Challenges in determining causality: an ongoing critique of Bendavid et al's "Assessing mandatory stay-at-home and business closure effects on the spread of COVID-19."

Lonni Besançon, Gideon Meyerowitz-katz, Emilio Zanetti Chini, Hermann Fuchs, Antoine Flahault

Challenges in determining causality: an ongoing critique of Bendavid et al's "Assessing mandatory stay-at-home and business closure effects on the spread of COVID-19."

Lonni Besançon PhD,[1*+] Gideon Meyerowitz-Katz,[2+] Emilio Zanetti Chini PhD,[3+] Hermann Fuchs Dr.rer.nat,[4+] Antoine Flahault PhD.[5+]

[1]Faculty of Information Technology, Monash University, Clayton, Australia,
+33689902815 lonni.besancon@gmail.com

[2]University of Wollongong, Wollongong, Australia
gideon.meyerowitzkatz@health.nsw.gov.au

[3]Department of Economics and Law, Sapienza University Rome, Italy,
+39 0649766354 emilio.zanettichini@uniroma1.it

[4]Independent
+496221869701 hxfuchs@t-online.de

[5]Institute of Global Health, Faculty of Medicine, University of Geneva,
Campus Biotech, Chemin des Mines 9, CH-1202 Geneva, Switzerland
+41786725063 Antoine.Flahault@unige.ch

[+]Authors contributed equally to the work
[*]Corresponding author

Dear Editors,

We are happy to respond to Bendavid et al. on the matter of their paper [1] (the authors, henceforth). Given the subject matter impacts on lives across the globe, we are pleased to have the opportunity to continue this worthwhile discussion. While the authors have written a response [2] to our initial concerns [3,4,5], we feel that it falls short in a number of key ways, and thus the paper still does not propose a useful assessment of the efficacy of Non-Pharmaceutical Interventions (NPIs) against COVID-19.

1. Sample Size and Assessment Criteria

We are confused by the authors' response to our questions regarding sample size and the inclusion/exclusion criteria they used. First, on sample size, while the authors have indeed combined regional estimates, even within the paper itself they agree that there are 16 primary comparisons between the total sample of 10 countries. The primary analysis, therefore, is indeed limited to the very small sample size of 10 (or perhaps 16) which remains a choice that significantly limits the analysis in important ways. An analogy to the argument of the authors, in the field of clinical trials, would be to argue that a study with 100 patients does not have a sample size of 100 patients since the drugs has been in the millions of cells of each patient, but that the results are presented aggregated by patient in the end. As also noted by the John Hopkins institute's review of the paper [6], while sub-national data analysis is one of the strengths of the initial manuscript, the fact that the authors only included 10 of the many countries with sub-national data available is one of the key limitations of the study.

Concerning the exclusion criteria used, the authors seem to point out in their own response to a contradiction. In their initial manuscript [1] the authors explained that they only included countries with sub-national data available. In their response [2], the authors note that they excluded countries with restrictive measures but few cases. This first highlights the fact that the exclusion criteria was not presented in the original manuscript [1]. Then, the authors argue that this exclusion is justified because there is "no evidence beyond the anecdotal" that restrictive NPIs can control cases, which makes very little sense considering that this is precisely the question the paper is presumably attempting to answer. Excluding these countries seems to be a clear example of confounding by indication [7]. If mrNPIs are indeed associated with fewer cases, but countries with very low numbers of cases are excluded, by definition the analysis will fail to find an effect of mrNPIs where one exists.
Finally, on the matter of sample size and exclusion criteria, the authors have not only excluded countries with few cases. It is fairly trivial to include other countries with many cases - such as Brazil - however such countries seem to also have been excluded. All of these points considered, it would seem that our initial criticism of the sample size and inclusion criteria still remain valid despite the authors' response.

2. Country Classification

The authors respond to the criticism that their decisions were arbitrary by simply disagreeing. Yet, the authors have not provided, in these two manuscripts, any rationale for the categorization that they have done nor have they given any coding scheme to classify countries should anyone wish to extend their analysis in future work. This is, it seems, an admission that the classification is arbitrary, or subjective. If these decisions were not arbitrary, it would be useful for the authors to publish a fulsome accounting of the difference between a more and less restrictive NPI country, with particular attention given to how sub-national units can be vastly different.

Indeed, this accounting seems extremely important more broadly for the paper and the argument from the authors. While they assert that their distinction "characterizes the countries well", there is, it seems, no factual basis to this claim. Without a rigorous examination of what makes a NPI "more" or "less" restrictive, and why each country was categorized as such, the analysis simply represents the opinions of the authors and has no underlying scientific rationale. The authors may consider these countries more or less restrictive, but unless they explain why and how these classifications came about, it is hard to garner meaning from the analysis. By many measures South Korea is in fact a "more" restrictive country. As explained in one of the letters [3], it had one of the longest school closures in the world, and school closures is considered as one of the strictest NPI as the recent heated debate over this measure has shown (e.g., see [8,9]). This idea is even reinforced by looking at the stringency index [10] for all specified countries as calculated by OurWorldInData [11], we can see that South Korea is one of the countries that implemented much stricter NPIs during the time period examined (see Figure 1). An even more compelling image is visible when looking at the Containment and Health Index [10] (see Figure 2) for which South Korea is now the second most restrictive country only behind Italy. Much like any index, these two have inherent limitations, but they provide an objective categorization of countries based on how restrictive their measures have been. When applied to the countries selected by the authors, these two indexes show, in addition to our initial arguments that South Korea had measures that would be considered in most countries

as restrictive, that the classification done by the authors does not hold in many regards. Since the authors have not yet provided in their initial article nor in their response their coding scheme for country classification, our argument that it is arbitrary or subjective thus stands. The authors may, of course, disagree with this categorization of South Korea as a mrNPI, but if so, they should provide an objective reason rather than simply dismissing the criticism.

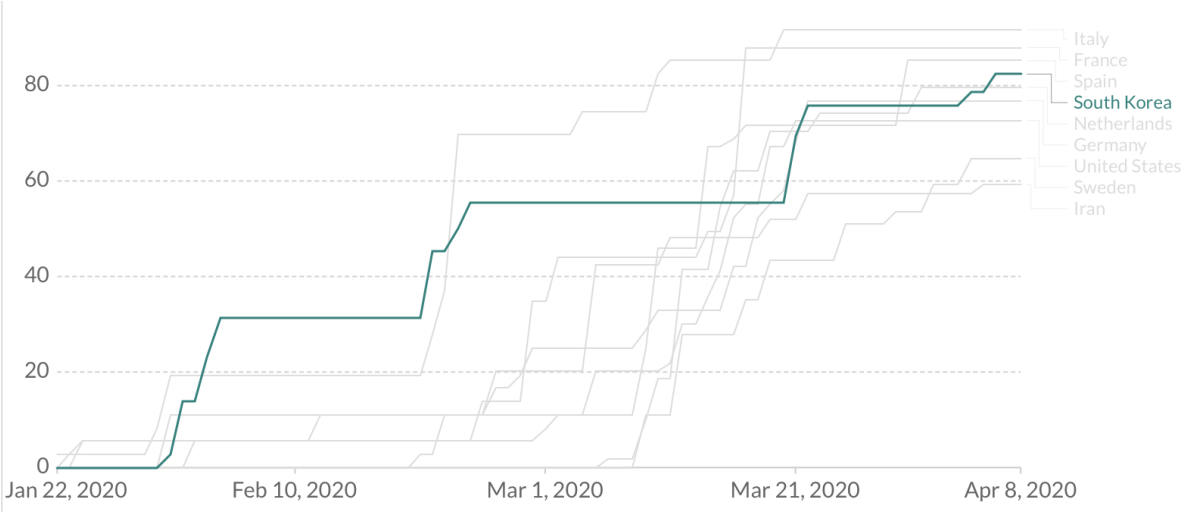

Figure 1: Stringency Index [11] of all countries included by the authors until the maximum cut-off date as specific in the supplementary materials of the original manuscript [1]. England is not included as OurWorldInData only provided Stringency Index data for United Kingdom. Image source:OurWorldInData [11].

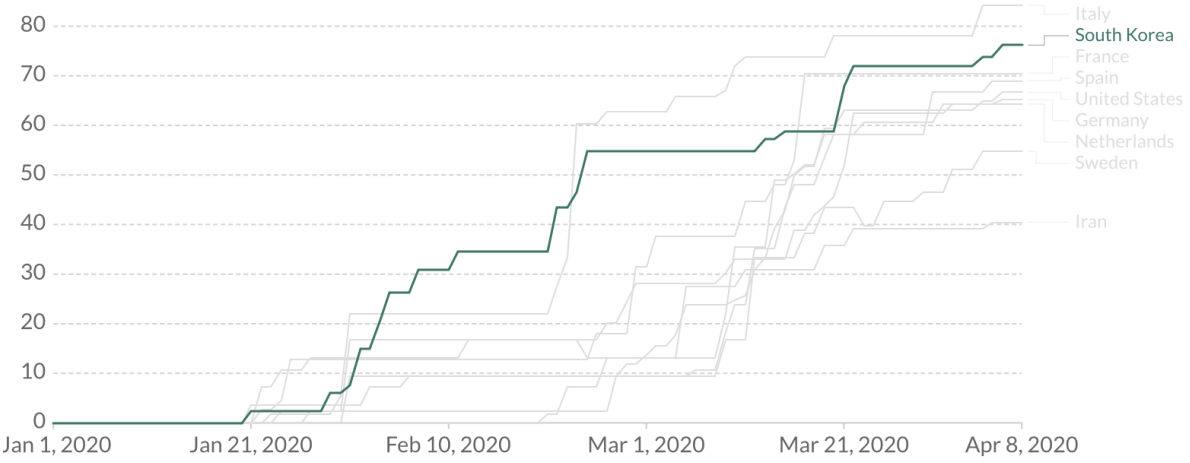

Figure 2: Containment and Health Index [12] of all countries included by the authors until the maximum cut-off date as specific in the supplementary materials of the original manuscript [1]. England is not included as OurWorldInData only provided Containment and Health Index data for the United Kingdom. Image source: OurWorldInData [12].

3. Issues in the "policy" variable

There are two points that we would like to rebut and question concerning the modeling used. First, in Section 4 of their reply, the authors solve the problem of the definition of "Policy" variable as dichotomous. Surprisingly, then they write that they "*implement panel regression model where coefficient on Policy$_{\{pcit\}}$ variables identify "breaks*" [the quotes are by authors] *in case growth patterns in each sub-national unit following the implementation of each NPI identified by specific Policy$_{\{pcit\}}$ variables rather than a difference-in-difference as suggested by Zanetti Chini*" [4]. The so called "breaks" (defined as "structural breaks" in econometric literature to distinguish a break that produces perduring effects in the path of the time series under investigation from other ones that can be explained by cyclical oscillations or pure noise) cannot be identified by the coefficient of Policy$_{\{pcit\}}$. This is a discrete-choice model for panel data, not a model for structural breaks. Structural breaks require completely different models and statistical treatment like spline and eventually have to be tested properly. In any case, it cannot be addressed by imputing, sic et simpliciter, this meaning to a coefficient.

Second, the authors explain in their response on the issues of timing and lags, identified in all three letters [3,4,5] that they do not make a difference. The authors point out that the "timing of each NPI in each subnational unit of each country is explicitly modeled in the *Policy*$_{\{pcit\}}$ variables". We think that their answers here miss the point of all three letters. There will not be a unique number of days between declaring an NPI and notable effects in the daily case numbers. Some responses are earlier, others later. Since a lot of factors such as individual behavioral responses have to be factored in (see e.g., [14,15,16,17]), there is a distribution of time lags leading to a smooth temporal onset of the effect. *Policy*$_{\{pcit\}}$ is a binary variable in the model and therefore attributes NPI-induced growth reductions prior to the day of switching *Policy* to 1 to the pre-NPI period, and takes the not yet fully developed reductions in the days after as the complete NPI effect. This decreases the effective pre-post difference, even if the day of switching has the lag equal to the mean value of the lag distribution.

4. Data cut-off

In their response, Bendavid et al. correctly state: "Fuchs worries about omitting the period of declining daily case numbers...". As a reason, he emphasized that this decline is claimed to be the main benefit of rigorous NPIs and provides the most prominent negative contributions to growth rates. In their original paper the authors defined such negative contributions as the signature of NPIs, and at the same time they suppress the most prominent negative contributions provided in the period after the start of rigorous NPIs, without explicit mention nor reasons. Mention now is supplied in their response: "The data that we include cover the period up to the elimination of rapid growth in the first wave", i. e. the period in which the daily case numbers form a sort of maximum, the subsequent descent being excluded, to the detriment of of the signature of rigorous NPIs. A foundation for the data cut-off is still missing.

5. Issues on cases reduction

The arguments in the Authors response on "not very implausible values" of 0.4 or - 028 logarithmic growth change due to rigorous NPIs miss the point. The largest beneficial growth change of - 0.28 conceded to rigorous NPIs by the authors analysis, in the original paper was denoted as "modest". This qualification is criticized by Fuchs as misleading. The logarithmic -0.28 growth change is equivalent to a factor 2 of reduction of daily case number

within 2.5 days and thus sufficient to neutralize the most dramatic exponential increase of Covid-19 cases observed - this is not a "modest" reduction by any reasonable definition. This quite beneficial value of -0.28 growth reduction is certainly not as exceptional as presented by the authors. that unilaterally attenuates the quantitative effect of rigorous NPIs, via the various approximations discussed above,.

6. Estimating the NPI effect

The authors' neglecting of the Diff-in-Diff is surprising, since, in the equation on p. 3 of the original paper, θ_0 are fixed effects of subnational units and δ_{ct} are country specific day-of-week fixed-effects. This is a canonical specification of a Diff-in-Diff estimation: subnational units of a certain country differ among them in levels but not in the trend, assumed by the authors as common in all the sub-national units of that country. As mentioned in Zanetti Chini's reply [4], this assumption is not sound, and this can be proved by looking at the data of Italian regions, for example.

If Diff-in-Diff is not used in this context, it is impossible to understand how the estimates of the model parameters have been made. Are these obtained by Least Squares? If so, what kind? Grouped? Pooled? Each one of these estimators relies on specific assumptions that need to be properly discussed in the context of the empirical strategy. Without this information, any code replication becomes useless, as the statistical methodology that drives the available coding is missing.

Moreover, the motivation that the authors give to the non-use of Diff-in-diff estimation (which, contrary to their response, is not a suggestion but an attempt to understand what precisely they have done) is not really a motivation. Namely they write: "*We do not pass a strong verdict on the role of parallel trend assumption for causal identification here, but note that if it were indeed critical, that would invalidate most assessments of NPI effects that use similar econometric approaches, since the baseline trends are unique and highly nonlinear in each subnational unit*". This sentence does not seem to make sense with respect to the uniqueness of the trend: aren't the authors using two units for each comparison, so that a small panel with i=2, hence with two individual trends can be constructed? Or are they computing a common trend among these two individuals? But yet again, how is this done? Is it via cointegration analysis? This is not explained in the submissions. Moreover the assertion is also inaccurate in the part of the nonlinearity. In fact, a substantial portion of the econometric literature addresses nonlinear panel data (and discrete-choice models), see e.g., [13, 18, 19].

Finally, the overconfidence in randomizations seems inappropriate. The authors write: *"Randomization has been increasingly used for assessing the impact of real-world policies, and the value of knowing the benefits of NPIs, especially those with large health and welfare costs, would be enormous*". Some of the past literature (e.g., [20]) argues in the opposite direction: in fact the estimates from experiments can be severely biased when the comparison is done using different models, so that the use of nonexperimental estimators is still fully justified.

7. Conclusion

Overall, we are forced to restate our previous position, which is that this paper does not allow us to meaningfully assess the efficacy of NPIs against COVID-19. It is not possible to know from this study whether restrictive NPIs work, do not work, or even how we might define a country's response as more or less "restrictive".

Conflict of interests:

The authors declare no conflict of interests.

Acknowledgements:

Emilio Zanetti Chini would like to thank Enrico Rettore for his support and suggestions.

References:

[1] Bendavid E, Oh C, Bhattacharya J, Ioannidis JPA. Assessing mandatory stay-at-home and business closure effects on the spread of COVID-19. Eur J Clin Invest. 2021 Jan 5:e13484. doi: 10.1111/eci.13484. Epub ahead of print. PMID: 33400268.

[2] Bendavid, E, Oh, C, Bhattacharya, J, Ioannidis, JPA. Response to Letters Re: 'Assessing mandatory stay- At- Home and business closure effects on the spread of COVID- 19'. *Eur J Clin Invest*. 2021; 00:e13553. https://doi.org/10.1111/eci.13553

[3] Besançon L, Meyerowitz-Katz G, Flahault A. Sample size, timing, and other confounding factors: toward a fair assessment of stay-at home orders. *Eur J Clin Invest*. 2021;12:e13518. https://doi.org/10.1111/eci.13518

[4] Zanetti Chini, E. Letter to Editor. *Eur J Clin Invest*. 2021;e13556. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/eci.13556

[5] Fuchs H. Comment on Bendavid E, Oh Ch, Battacharya J, Ioannidis JPA Assessing mandatory stay-at-home and business closure effects on the spread of covid-19. *Eur J Clin Invest*. 2021;4:e13529. https://doi.org/10.1111/eci.13529

[6] Assessing mandatory stay-at-home and business Closure effects on the spread of Covid-19: NCRC. (2021, March 13). Retrieved April 04, 2021, from https://ncrc.jhsph.edu/research/assessing-mandatory-stay-at-home-and-business-closure-effects-on-the-spread-of-covid-19/

[7] Confounding by indication. (2019, July 12). Retrieved April 04, 2021, from https://catalogofbias.org/biases/confounding-by-indication/

[8] Vogel, G. (2021). Data in paper about Swedish schoolchildren come under fire. *Science*, *371*(6533), 973–974. https://doi.org/10.1126/science.371.6533.973

[9] (2021) Correspondence: Open Schools, Covid-19, and Child and Teacher Morbidity in Sweden. *New England Journal of Medicine*. https://dx.doi.org/10.1056/NEJMc2101280

[10] Hale, T., Angrist, N., Goldszmidt, R. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat Hum Behav*(2021). https://doi.org/10.1038/s41562-021-01079-8

[11] Covid Stringency Index. Retrieved April 4, 20201, from https://ourworldindata.org/grapher/covid-stringency-index?tab=chart&time=2020-01-22..2020-04-08
[12] Containment and Health Index. Retrieved April 4, 20201, from https://ourworldindata.org/grapher/covid-containment-and-health-index?tab=chart&time=earliest..2020-04-08
[13] Greene, William. "Panel data models for discrete choice." *The Oxford handbook of panel data*. 2015.
[14] Chernozhukov V, Kasahara H, Schrimpf P. Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S. *J Econom*. 2021; 220(1): 23-62. https://doi.org/10.1016/j.jeconom.2020.09.003
[15] Goolsbee, A., & Syverson, C. (2021). Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020. *Journal of Public Economics*, *193*, 104311. https://doi.org/10.1016/j.jpubeco.2020.104311
[16] Gupta, S., Simon, K., & Wing, C. (2020). Mandated and Voluntary Social Distancing during the COVID-19 Epidemic. *Brookings Papers on Economic Activity,* 269-315. doi:10.2307/26996643
[17] Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., ... & Kulveit, J. (2020). Inferring the effectiveness of government interventions against COVID-19. Science. Doi: http://dx.doi.org/10.1126/science.abd9338
[18] Honoré, B.E. and Kyriazidou, E. (2000), Panel Data Discrete Choice Models with Lagged Dependent Variables. Econometrica, 68: 839-874. Doi: https://doi.org/10.1111/1468-0262.00139
[19] Arellano, M., R. Blundell, and S. Bonhomme. "Earnings and consumption dynamics: a nonlinear panel data framework." *Econometrica* 85.3 (2017): 693-734.
[20] Heckman, James J., and V. Joseph Hotz. "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training." *Journal of the American statistical Association* 84.408 (1989): 862-874.