Matters arising

The effect of interventions on COVID-19

https://doi.org/10.1038/s41586-020-3025-y

Received: 15 June 2020

Accepted: 13 November 2020

Published online: 23 December 2020

Check for updates

Kristian Soltesz[1]✉, Fredrik Gustafsson[2], Toomas Timpka[3,4], Joakim Jaldén[5], Carl Jidling[6], Albin Heimerson[1], Thomas B. Schön[6], Armin Spreco[3,4], Joakim Ekberg[3,4], Örjan Dahlström[7], Fredrik Bagge Carlson[8], Anna Jöud[9,10] & Bo Bernhardsson[1]

ARISING FROM S. Flaxman et al. *Nature* https://doi.org/10.1038/s41586-020-2405-7 (2020)

Flaxman et al.[1] took on the challenge of estimating the effectiveness of five categories of non-pharmaceutical intervention (NPI)—social distancing encouraged, self isolation, school closures, public events banned, and complete lockdown—on the spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). On the basis of mortality data collected between January and early May 2020, they concluded that only one of these, the lockdown, had been effective in 10 out of the 11 European countries that were studied. However, here we use simulations with the original model code to suggest that the conclusions of Flaxman et al. with regard to the effectiveness of individual NPIs are not justified. Although the NPIs that were considered have indisputably contributed to reducing the spread of the virus, our analysis indicates that the individual effectiveness of these NPIs cannot be reliably quantified.

Flaxman et al.[1] presented a method to estimate the effects of NPIs on the time-varying reproduction number (R_t) of SARS-CoV-2 infection. Data from 11 European countries were pooled on the basis of the assumption that the effects of NPIs on R_t are not country-specific: the factor of relative change in R_t resulting from a particular NPI was assumed to be independent of the country in which the NPI was implemented.

Some country-specific flexibility was, however, provided through the basic reproduction number (R_0) being country-specific. More notably, additional flexibility was introduced by ascribing a country-specific effect to the NPI that was introduced last in each country. This replaced the parameterization in a preprint version (Imperial College Report 13)[2], in which a country-specific effect was instead assigned to the lockdown NPI.

Our criticism concerns the final published version of the model[1,3]. Previous iterations of the model are not explicitly considered, but we reference them for two purposes: (1) to demonstrate the sensitivity of the final published model to subtle and realistic alterations in parameter values; (2) to illustrate how the modelling choices appear to lack motivation other than to introduce flexibility, which masks sensitivity issues pertaining to the fundamental structure of the model. As made evident below, we believe the core problem is that the death data are not descriptive enough to support the conclusions of Flaxman et al., which were based on simulation results obtained using an over-flexible model.

Of the 11 modelled countries, Sweden is worthy of particular attention, given that it was the only country in which no lockdown took place. As we have previously shown[4], the estimated effects of NPIs change markedly when the model is not allowed to give the Swedish data the special treatment that the country-specific last NPI parameter enables. The country-specific last NPI parameter is needed to explain the decrease of R_t supported by the Swedish death data, and to provide a good model fit despite the absence of a lockdown in Sweden.

Figure 1 shows the outcome for Sweden when executing the model[3,5] either with (Fig. 1a) or without (Fig. 1b) the last NPI adjustment in place. With the last NPI adjustment in place, the public events ban results in a mean reduction of R_t of 71% (95% credible interval: 59–81%) in Sweden, which contrasts with the negligible effect of the public events ban in the other 10 countries (less than 2% mean reduction of R_t and less than 15% with 95% credibility). Notably, the estimated effectiveness of the public events ban in Sweden is comparable to that of lockdown in the 10 countries in which one was implemented. As lockdown was the last intervention in most countries, its estimated effect comprises a pooled effect (82% mean reduction of R_t) and a separate country-specific 'last NPI' effect (mean change in R_t of between −24% and 18% for the countries considered).

The result above—that is, the public events ban and the lockdown being mutually effective in Sweden and 10 other European countries—was not addressed by Flaxman et al, which is noteworthy as this result undermines the conclusion of lockdown being especially effective. Furthermore, without the introduction of the last intervention parameter after the publication of the preprint[2], the inconsistency would have been readily visible in reported plots (Fig. 1b).

It seems unlikely to be a result of circumstance that lockdown was implemented in the 10 countries in which it had a large effect on R_t, and omitted in the single country in which the public events ban instead had a similar effect (sufficient to drive R_t below 1). An alternative hypothesis is that the infection-to-death distribution used by the model, combined with the death data that were available by early May, makes the model ascribe almost all of the reduction in R_t to the last intervention that was implemented in each country. This hypothesis is supported by executing the model code[3,5] with different interventions being defined as having occurred last in the country in which no lockdown occurred (Sweden), as shown in Fig. 2.

Exchanging the last intervention for a different one is not merely interesting from a theoretical perspective. For example, it is hard to judge whether transitioning to online teaching at high school and university levels, while keeping elementary schools and preschools open, constitutes a school closure or not. Similarly, the crowd-size limit associated with the public events ban NPI remains a parameter to be decided by the modeller. Early versions of the model defined the public events ban to have taken place in Sweden on 12 March 2020, when gatherings exceeding 500 persons were prohibited. This was later changed to 29 March 2020, when gatherings exceeding 50 persons were prohibited. These subtle alterations of the definitions alter which NPI, of school closure, public events ban, or social distancing encouraged, was the last to be implemented in Sweden. In each case, the model uses the last intervention to explain the majority of the drop

[1]Department of Automatic Control, Lund University, Lund, Sweden. [2]Department of Electrical Engineering, Linköping University, Linköping, Sweden. [3]Department of Public Health, and Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden. [4]Unit for Public Health and Statistics, Region Östergötland, Linköping, Sweden. [5]Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. [6]Department of Information Technology, Uppsala University, Uppsala, Sweden. [7]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden. [8]Acoustic Research Laboratory, National University of Singapore, Singapore, Singapore. [9]Department of Laboratory Medicine, Lund University, Lund, Sweden. [10]Department of Research and Development, Skåne University Hospital, Lund, Sweden. ✉e-mail: kristian.soltesz@control.lth.se

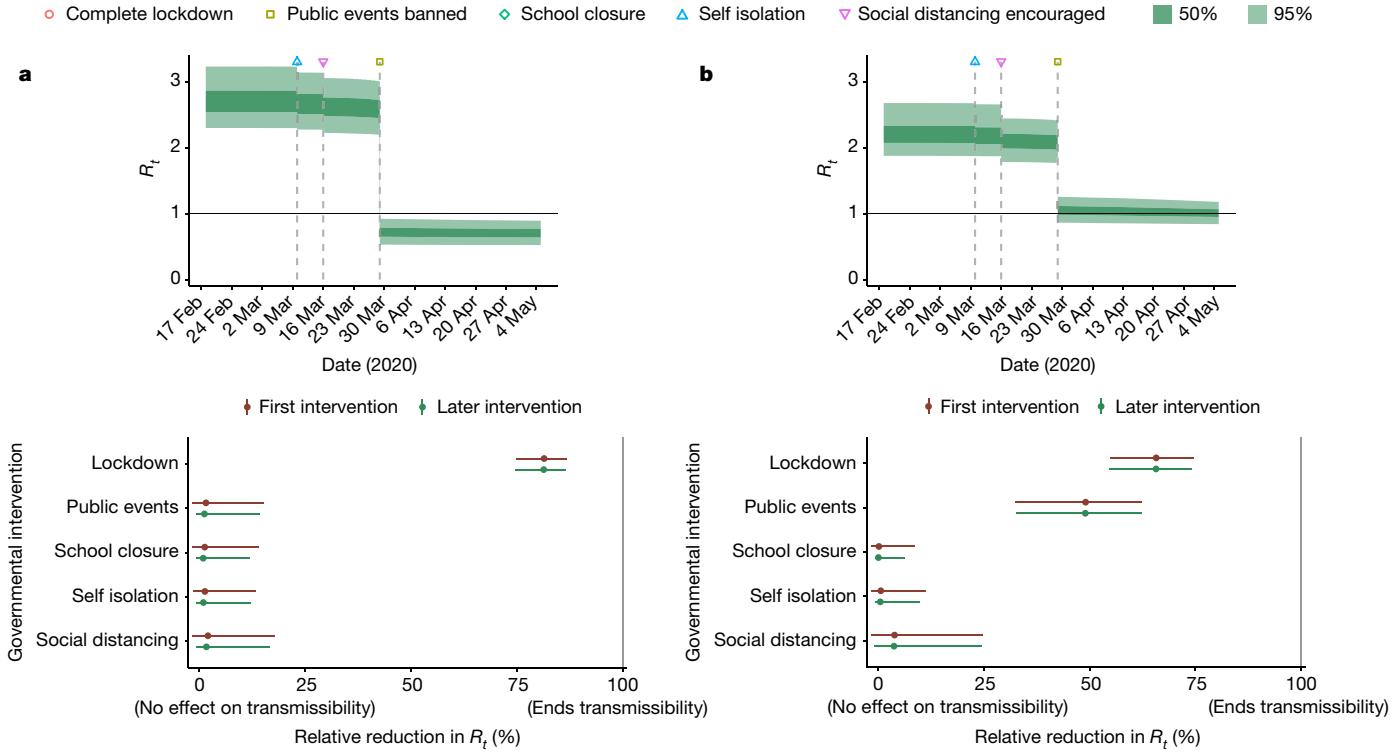


Fig. 1 | Estimated effectiveness of the public events ban in Sweden. Top, posterior credible intervals for the reproduction number R_t in Sweden. Bottom, effectiveness of the pooled interventions in the 11 modelled European countries. **a**, Reproduced results from Flaxman et al.[1], using the original model code[3,5], including a country-specific effectiveness parameter for the last NPI to be implemented in each country. This corresponds to a country-specific effectiveness for the public events ban in Sweden and for lockdown in the other 10 countries. **b**, Results using the same code, but with the 'last NPI' parameters replaced with country-specific parameters for the lockdown NPI, as in the preceding report[2]. This change does not affect the 10 countries for which lockdown was the last NPI, but for Sweden it removes the flexibility of a last NPI parameter, which is needed to explain the R_t value supported by the Swedish death data.

of R_t to below 1, which is needed to stay consistent with the decrease in reported deaths.

As mentioned above, our analyses were conducted using the original model implementation[3,5] referenced from the final published paper[1], and we have considered the definitions of NPIs reported in the preceding versions of the model[1–3] solely to highlight how small and plausible perturbations of these definitions can result in a lack of practical identifiability, in the statistical sense. Identifiability issues have to some extent been acknowledged by the authors; Flaxman et al. state that "The close spacing of interventions in time [...] means that the individual effects of the other interventions are not identifiable"[1]. However, this is overshadowed by the subsequent presentation of credible intervals for the effects of the different NPIs, and the claim that "Lockdown has an identifiable large effect on transmission (81% (75–87%) reduction)"[1]. We believe that the

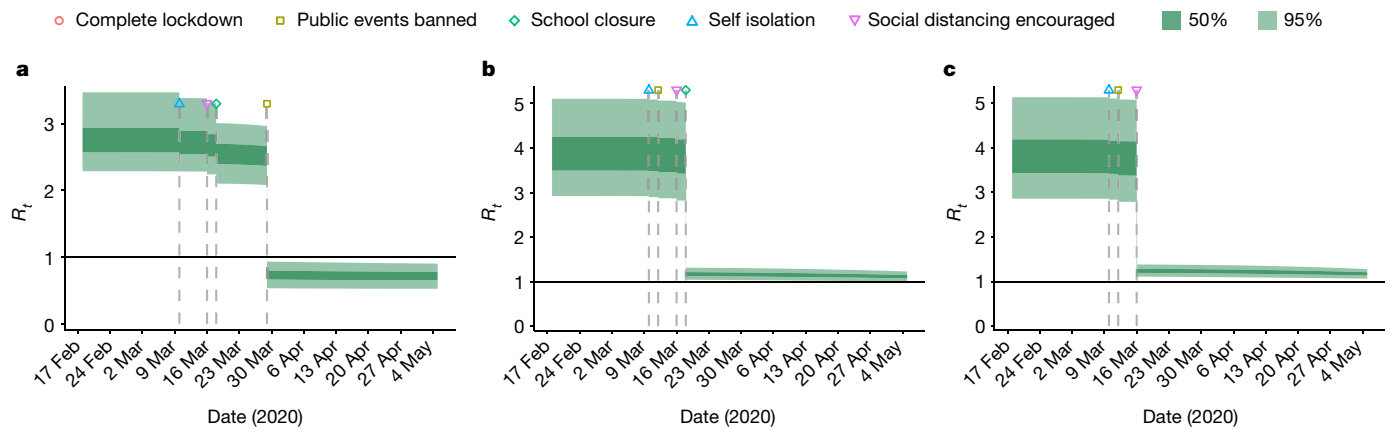


Fig. 2 | The effects of interventions on virus spread in Sweden, with slightly varying definitions of the interventions. a, School closure defined to have taken place on 18 March 2020; public events ban defined to have taken place on 29 March 2020. **b**, Same as **a**, but with the public events ban moved back to 12 March 2020. **c**, Same as **b**, but with school closure defined not to have taken place. As expected, the visual appearance of the plots is similar, with the last intervention contributing most to the reduction of virus spread. This is problematic, as the last intervention differs between **a**, **b** and **c**, with each relying on equally motivated NPI implementation dates that were introduced by Flaxman et al.[1,2] in different versions of the model code[3]. The conclusion is that subtle changes in the definitions of NPIs result in a great deal of variation in the estimated effectiveness of the NPI categories considered.

basis of this claim is unclear. As seen in the supplementary videos of the *Nature* article[1], the credible intervals narrow as more data become available, further hiding the identifiability problems of the underlying model and potentially giving the results a false sense of reliability.

Our point here is not to argue whether or not a school closure took place in Sweden, or what the most appropriate crowd-size limit is. Instead, our findings highlight that the model presented by Flaxman et al. is very sensitive to reasonable, minor changes in the input data. As indicated by our simulation examples, and further supported by our previous analyses[4], there is a fundamental problem with the identifiability of the effectiveness of individual NPIs, including the lockdown. This problem is caused by the close temporal spacing between the implementation of these NPIs throughout Europe. In particular, we note in relation to the lockdown NPI that an estimated value that is considerably larger than zero should not be confused with statistical identifiability of the corresponding parameter.

Although we fully support the ambition of Flaxman et al.[1]—to estimate the effectiveness of different NPIs from the available data—we find the underlying modelling approach problematic. Flexible parameterization leads to issues with identifiability, which are masked by model assumptions. In particular, we find it questionable to designate a country-specific effectiveness parameter to the last NPI that was introduced in each country. Besides the problems illustrated in Fig. 2, with large variations in the estimated effectiveness of NPIs, this prohibits prospective use of the model, as it is unknown at any given time whether the latest NPI will also be the last to be implemented in a particular country.

We conclude that the model[1,3] is in effect too flexible, and therefore allows the data to be explained in various ways. This has led the authors to go beyond the data in reporting that particular interventions are especially effective. This kind of error—mistaking assumptions for conclusions—is easy to make, and not especially easy to catch, in Bayesian analysis. As NPIs are revoked, and possibly reintroduced over an extended period of time, more data will become available and practical identifiability of the separate effects of NPIs may be obtained. Until then, we suggest that the model[1,3], and its conclusion that all NPIs apart from lockdown have been of low effectiveness, should be treated with caution with regard to policy-making decisions.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

A fork of the original code and associated data, which was used to generate the figures presented here, is provided in a separate GitHub repository[5]. This fork is based on the GitHub repository commit 885466d of the original code[3], in which the README file states that it was "the exact code that was used [in the *Nature* article[1]]". We have, however, noticed discrepancies between the original code[3] and the figures in the article[1]. For example, the code that was used to generate figure 1 in Flaxman et al.[1] defines the self-isolation NPI as having been implemented as the last NPI in Spain on 17 March 2020, whereas the code in the commit defines this date as 14 March 2020.

1. Flaxman, S. et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
2. Flaxman, S. et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. https://doi.org/10.25561/77731 (Imperial College London, 2020).
3. Imperial College London. covid19model. Available at https://github.com/ImperialCollegeLondon/covid19model (2020).
4. Soltesz, K. et al. On the sensitivity of non-pharmaceutical intervention models for SARS-CoV-2 spread estimation. Preprint at https://doi.org/10.1101/2020.06.10.20127324 (2020).
5. Heimerson, A. covid19model fork. Available at https://github.com/albheim/covid19model (2020).

Acknowledgements We acknowledge Ericsson Research for hosting our model runs in their data centre. This work has been partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation; the ELLIIT strategic research area on IT and mobile communications; the Swedish Research Council (grant reference number 2017-04989); the Swedish Foundation for Strategic Research (SSF) via the project ASSEMBLE (grant reference number RIT15-0012); and ALF Grants, Region Östergötland.

Author contributions Study conceptualization: B.B., F.G., J.J. and K.S. Analysis and interpretation of model weaknesses: B.B., F.B.C., J.J. and K.S. Code preparation and execution: A.H. and C.J. Background literature review: B.B., A.J., A.S. and T.T. Manuscript writing and reviewing: B.B., O.D., J.E., J.J., A.J., T.B.S., K.S., A.S. and T.T.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-3025-y.
Correspondence and requests for materials should be addressed to K.S.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Kristian Soltesz

Last updated by author(s): Oct 7, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	The probabilistic programming language STAN has been used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We have used data from http://ec.europa.eu/, fetched from https://github.com/ImperialCollegeLondon/covid19model. All data has been included in our openly accessible Github repository https://github.com/albheim/covid19model, providing full transparency and reproducability.

Field-specific reporting

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Reply to: The effect of interventions on COVID-19

https://doi.org/10.1038/s41586-020-3026-x

Published online: 23 December 2020

Check for updates

Seth Flaxman[1,3], Swapnil Mishra[2,3], James Scott[1,3], Neil Ferguson[2,3], Axel Gandy[1,3] & Samir Bhatt[2,3]✉

REPLYING TO K. Soltesz et al. *Nature* https://doi.org/10.1038/s41586-020-3025-y (2020)

The accompanying Comment[1] concerns our original paper, Flaxman et al.[2], in which we introduced a Bayesian hierarchical model to estimate the transmission intensity (in terms of the time-varying reproduction number, R_t) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from observed counts of coronavirus disease 2019 (COVID-19)-related deaths. We parameterized R_t in terms of a binary set of government-mandated non-pharmaceutical interventions (NPIs), with the motivation of examining how effective NPIs were at controlling the transmission of SARS-CoV-2. We concluded that the NPIs that were widely used across Europe successfully drove R_t below 1, thus controlling the epidemic. However, we were unable to disentangle the effect sizes of the NPIs we considered, except for concluding that lockdown had a stronger effect than the other NPIs.

We start by giving some background on the evolution of the paper. Our first preprint, released as Imperial College Report 13[3], was based on data up to 28 March 2020 and used a simpler model in which the effect size of each intervention on transmission is the same across countries (here referred to as a full pool model; in our published paper[2] we use this model for the single-country analyses reported in supplementary discussion 8 of the paper). As more data became available (Flaxman et al.[2] uses data up to 4 May 2020), more heterogeneity between countries became evident and we therefore extended the full pool model.

This extended model, which is the one used in Flaxman et al.[2], includes a random effect, with the aim of capturing country-specific variation in the effectiveness of the last government-mandated intervention or interventions; for example, lockdown in Italy, lockdown and a ban on public events in the UK, and a ban on public events in Sweden (see Extended Data Fig. 1). Random effects are common components of statistical models to account for heterogeneity not explained by covariates[4–6].

The focus of Soltesz et al.[1] is the size of the random effect assigned by our model to the last intervention in Sweden. Specifically, a large random effect is needed to explain the Swedish data, and this could have been more explicitly stated in our original paper. Soltesz et al.[1] claim that the difference between effect sizes in a full pool model and in Flaxman et al.[2] points to our model having little practical statistical identifiability. On this basis, Soltesz et al.[1] question whether the effectiveness of lockdown can be resolved to the degree our paper stated.

The main goal of our paper was to examine multiple countries to see what worked in most places, not to explain the trajectory of the epidemic in each individual country. Although we feel that Soltesz et al.[1] raise an interesting point, we stand by our assessment that the effectiveness of NPIs can in principle be identified when looking at what worked in most countries, subject of course to the available data.

Here we present further analyses that support our finding reported in Flaxman et al.[2] that lockdown was an identifiable intervention with a major effect. We accept that additional covariates beyond the timing of mandatory measures are likely to be needed to provide a fully satisfactory explanation of the trajectory of the epidemic in Sweden, as that country relied on voluntary social distancing measures rather than government-mandated interventions.

Because our goal was to estimate which NPIs worked consistently in most countries, we argue that an analysis of the effectiveness of NPIs should be robust to leaving any one country out. In Extended Data Fig. 1 of this Reply we compare results from the full pool model (used by Soltesz et al.[1]), the model used in Flaxman et al.[2], and a partial pool model, removing one country at a time from the input data. In the partial pool model[4,6,7], all NPIs have both a random effect component shared between all countries and a country-specific random effect (via a Gaussian shrinkage prior).

In the full pool model, results for effect sizes are dependent on whether Sweden is included, hence Sweden has a very high statistical influence[8]. As seen in Extended Data Fig. 1, when Sweden is left out of the full pool model, we recover the results from Flaxman et al.[2], but when Sweden is included the estimates change markedly. This happens because the full pool model attributes a large effect size to the ban on public events to explain the Swedish death data.

However, both the model we considered in Flaxman et al.[2] and the partial pool model discussed here show consistent effect sizes across all 'held-out' (that is, excluding a given country from fitting) countries. (For space, only the UK, Italy and Sweden are shown in Extended Data Fig. 1.) This explains our choice to move from a full pool model, which is the one used by Soltesz et al.[1], to the model used in Flaxman et al.[2]

The partial pool model is what we recommend (and are currently adopting) for such analyses in future. Partial pooling allows all interventions to have a shared effect and an effect specific to each country for each intervention. Thus, it stands somewhere between a full pool model and 11 separate models, with the data informing this location. These choices mean the partial pool model has no specific affinity towards a country or a specific intervention.

To further explore issues around identifiability at an individual country level versus across countries, in Fig. 1 we present the effects of NPIs for each country from separate country-specific models, a meta-analysis of these effects, and the estimates from our various joint models. In summary, we see that although the overall mean effect for lockdown is lower in the meta-analysis, it is still the only NPI with an identifiable effect size. The individual country fits provide insight into why this occurs; the only intervention that is consistently significant

[1]Department of Mathematics, Imperial College London, London, UK. [2]MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), Imperial College London, London, UK. [3]These authors contributed equally: Seth Flaxman, Swapnil Mishra, James Scott, Neil Ferguson, Axel Gandy, Samir Bhatt. ✉e-mail: s.bhatt@imperial.ac.uk

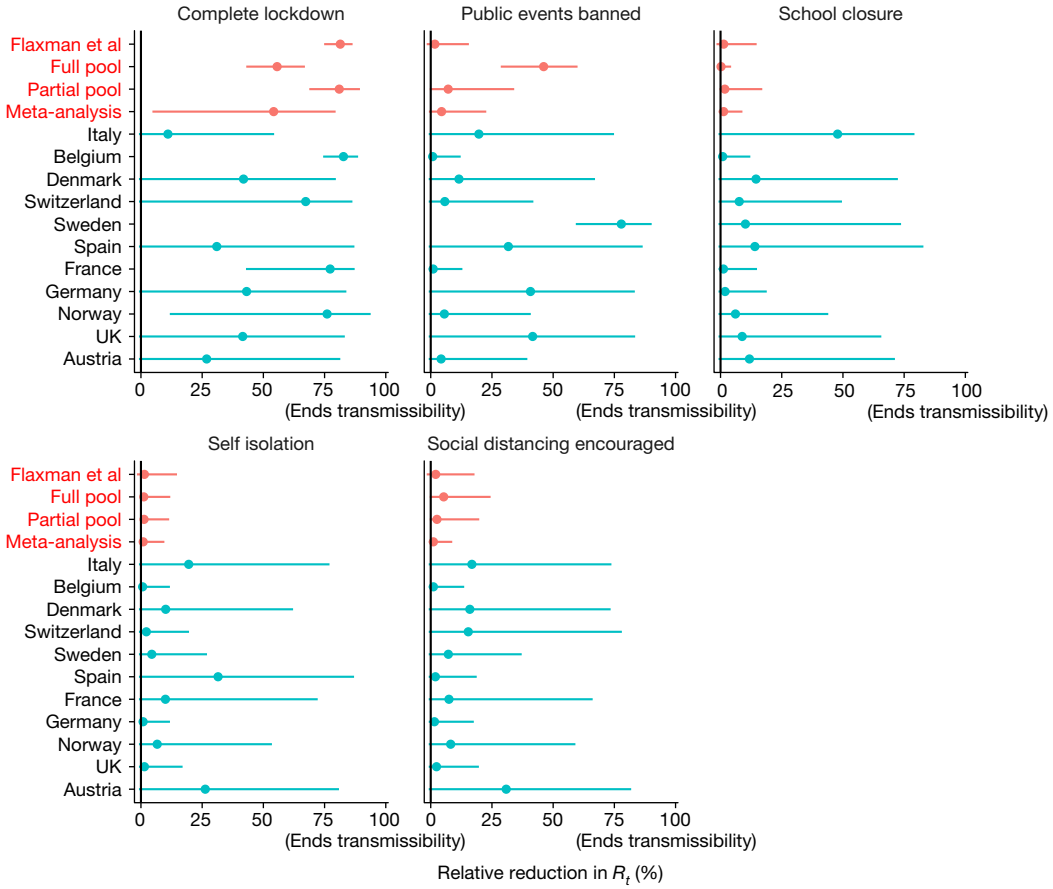


Fig. 1 | Inferred intervention effect sizes. The *x* axis shows the relative reduction in transmission. Rows show model predictions for our published model (Flaxman et al.[2]), the model from Soltesz et al.[1] (full pool), a generalized version of our published model (partial pool) and fits to individual countries (reported in supplementary discussion 8 of our original paper[2]; the model is the same one considered by Soltesz et al.[1]). We also include the mean effect size derived from a meta-analysis (mean across countries for each individual sample) of the individual country effects.

is lockdown (and the banning of public events in Sweden, as discussed in the legend of Extended Data Fig. 1).

Considering the single-country models, we see that the effectiveness of lockdown is not merely the result of a modelling choice on our part. In countries such as Italy, no intervention is estimated to be significantly more effective than any other. The lack of identifiability is not a feature inherent to our model, but a limitation of the data available at the time, as we noted in our paper[2]. In particular, although we noted the close spacing of interventions in time, in a few countries lockdowns and the banning of public events coincided exactly (for example, in the UK). The result is that in the separate country analyses and full pooling (Soltesz et al.[1]), there is a strong posterior correlation between the effects of these two NPIs (Pearson correlation of −0.59 in separate country analyses; −0.67 in full pooling analysis): when one has a large effect, the other by necessity has a small effect.

It is crucial to note here that Soltesz et al.[1] are correct that the relative effect of different interventions cannot be disentangled for a single country treated in isolation. This probably reflects the limitation of using time series of deaths to infer transmission changes, given the high mean and variance of the distribution of the delay from infections to deaths. However, when looking across multiple countries, all aggregate models suggest that the lockdown intervention has an identifiable effect. This is true for all models considered, including the full pool model of Soltesz et al.[1], in which the posterior probability that lockdown is the most effective intervention is 76%, as compared with 96% in the meta-analysis and 100% in both partial pooling models. Therefore, by simultaneously analysing trends in multiple countries, our model has the ability to resolve an identifiable signal of the effect of lockdown.

To further reinforce this point, we also undertook a simulation study examining the extent to which the timing and ordering of the interventions used fundamentally limit the ability to infer effect sizes reliably.

We used our model to simulate synthetic epidemics for all 11 countries, keeping the original timing and ordering of interventions and the same initialization priors, but assigning hypothetical effect sizes to each intervention. We assigned small effect sizes (5% with a tight prior) to all but one NPI, giving the remaining one an effect size with a mean of 59%, also with a tight prior, across countries. In addition, to better reflect reality, we simulate another, country-varying NPI, at a random time, which we treat as unobserved in our model. This unknown and unobserved NPI has a diffuse prior bounded between 0% and 100%, with a mean of 27%, and it is included to assess whether an omitted variable (for example, representing spontaneous behaviour change in response to government messaging) could bias the effect-size estimates of our modelled NPIs. We keep the dates for NPIs the same as the ones in the real data to account for concerns raised about the possible effects of coincident timing on the identifiability of effect sizes.

Next, we fitted the Flaxman et al. model[2] to these simulated datasets (20 different simulations for each setting). As shown in Fig. 2, the estimates from the Flaxman et al. fitted model[2] (without any information about the unobserved NPI) are in agreement with the NPI effect sizes that were used to generate the data. This analysis provides further evidence that the results we found were not merely artefacts of the modelling approach; if there is a strong signal in the data for a specific NPI, our model can recover it.

However, this does not on its own show that the converse is necessarily true. To evaluate competing explanations for the observed dynamics of transmission, additional empirical evidence—such as NPI efficacy or alternative epidemiological explanations—is needed.

In summary, we believe that the additional evidence we present here confirms that the key conclusion from our paper is robust: within our model we can conclude that all NPIs together brought the epidemic under control; lockdown had the strongest effect and was identifiable; and the precise effect of the other NPIs was not identifiable.

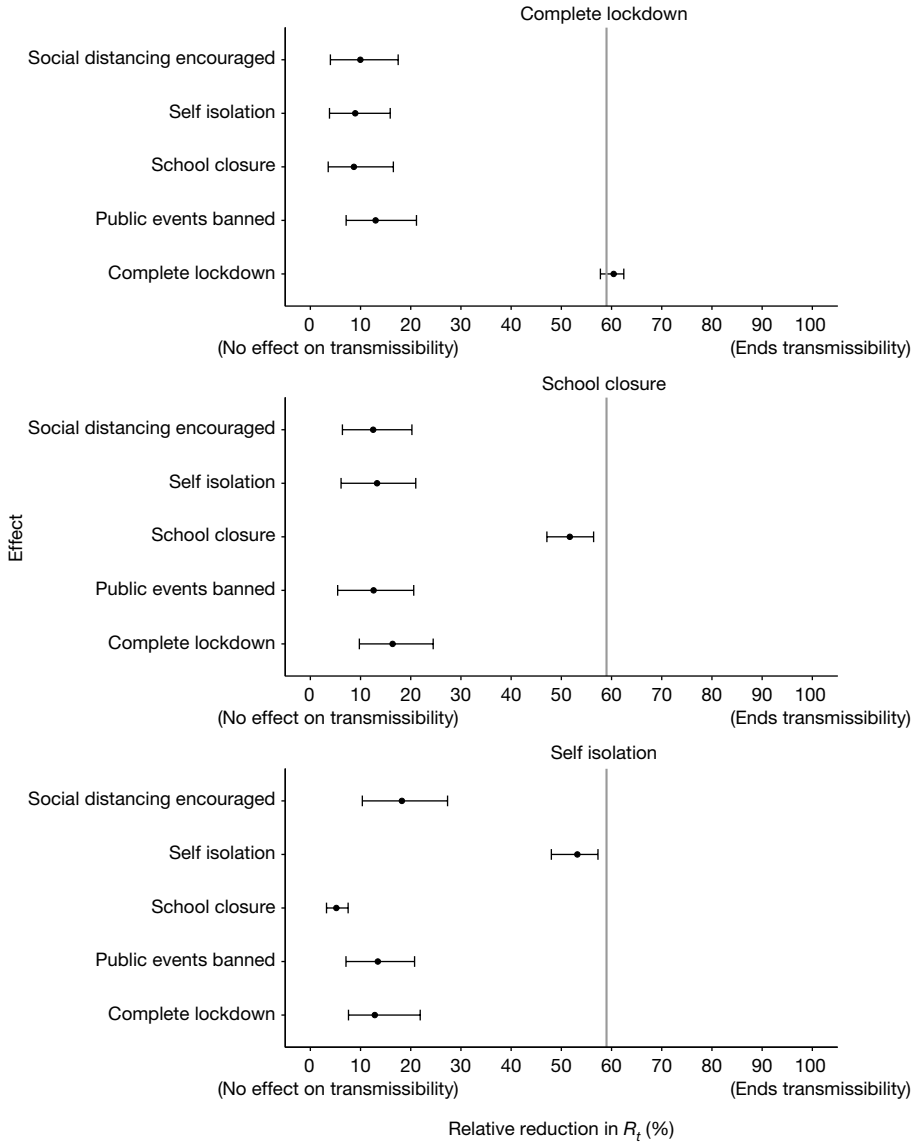


Fig. 2 | Estimated effect sizes from simulated data. Top to bottom, three separate simulations for lockdown, school closure and self isolation, with a mean of 59% effect size (grey lines), were repeated 20 times each. In each panel, effect sizes from the Flaxman et al. model[2] fitted to the 20 simulations are plotted as the mean point estimate with 95% intervals from the 20 runs.

Although our work shows that lockdowns had the largest effect size, we did not and do not claim that they were the only path to controlling the virus; merely that among the NPIs we considered, lockdown was the most effective single measure. We of course acknowledge that improvements could be made to our model, such as including random processes, partial pooling (see above) or more prior analysis. Improved models and more granular information on NPIs and population behaviour will in future hopefully give a more nuanced understanding of which measures—whether mandatory or voluntary—contributed most to reductions in transmission.

Data availability

No new data are used in this response; all data are available in the original publication.

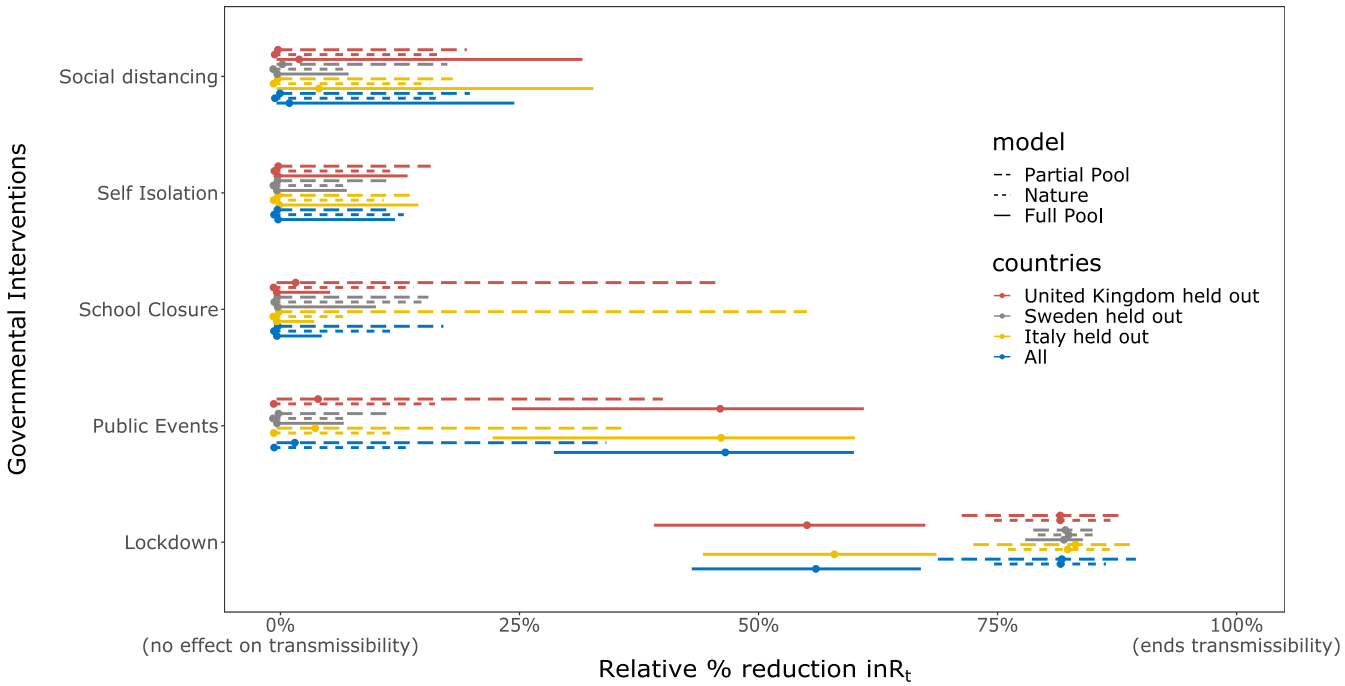
1. Soltesz, K. et al. The effect of interventions on COVID-19. *Nature* https://doi.org/10.1038/s41586-020-3025-y (2020).
2. Flaxman, S. et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature***584**, 257–261 (2020).
3. Flaxman, S. et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. https://doi.org/10.25561/77731 (Imperial College London, 2020).
4. Gelman, A. & Hill, J. *Data Analysis Using Applied Regression and Multilevel/Hierarchical Models* (Cambridge University Press, 2006).
5. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* 2nd edn (Chapman & Hall/CRC, 2003).
6. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics***38**, 963–974 (1982).
7. Efron, B. & Morris, C. Data analysis using Stein's estimator and its generalizations. *J. Am. Stat. Assoc.* **70**, 311–319 (1975).
8. Belsley, D. A., Kuh, E. & Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (John Wiley & Sons, 2005).

Acknowledgements We would like to acknowledge Microsoft AI for Health for their grant of Azure Cloud Computing.

Author contributions All authors analysed, wrote and drafted this response. The work done in this response was performed by a subset of the original authorship, and therefore many of the original authors have been excluded. A new author who contributed to the R package used in this work has been added (J.S.).

Competing interests The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.B.



Extended Data Fig. 1 | Effect-size plot for each intervention for a partial pool model, the Flaxman et at. model and the Soltesz et al. full pool model. Data are mean and 95% credible intervals. The blue lines show the fits with all 11 countries; additional lines are from holding countries out and refitting. For example, the red line for shows a fit in which the UK data are not used, and the model is fitted to the 10 remaining countries. These results show that the full pool model is not robust to outliers.