costs of Conflict: A Case Study of the Basque Country', *The American Economic Review* 93(1): 113-132.

Alfano, V., Ercolano, S. and Cicatiello, L. (2020) 'A Synthetic Control Method Analysis of Schools Opening and Covid-19 Outbreak in Italy', *CESifo Working Papers.* Munich: Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute.

Altiparmakis, A., Bojar, A., Brouard, S., et al. (2021) 'Pandemic politics: policy evaluations of government responses to COVID-19', *West European Politics*. DOI: 10.1080/01402382.2021.1930754. 1-21.

Andersson, S. and Aylott, N. (2020) 'Sweden and Coronavirus: Unexceptional Exceptionalism', *Social Sciences* 9(12).

Angrist, J.D. and Pischke, J.r.-S. (2009) *Mostly harmless econometrics : an empiricist's companion.* Princeton: Princeton University Press.

Auger, K.A., Shah, S.S., Richardson, T., et al. (2020) 'Association Between Statewide School Closure and COVID-19 Incidence and Mortality in the US', *Journal of the American Medical Association* 324(9): 859-870.

Bandelow, N.C., Hassenteufel, P. and Hornung, J. (2021) 'Patterns of Democracy Matter in the COVID-19 Crisis', *International Review of Public Policy* 3(1).

Beck, N. and Katz, J.N. (1995) 'What to do (and not to do) with Time-Series Cross-Section Data', *The American Political Science Review* 89(3): 634-647.

Blackwell, M., Iacus, S., King, G., et al. (2009) 'cem: Coarsened exact matching in Stata', *The Stata Journal* 9(4): 524-546.

Bol, D., Giani, M., Blais, A., et al. (2021) 'The effect of COVID-19 lockdowns on political support: Some good news for democracy?', *European Journal of Political Research* 60(2): 497-505.

Bovens, M., Hart, P.t. and Kuipers, S. (2006) The politics of policy evaluation. In: Moran M, Rein M and Goodin RE (eds) *The Oxford handbook of public policy*. Oxford: Oxford University Press, pp.319-335.

Brady, H.E. (2008) Causation and Explanation in Social Science. In: Box-Steffensmeier JMB, Henry E. and Collier D (eds) *Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Brandal, L.T., Ofitserova, T.S., Meijerink, H., et al. (2021) 'Minimal transmission of SARS-CoV-2 from paediatric COVID-19 cases in primary schools, Norway, August to November 2020', *Euro Surveill* 26(1).

Busa, F., Bardanzellu, F., Pintus, M.C., et al. (2021) 'COVID-19 and School: To Open or Not to Open, That Is the Question. The First Review on Current Knowledge', *Pediatric Reports* 13(2): 257-278.

Capano, G. (2020) 'Policy design and state capacity in the COVID-19 emergency in Italy: if you are not prepared for the (un)expected, you can be only what you already are', *Policy and Society* 39(3): 326-344.

Capano, G., Howlett, M., Jarvis, D.S.L., et al. (2020) 'Mobilizing Policy (In)Capacity to Fight COVID-19: Understanding Variations in State Responses', *Policy and Society* 39(3): 285-308.

Cavallo, E., Galiani, S., Noy, I., et al. (2013) 'Catastrophic Natural Disasters and Economic Growth', *Review of Economics and Statistics* 95(5): 1549-1561.

Ceron, M., Palermo, C.M. and Grechi, D. (2021) 'Covid-19 Response Models and Divergences Within the EU: A Health Dis-Union', *Statistics, Politics and Policy* 0(0).

Cerulli, G. (2020) 'Nonparametric synthetic control using the npsynth command', *The Stata Journal: Promoting communications on statistics and Stata* 20(4): 844-865.

Coppola, D.P. and Ryan, B.J. (2020) 'Color-coding for COVID-19 Management: An Idea Whose Time has Come', *Journal of Homeland Security and Emergency Management* 17(2).

Costalli, S. and Negri, F. (2021) 'Looking for twins: how to build better counterfactuals with matching', *Italian Political Science Review/Rivista Italiana di Scienza Politica* 51(2): 215-230.

Cunha, M., Domingos, A., Rocha, V., et al. (2021) 'How many could have been saved? Effects of social distancing on COVID-19', *Revista De Administracao Publica* 55(1): 12-26.

Cunningham, S. (2021) *Causal Inference. The mixtape.* New Haven: Yale University Press.

Davies, N.G., Kucharski, A.J., Eggo, R.M., et al. (2020) 'Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study', *The Lancet Public Health* 5(7): e375-e385.

Dunlop, C.A., Ongaro, E. and Baker, K. (2020) 'Researching COVID-19: A research agenda for public policy and administration scholars', *Public Policy and Administration* 35(4): 365-383.

Folkhälsomyndigheten (2020) Covid-19 in schoolchildren. A comparison between Finland and Sweden.

Friedson, A.I., McNichols, D., Sabia, J.J., et al. (2021) 'Public Policy Implications of the COVID-19 Pandemic', *Journal of Policy Analysis and Management* 40(1): 258-+.

Gandini, S., Rainisio, M., Iannuzzo, M.L., et al. (2021) 'A cross-sectional and prospective cohort study of the role of schools in the SARS-CoV-2 second wave in Italy', *The Lancet Regional Health - Europe* 5.

Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel Hierarchical Models* Cambridge: Cambridge University Press.

Gordon, D.V., Grafton, R.Q. and Steinshamn, S.I. (2021) 'Cross-country effects and policy responses to COVID-19 in 2020: The Nordic countries', *Economic Analysis and Policy* 71: 198-210.

Hale, T., Angrist, N., Goldszmidt, R., et al. (2021) 'A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)', *Nature Human Behaviour* 5(4): 529-538.

Herrera, M. and Godoy-Faúndez, A. (2021) 'Exploring the Roles of Local Mobility Patterns, Socioeconomic Conditions, and Lockdown Policies in Shaping the Patterns of COVID-19 Spread', *Future Internet* 13(5).

Iacus, S.M., King, G. and Porro, G. (2011) 'Multivariate Matching Methods That Are Monotonic Imbalance Bounding', *Journal of the American Statistical Association* 106(493): 345-361.

Iacus, S.M., King, G. and Porro, G. (2012) 'Causal Inference without Balance Checking: Coarsened Exact Matching', *Political Analysis* 20(1): 1-24.

Isphording, I.E., Lipfert, M. and Pestel, N. (2020) 'School Re-Openings after Summer Breaks in Germany Did Not Increase SARS-CoV-2 Cases', *Discussion paper.* Bonn: IZA Institute for the Study of Labor.

Iwata, K., Doi, A. and Miyakoshi, C. (2020) 'Was school closure effective in mitigating coronavirus disease 2019 (COVID-19)? Time series analysis using Bayesian inference', *International Journal of Infectious Diseases* 99: 57-61.

Jørgensen, F., Bor, A., Lindholt, M.F., et al. (2021) 'Public support for government responses against COVID-19: assessing levels and predictors in eight Western democracies during 2020', *West European Politics*. DOI: 10.1080/01402382.2021.1925821. 1-30.

King, G., Keohane, R.O. and Verba, S. (1994) *Designing Social Inquiry. Scientific Inference in Qualitative Research.* Princeton: Princeton University Press.

King, G., Lucas, C. and Nielsen, R.A. (2017) 'The Balance-Sample Size Frontier in Matching Methods for Causal Inference', *American Journal of Political Science* 61(2): 473-489.

Krishnaratne, S., Pfadenhauer, L.M., Coenen, M., et al. (2020) 'Measures implemented in the school setting to contain the COVID-19 pandemic: a scoping review', *Cochrane Database Syst Rev* 12: CD013812.

Li, Y., Campbell, H., Kulkarni, D., et al. (2021) 'The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries', *The Lancet Infectious Diseases* 21(2): 193-202.

Liu, Y., Morgenstern, C., Kelly, J., et al. (2021) 'The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories', *BMC Medicine* 19(1).

Ludvigsson, J.F. (2020) 'Children are unlikely to be the main drivers of the COVID-19 pandemic - A systematic review', *Acta Paediatrica* 109(8): 1525-1530.

Manica, M., Guzzetta, G., Riccardo, F., et al. (2021) 'Impact of tiered restrictions on human activities and the epidemiology of the second wave of COVID-19 in Italy', *Nat Commun* 12(1): 4570.

McConnell, A. and Stark, A. (2021) 'Understanding policy responses to COVID-19: the stars haven't fallen from the sky for scholars of public policy', *Journal of European Public Policy* 28(8): 1115-1130.

Mill, J.S. (1843) *A System of Logic, Ratiocinative and Inductive, being a Connected View of the Princip0les of Evidence, and the Methods of Scientific Investigation.* London: John W. Parker.

Mitze, T., Kosfeld, R., Rode, J., et al. (2020) 'Face Masks Considerably Reduce Covid-19 Cases in Germany. A Synthetic Control Method Approach', *CESifo Working Papers.* Munich: Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute.

Negri, F. (2022) Correlation is not causation, yet... Matching and weighting for better counterfactuals. In: Damonte A and Negri F (eds) *Causality in Policy Studies. A Pluralist Toolbox*. Cham: Springer Nature.

Nielsen, R.A. (2020) Statistical Matching with Time-Series Cross-Sectional Data: Magic, Malfeasance, or Something in between? . In: Curini L and Franzese RJ (eds) *The SAGE Handbook of Research Methods in Political Science and International Relations*. London: Sage, pp.805-821.

O'Flynn, J. (2020) 'Confronting the big challenges of our time: making a difference during and after COVID-19', *Public Management Review*. DOI: 10.1080/14719037.2020.1820273.

Park, Y.J., Choe, Y.J., Park, O., et al. (2020) 'Contact Tracing during Coronavirus Disease Outbreak, South Korea, 2020', *Emerging Infectious Diseases* 26(10): 2465-2468.

Pashakhanlou, A.H. (2021) 'Sweden's coronavirus strategy: The Public Health Agency and the sites of controversy', *World Medical Health Policy*. Epub ahead of print 2021/07/07. DOI: 10.1002/wmh3.449.

Paul, L.A. (2009) Counterfactual Theories. In: Beebee H, Hitchcock C and Menzies P (eds) *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Pierre, J. (2020) 'Nudges against pandemics: Sweden's COVID-19 containment strategy in perspective', *Policy and Society* 39(3): 478-493.

Plumper, T., Troeger, V.E. and Neumayer, E. (2019) 'Case selection and causal inferences in qualitative comparative research', *PLoS One* 14(7): e0219727.

Rauscher, E. and Burns, A. (2021) 'Unequal Opportunity Spreaders: Higher COVID-19 Deaths with Later School Closure in the United States', *Sociological Perspectives*. DOI: 10.1177/07311214211005486.

Sartori, G. (1970) 'Concept Misformation in Comparative Politics', *The American Political Science Review* 64(4): 1033-1053.

Sartori, G. (1991) 'Comparing and Miscomparing', *Journal of Theoretical Politics* 3(3): 243-257.

Silverman, M., Sibbald, R. and Stranges, S. (2020) 'Ethics of COVID-19-related school closures', *Canadian Journal of Public Health* 111(4): 462-465.

Stage, H.B., Shingleton, J., Ghosh, S., et al. (2021) 'Shut and re-open: the role of schools in the spread of COVID-19 in Europe', *Philosophical Transactions of the Royal Society B-Biological Sciences* 376.

Stone, D. (2020) *Counting. How We Use Numbers to Decide What Matters.* New York: Liveright.

Tian, T., Tan, J., Luo, W., et al. (2021) 'The Effects of Stringent and Mild Interventions for Coronavirus Pandemic', *Journal of the American Statistical Association* 116(534): 481-491.

Toshkov, D., Carroll, B. and Yesilkagit, K. (2021) 'Government capacity, societal trust or party preferences: what accounts for the variety of national policy responses to the COVID-19 pandemic in Europe?', *Journal of European Public Policy*. DOI: 10.1080/13501763.2021.1928270. 1-20.

Tosi, D. and Campi, A.S. (2021) 'How Schools Affected the COVID-19 Pandemic in Italy: Data Analysis for Lombardy Region, Campania Region, and Emilia Region', *Future Internet* 13(109): 1-12.

Vezzoni, C., Ladini, R., Molteni, F., et al. (2020) 'Investigating the social, economic and political consequences of Covid-19: A rolling cross-section approach', *Survey Research Methods* 14(2).

Weible, C.M., Nohrstedt, D., Cairney, P., et al. (2020) 'COVID-19 and the policy sciences: initial reactions and perspectives', *Policy Sciences* 53(2): 225-241.

Wenander, H. (2021) 'Sweden: Non-binding Rules against the Pandemic - Formalism, Pragmatism and Some Legal Realism', *European Journal of Risk Regulation* 12(1): 127-142.

Yang, X. (2021) 'Does city lockdown prevent the spread of COVID-19? New evidence from the synthetic control method', *Global Health Research and Policy* 6(1).

Zhao, T., Tahir, N., Sezgin, M.H., et al. (2020) 'Institutional Origins of Protective COVID-19 Public Health Policy Responses: Informational and Authority Redundancies and Policy Stringency', *Journal of Political Institutions and Political Economy* 1(4): 585-613.