The effectiveness of eight nonpharmaceutical interventions against COVID-19 in 41 countries

Jan M. Brauner MD[@,*,1,2], Sören Mindermann[*,1], Mrinank Sharma[*,3], Anna B. Stephenson[4], Tomáš Gavenčiak PhD[5], David Johnston[6,7], Gavin Leech[8], John Salvatier[7], George Altman MBChB[9], Alexander John Norman[5], Joshua Teperowski Monrad[2], Tamay Besiroglu[10], Hong Ge PhD[11], Vladimir Mikulik[5], Meghan A. Hartwick PhD[12], Prof Yee Whye Teh PhD[13], Prof Leonid Chindelevitch PhD[14], Prof Yarin Gal PhD[+,1], Jan Kulveit PhD[+,2]

[1]*OATML, Department of Computer Science, University of Oxford, Oxford, United Kingdom.*
[2]*Future of Humanity Institute, University of Oxford, Oxford, United Kingdom.*
[3]*Department of Computer Science, University of Oxford, Oxford, United Kingdom.*
[4]*Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA.*
[5]*No affiliation.*
[6]*College of Engineering and Computer Science, Australian National University, Australia.*
[7]*Quantified Uncertainty Research Institute, California, USA.*
[8]*School of Computer Science, University of Bristol, Bristol, United Kingdom.*
[9]*School of Medical Sciences, University of Manchester, Manchester, United Kingdom.*
[10]*Faculty of Economics, University of Cambridge, Cambridge, United Kingdom.*
[11]*Engineering Department, University of Cambridge, Cambridge, United Kingdom.*
[12]*Tufts Initiative for the Forecasting and Modeling of Infectious Diseases, Tufts University, Boston, USA.*
[13]*Department of Statistics, University of Oxford, Oxford, United Kingdom.*
[14]*Computational Epidemiology Lab, School of Computing Science, Simon Fraser University, Burnaby, Canada.*

Abstract

Background. Governments are attempting to control the COVID-19 pandemic with non-pharmaceutical interventions (NPIs). However, it is still largely unknown how effective different NPIs are at reducing transmission. Data-driven studies can estimate the effectiveness of NPIs while minimizing assumptions, but existing analyses lack sufficient data and validation to robustly distinguish the effects of individual NPIs.

Methods. We collect chronological data on NPIs in 41 countries between January and May 2020, using independent double entry by researchers to ensure high data quality. We estimate NPI effectiveness with a Bayesian hierarchical model, by linking NPI implementation dates to national case and death counts. To our knowledge, this is the largest and most thoroughly validated data-driven study of NPI effectiveness to date.

[@]Correspondence to jan.brauner@eng.ox.ac.uk
[*]Equal contribution
[+]Contributed equally to senior authorship
This work was conducted in association with the EpidemicForecasting.org project

Results. We model each NPI's effect as a multiplicative (percentage) reduction in the reproduction number R. We estimate the mean reduction in R across the countries in our data for eight NPIs: mandating mask-wearing in (some) public spaces (2%; 95% CI: -14%–16%), limiting gatherings to 1000 people or less (2%; -20%–22%), to 100 people or less (21%; 1%–39%), to 10 people or less (36%; 16%–53%), closing some high-risk businesses (31%; 13%–46%), closing most nonessential businesses (40%; 22%–55%), closing schools and universities (39%; 21%–55%), and issuing stay-at-home orders (18%; 4%–31%). These results are supported by extensive empirical validation, including 15 sensitivity analyses.

Conclusions. Our results suggest that, by implementing effective NPIs, many countries can reduce R below 1 without issuing a stay-at-home order. We find a surprisingly large role for school and university closures in reducing COVID-19 transmission, a contribution to the ongoing debate about the relevance of asymptomatic carriers in disease spread. Banning gatherings and closing high-risk businesses can be highly effective in reducing transmission, but closing most businesses only has limited additional benefit.

Introduction

Worldwide, governments have mobilised vast resources to fight the COVID-19 pandemic. A wide range of nonpharmaceutical interventions (NPIs) has been deployed, including drastic measures like stay-at-home orders and the closure of all nonessential businesses. Recent analyses show that these large-scale NPIs are jointly effective at reducing the virus' effective reproduction number,[1] but the effects of individual NPIs are still largely unknown. As time progresses and more data become available, we can move beyond estimating the combined effect of a bundle of NPIs and begin to understand the effects of individual interventions. This can help governments efficiently control the epidemic, while removing less effective NPIs, to ease the burden put on the population.

A promising way to estimate NPI effectiveness is data-driven, cross-country modelling: inferring effectiveness by relating the NPIs implemented in different countries to the course of the epidemic in these countries. To disentangle the effects of individual NPIs, we need to leverage data from multiple regions with diverse sets of interventions in place. With some exceptions,[1–4] previous data-driven studies focus on single NPIs or single geographical regions (Table F.4). In contrast, we evaluate the impact of eight NPIs on the epidemic's growth in 34 European and 7 non-European countries. To our knowledge, this is the largest data-driven study of NPI effects on COVID-19 transmission to date. The data gathered is publicly available.

To isolate the effect of individual NPIs, we also require sufficiently diverse data. If all countries implemented the same set of NPIs on the same day, the individual effect of each NPI would be unidentifiable. However, the COVID-19 response was far less coordinated: countries implemented different sets of NPIs, at different times, in different orders (Figure 1).

Even with diverse data from many countries, estimating NPI effects remains a challenging task. First, many components of a model, such as epidemiological parameters and interactions between NPIs, are only known with high uncertainty. Two recent replication studies demonstrated that NPI effectiveness estimates can be highly sensitive to arbitrary modelling decisions,[5] especially when based on insufficient data.[6] Second, the data are retrospective and observational, meaning that unobserved factors could confound the results. Third, large-scale public NPI datasets suffer from frequent inconsistencies[7] and missing data.[8] For these reasons, the data and the model should be carefully validated. Insufficiently validated results should not be used to guide policy decisions. We perform, to our knowledge, by far the most extensive validation of any COVID-19 NPI effectiveness results to date, a crucial but largely absent or incomplete element of NPI effectiveness studies.[5]

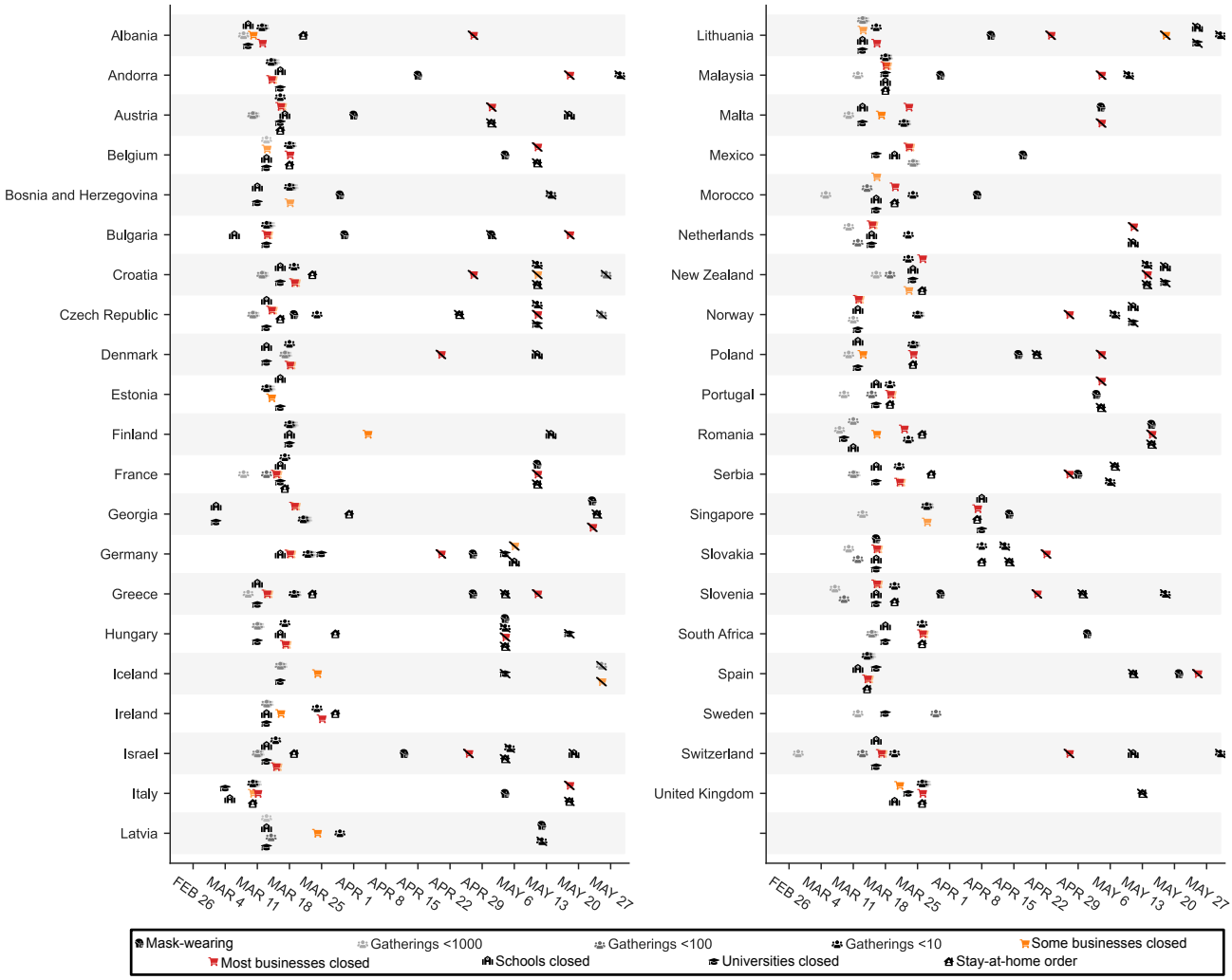

Figure 1: Timing of NPI implementations in early 2020. Crossed-out symbols signify when an NPI was lifted. Detailed definitions of the NPIs are given in Table 1.

Methods

Dataset

We analyse the effects of NPIs (Table 1) in 41 countries[a] (see Figure 1). We recorded NPI implementations when the measures were implemented nationally or in most regions of a country (affecting at least three fourths of the population). For each country, the window of analysis starts on the 22nd of January and ends after the first NPI was lifted, or on the 30th of May 2020, whichever was earlier. The reason to end the analysis after the first major reopening[b] was to avoid a distribution shift. For example, when schools reopened, it was often with safety measures, such as smaller class sizes and distancing rules. It is therefore expected that contact patterns in schools will have been different before school closure compared to after reopening. Modelling this difference explicitly is left for future work. Data on confirmed COVID-19 cases and deaths were taken from the Johns Hopkins CSSE COVID-19 Dataset.[9] The data used in this study, including sources, are available online here.

[a]The countries were selected for their number of cumulative cases surpassing a minimum threshold (at the time of modelling), the availability of reliable data on NPIs, and the trustworthiness of their reporting of deaths. Finally, we excluded very large countries like China, the US, and Canada, for ease of data collection, as these would require more locally fine-grained data.

[b]Concretely, the window of analysis extended until three days after the first reopening for confirmed cases, and 13 days after the first reopening for deaths. These values correspond to the 5% quantile of the infection-to-confirmation/death distributions, ensuring that less than 5% of the new infections on the reopening day were still observed in the window of analysis.

Table 1: NPIs included in the study. Appendix G details how edge cases in the data collection were handled.

NPI	Description
Mask-wearing mandatory in (some) public spaces	A country has mandated mask usage in the public, sometimes limited to just some public spaces (which the government deems to have a high risk of infection). For example, some countries mandated mask-wearing in most or all indoor public spaces but not outdoors.
Gatherings limited to 1000 people or less	A country has set a size limit on gatherings. The limit is at most 1000 people (often less), and gatherings above the maximum size are disallowed. For example, a ban on gatherings of 500 people or more would be classified as "gatherings limited to 1000 or less", but a ban on gatherings of 2000 people or more would not.
Gatherings limited to 100 people or less	A country has set a size limit on gatherings. The limit is at most 100 people (often less).
Gatherings limited to 10 people or less	A country has set a size limit on gatherings. The limit is at most 10 people (often less).
Some businesses closed	A country has specified a few kinds of customer-facing businesses that are considered "high risk" and need to suspend operations (blacklist). Common examples are restaurants, bars, nightclubs, cinemas, and gyms. By default, businesses are not suspended.
Most nonessential businesses closed	A country has suspended the operations of many customer-facing businesses. By default, customer-facing businesses are suspended unless they are designated as essential (whitelist).
Schools closed	A country has closed most or all schools.
Universities closed	A country has closed most or all universities and higher education facilities.
Stay-at-home order (with exemptions)	An order for the general public to stay at home has been issued. This is mandatory, not just a recommendation. Exemptions are usually granted for certain purposes (such as shopping, exercise, or going to work), or, more rarely, for certain times of the day. In practice, a stay-at-home order was often accompanied by other NPIs such as businesses closures. However, a stay-at-home order does not in principle entail these other NPIs, but only the (additional) order to generally stay at home except for exemptions.

Data collection

We collected data on the start and end date of NPI implementations, from the start of the pandemic until the 30th of May 2020. Before collecting the data, we experimented with several public NPI datasets, finding that they were not complete enough for our modelling and contained incorrect dates.[c] By focusing on a smaller set of countries and NPIs than these datasets, we were able to enforce strong quality controls: We used independent double entry and manually compared our data to public datasets for cross-checking.

First, two authors independently researched each country and entered the NPI data into separate spreadsheets. The researchers manually researched the dates using internet searches: there was no automatic component in the data gathering process. The average time spent researching each country per researcher was 1.5 hours.

Second, the researchers independently compared their entries to the following public datasets and, if there were conflicts, visited all primary sources to resolve the conflict: the EFGNPI database,[10] the Oxford COVID-19 Government Response Tracker,[11] and the mask4all dataset.[12]

Third, each country and NPI was again independently entered by one to three paid contractors, who were provided with a detailed description of the NPIs and asked to include primary sources with their data. A researcher then resolved any conflicts between this data and one (but not both) of the spreadsheets.

Finally, the two independent spreadsheets were combined and all conflicts resolved by a researcher. The final dataset contains primary sources (government websites and/or media articles) for each entry.

Data Preprocessing

Data on cases and deaths are noisy. Many countries preferentially report deaths and cases on certain days of the week. For example, there are days with zero newly confirmed cases even though there had been several hundred reported cases on the previous day. We therefore smooth the data using a moving average (on a linear scale) over 2 days into the past and future. When the case count is small, a large fraction of cases may be imported from other

[c] We evaluated the following datasets:

- Epidemic Forecasting Global NPI Database[10]
- Oxford COVID-19 Government Response Tracker (OxCGRT)[11]
- ACAPS #COVID19 Government Measures Dataset

Note that these datasets are under continuous development. Many of the mistakes found will already have been corrected. We know from our own experience that data collection can be very challenging. We have the fullest respect for the people behind these datasets. In this paper, we focus on a more limited set of countries and NPIs than these datasets contain, allowing us to ensure higher data quality in this subset. Given our experience with public datasets and our data collection, we encourage fellow COVID-19 researchers to independently verify the quality of public data they use, if feasible.

countries and the testing regime may change rapidly. To prevent this from biasing our model, we neglect case numbers before a country has reached 100 confirmed cases and death numbers before a country has reached 10 deaths. We include all preprocessing steps in our sensitivity analysis (Appendix C.2).

Model

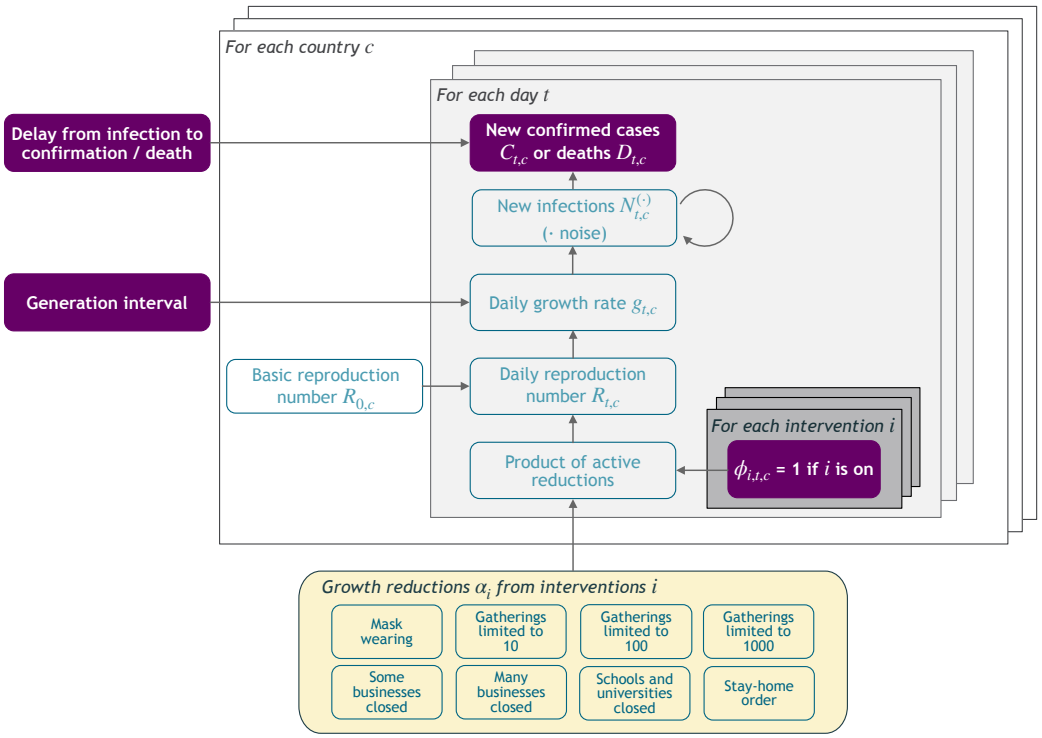

Figure 2: Model Overview. Purple nodes are observed or have a fixed distribution. From bottom to top: The effectiveness of intervention i is represented by α_i. On each day t, a country's daily reproduction number $R_{t,c}$ depends on the country's basic reproduction number $R_{0,c}$ and the active NPIs. The active NPIs are encoded by $\Phi_{i,t,c}$, which is 1 if NPI i is active in country c at time t, and 0 otherwise. $R_{t,c}$ is transformed into the daily growth rate $g_{t,c}$, which also depends on the generation interval. The growth rate is used to compute the new infections $N^{(C)}_{t,c}$ and $N^{(D)}_{t,c}$ that will later be registered as confirmed cases $C_{t,c}$ and deaths $D_{t,c}$ respectively, after a delay. Our model uses both death and case data: it splits all nodes above the daily growth rate $g_{t,c}$ into separate branches for deaths and confirmed cases.

Our model uses case and death data from each country to 'backwards' infer the number of new infections at each point in time, which is itself used to infer the reproduction numbers. NPI effects are then estimated by relating the daily reproduction numbers to the active NPIs, across all days and countries. This relatively simple, 'data-driven' approach allows us to sidestep assumptions about contact patterns and intensity, infectiousness of different age groups, and so forth, that are typically required in modelling studies. Our semi-mechanistic Bayesian hierarchical model is based on that of Flaxman et al.,[1] extended to use both case and death data. This increases the amount of data from which we can extract NPI effects, reduces distinct biases of case and death reporting, and reduces the bias of only including

countries that have many deaths. Additionally, as we do not aim to infer the total number of COVID-19 infections, we do not assume a specific infection fatality rate (IFR) or ascertainment rate (rate of testing). We proceed by summarising the model (Figure 2). A detailed description is given in Appendix A. Code is available online here.

The growth of the epidemic is determined by the time- and country-specific reproduction number $R_{t,c}$, which depends on: a) the (unobserved) basic reproduction number $R_{0,c}$ given no active NPIs and b) the active NPIs at time t. $R_{0,c}$ accounts for all time-invariant factors that affect transmission in country c, such as differences in demographics, population density, culture, and health systems.[13] We assume that the effect of each NPI on $R_{t,c}$ is stable across countries and time. The effectiveness of NPI i is represented by a parameter α_i, over which we place a symmetric prior with mean zero, allowing both positive and negative effects. Following Flaxman et al. and others,[1–3] each NPI's effect on $R_{t,c}$ is assumed to independently affect $R_{t,c}$ as a multiplicative factor:

$$R_{t,c} = R_{0,c} \prod_{i=1}^{I} \exp\left(-\alpha_i \phi_{i,t,c}\right), \tag{1}$$

where $\phi_{i,c,t} = 1$ indicates that NPI i is active in country c on day t ($\phi_{i,c,t} = 0$ otherwise), and I is the number of NPIs. The multiplicative effect encodes the plausible assumption that NPIs have a smaller absolute effect when $R_{t,c}$ is already low. We discuss the meaning of effectiveness estimates given NPI interactions in the Results section.

In the early phase of an epidemic, the number of new daily infections grows exponentially. During exponential growth, there is a one-to-one correspondence between the daily growth rate and $R_{t,c}$.[14] The correspondence depends on the generation interval (the time between successive infections in a chain of transmission), which we assume to have a Gamma distribution with mean 6.67 days.[1,15,16] We model the daily new infection count separately for confirmed cases and deaths, representing those infections which are later reported and those which are later fatal. However, both infection numbers are assumed to grow at the same daily rate in expectation, allowing the use of both data sources to estimate each α_i. The infection numbers translate into reported confirmed cases and deaths after a stochastic delay, which is assumed to be equal across countries. The delay is the sum of two independent gamma distributions, assumed to be equal across countries: the incubation period and the delay from onset of symptoms to confirmation. We use previously published empirical distributions from China and Italy,[16–19] which mutually agree, and give a mean infection-to-confirmation delay of 10.35 days. Similarly, the infection-to-death delay is the sum of the incubation period and the (gamma distributed) delay from onset of symptoms to death,[17,20] which sum up to a mean delay of 22.9 days. Finally, both the reported deaths and cases follow a negative binomial noise distribution an inferred dispersion parameter, as in related NPI models.[1,3]

Using a Markov chain Monte Carlo (MCMC) sampling algorithm,[21] this model infers posterior distributions of each NPI's effectiveness while accounting for cross-country variations in testing, reporting, and fatality rates. However, it relies on the key assumptions that NPIs

have equal effects across countries and time, and that changes in R are due to the observed NPIs. To analyse the extent to which modelling choices affect the results, our sensitivity analysis includes all epidemiological parameters, prior distributions, and many of the structural assumptions introduced above (Appendix B.2 and Appendix C). MCMC convergence statistics are given in Appendix C.

Results

NPI Effectiveness

Our model enables us to estimate the individual effectiveness of each NPI, expressed as a *percentage reduction in R*. As in related work,[1–3] this percentage reduction is modelled as constant over countries and time, and independent of the other implemented NPIs. In practice, however, NPI effectiveness may depend on other implemented NPIs and local circumstances. Thus, our effectiveness estimates ought to be interpreted as the *effectiveness averaged over the contexts in which the NPI was implemented, in our data*.[5] Our results thus give the average NPI effectiveness across typical situations that the NPIs were implemented in. Figure 3 (bottom left) visualizes which NPIs typically co-occurred, aiding interpretation.

The mean percentage reduction in R (with 95% credible interval) associated with each NPI is as follows (Figure 3): mandating mask-wearing in (some) public spaces: 2% (-14%–16%), limiting gatherings to 1000 people or less: 2% (-20%–22%), to 100 people or less: 21% (1%–39%), to 10 people or less: 36% (16%–53%), closing some high-risk businesses: 31% (13%–46%), closing most nonessential businesses: 40% (22%–55%), closing schools and universities: 39% (21%–55%), and issuing stay-at-home orders: 18% (4%–31%).

Some NPIs frequently co-occur, i.e., are *collinear*. However, we are able to isolate the effects of individual NPIs since the collinearity is imperfect and our dataset is large. For every pair of NPIs, we observe one of them without the other for 748 country-days on average (Appendix D.3). The minimum number of country-days for any NPI pair is 143 (for limiting gatherings to 1000 or 100 attendees). Additionally, under excessive collinearity, and insufficient data to overcome it, individual effectiveness estimates are highly sensitive to variations in the data and model parameters.[22] High sensitivity prevented Flaxman et al.,[1] who had a smaller dataset, from disentangling NPI effects.[6] Our estimates are substantially less sensitive (see below). Finally, the posterior correlations between the effectiveness estimates are weak, suggesting manageable collinearity (Appendix D.4).

Although the correlations between the individual estimates are weak, we should take them into account when evaluating combined effects of NPIs. For example, if two NPIs frequently co-occur, there may be more certainty about the combined effect than about the two individual effects. Figure 4 shows the combined effectiveness of the sets of NPIs that are most common in our data. All NPIs together reduce R by 82% (79%–85%). Across our countries, the mean R without any NPIs (i.e. R_0) is 3.8, matching the mean result of Flaxman et al.[1] (Table D.2 reports R_0 for all countries) Starting from this number, the esti-

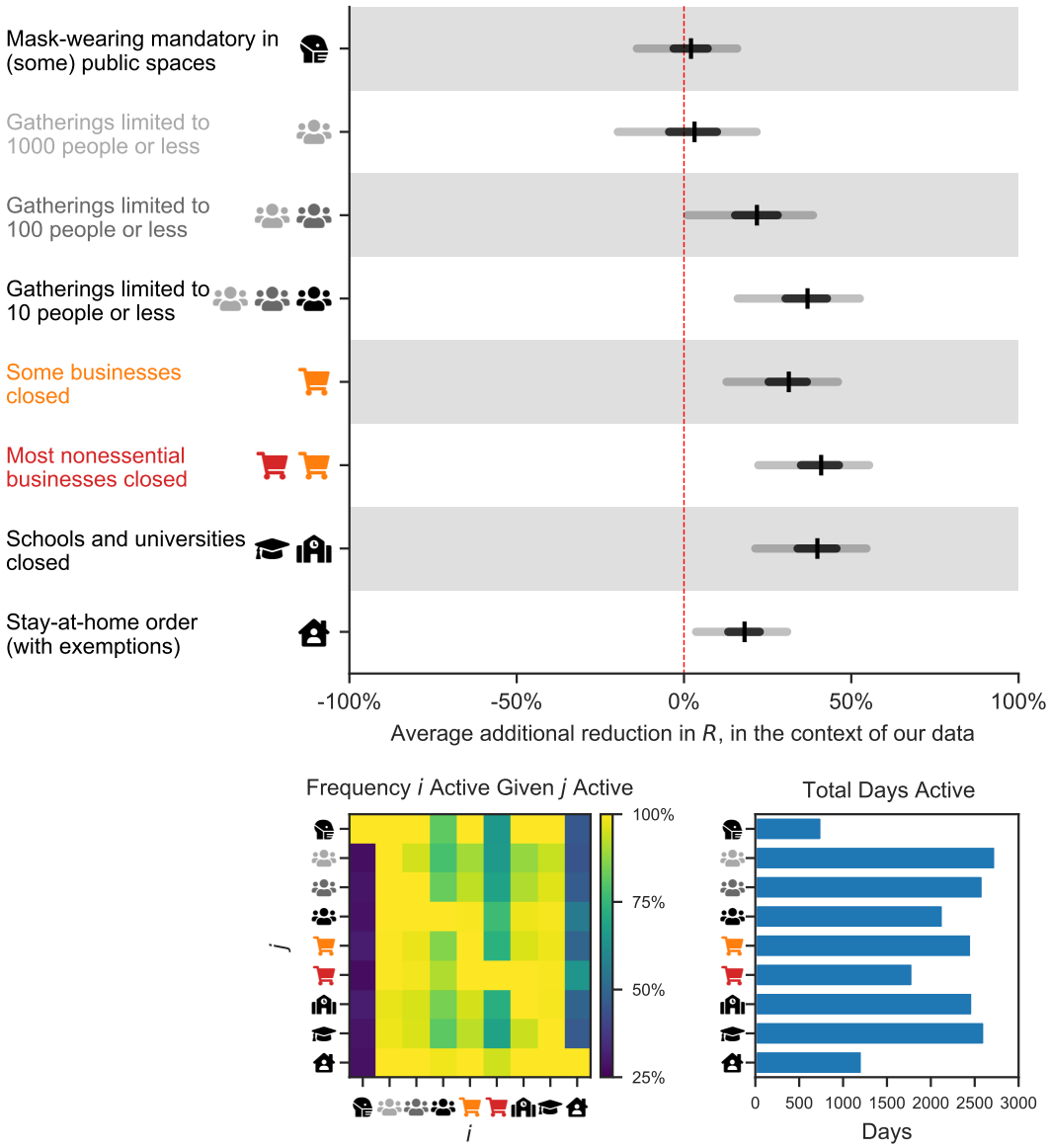

Figure 3: ***Top:*** **NPI effects. The Figure shows the average percentage reductions in R as observed in our data (or, in terms of the model, the posterior marginal distributions of $1-\exp(-\alpha_i)$), with median, 50% and 95% credible intervals. A negative 1% reduction refers to a 1% increase in R. Cumulative effects are shown for hierarchical NPIs (gathering bans and business closures) i.e., the result for *Most nonessential businesses closed* shows the cumulative effect of two NPIs with separate parameters and symbols - closing some (high-risk) businesses, and additionally closing most remaining (non-high-risk, but nonessential) businesses given that some businesses are already closed. Finally, we show the joint effect of closing both schools and universities because the dates of school and university closures nearly perfectly coincide in our data and we cannot meaningfully isolate their individual effects (Appendix D.2).** ***Bottom Left:*** **Conditional activation matrix. Cell values indicate the frequency that NPI i (x-axis) was active given that NPI j (y-axis) was active. E.g., schools were always closed whenever a stay-at-home order was active (bottom row, third column from the right), but not vice versa.** ***Bottom Right:*** **Total number of days each NPI was active across all countries.**

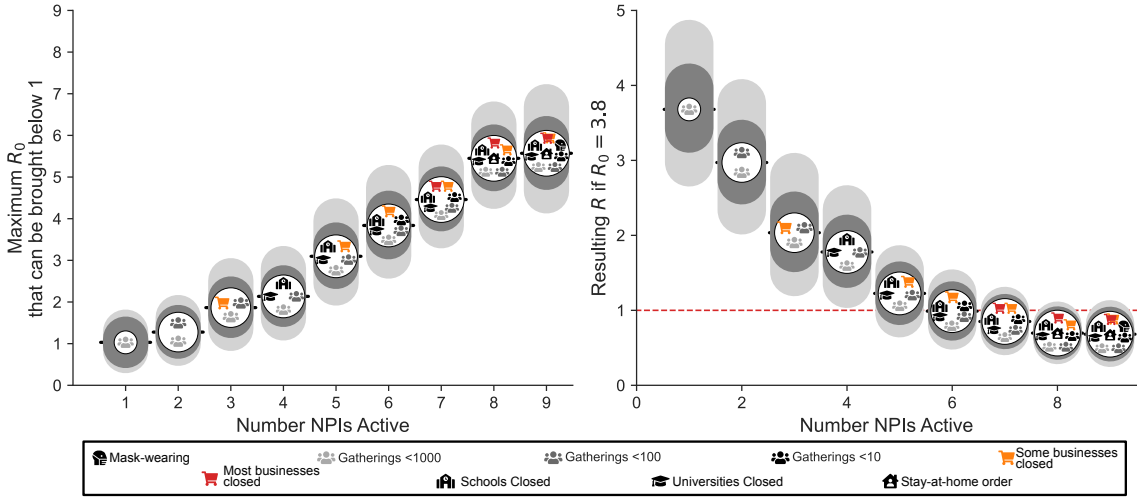

Figure 4: Combined NPI effectiveness for the most common sets of NPIs in our data, by size of the NPI set. Shaded regions denote 50% and 95% credible intervals. *Left*: Maximum R_0 that can be reduced to below 1 for each set of NPIs. *Right*: Predicted R after implementation of each set of NPIs, assuming $R_0 = 3.8$. Readers can interactively explore the effects of all sets of NPIs at http://epidemicforecasting.org/calc.

mated R can be reduced below 1 by closing schools and universities, high-risk businesses, and limiting gathering sizes. Readers can interactively explore the effects of sets of NPIs at http://epidemicforecasting.org/calc. A CSV file containing the joint effectiveness of all possible NPI combinations is available online here.

Validation

We perform a range of experiments to study the robustness and calibration of our NPI effectiveness estimates (Appendix B, with further experiments in Appendix C). We analyse how the model extrapolates to unseen countries and periods, and perform multiple sensitivity analyses. Amongst other things, we analyse how results change if we vary epidemiological parameters or vary the data (using only deaths *or* confirmed cases as observations; excluding countries from the data). To investigate our key assumptions, we show results for several alternative models (structural sensitivity), analyze the role of NPI timing, and examine possible confounding of our estimates by unobserved factors influencing R. Figure 5 summarises these analyses by showing how each NPI's effectiveness is ranked compared to other NPIs and how its rank is distributed across all experiment conditions. The strong agreement across these many analyses increases our confidence in the results while also showing that the precise effectiveness estimates come with additional uncertainty.

Discussion

We use a data-driven approach to estimate the effects of eight nonpharmaceutical interventions on COVID-19 transmission in 41 countries. All eight NPIs together reduce R by 82%

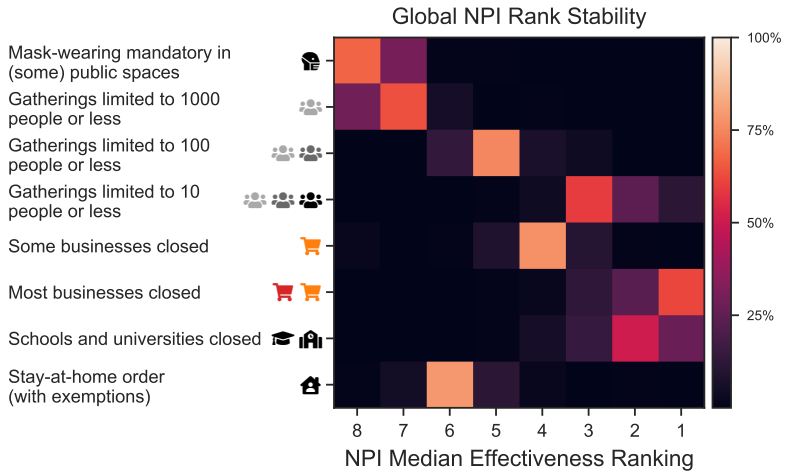

Figure 5: Ranking of NPIs by median effectiveness, across all sensitivity analyses (15 sensitivity analyses with a total of 96 experiment conditions). In each analysis, NPI effects were estimated under several different plausible variations of the model or the data (Appendix B and Appendix C). The colour indicates in which fraction of all experiment conditions an NPI occupied a given rank. The sensitivity analyses aggregated in this Figure are shown in Figures B.8, B.9, C.11, C.12, C.13, and B.7A.

(79%–85%). This finding is in strong agreement with the joint effect estimated in eleven countries by Flaxman et al.[1] and contributes to the mounting evidence that NPIs can be effective at mitigating and suppressing outbreaks of COVID-19. Furthermore, our results suggest that some NPIs outperform others. While the exact effectiveness estimates vary mildly, the qualitative conclusions discussed below are robust across 15 sensitivity analyses.

Business closures and gathering bans both seem effective at reducing COVID-19 transmission. Closing only high-risk businesses (mean reduction in R: 31%) appears only somewhat less effective than closing most nonessential businesses (40%), making it the more promising policy option in some circumstances. Limiting gatherings to 10 people or less (36%) was more effective than limits up to 100 (21%) or 1000 people (2%). This may reflect that small gatherings are common. As previously discussed, we estimate the average *additional* effect each NPI had in the contexts where it was implemented. When countries introduced stay-at-home orders, they nearly always also banned gatherings and closed schools, universities, and nonessential businesses if they had not done so already. Flaxman et al.[1] and Hsiang et al.[4] add the effect of these distinct NPIs to the effectiveness of stay-at-home orders, and accordingly find a large effect. In contrast, we and Banholzer et al.[3] isolate the *additional* effect of ordering the population to stay at home, and instead find a smaller effect (18%). A typical country can reduce R to below 1 without a stay-at-home order (Figure 4) provided other NPIs stay active.

Mandating mask-wearing in various public spaces had a small positive effect on average in the countries we studied (2%). This does not rule out that mask-wearing has a larger effect in other contexts. In our data, mask-wearing was only mandated when other NPIs had already reduced public interactions. When most transmission occurs in private spaces,

wearing masks in public is expected to be less effective. This might explain why a larger effect was found in studies that included China and South Korea, where mask-wearing was introduced earlier.[2,23] While there is an emerging body of literature indicating that mask-wearing can be effective in reducing transmission, the bulk of evidence comes from healthcare settings.[24] In non-healthcare settings, risk compensation[25] and open questions about different types of masks play a larger role, potentially reducing effectiveness. While our results cast doubt on reports that mask-wearing is the main determinant shaping a country's epidemic,[23] the policy still seems promising given all available evidence, due to its comparatively low economic and social costs. Its effectiveness may increase as other NPIs are lifted and public interactions return.

We find a surprisingly large effect for school and university closures: an average 39% reduction in R. This finding is remarkably robust across different model structures, variations in the data, and epidemiological assumptions (Figure B.7). It remains robust when controlling for NPIs excluded from our study and the onset time of major government intervention (Figures B.9 and C.10). Since school and university closures almost perfectly coincide in the countries we study, an approach such as ours cannot distinguish their individual effects (Appendix D.2). Furthermore, it cannot distinguish direct and indirect effects, such as forcing parents to stay at home or causing broader behaviour changes by increasing public concern.

Previous evidence on school and university closures is mixed.[1,3,26] Early data suggest that children and young adults are equally susceptible to infection but have a notably lower observed case rate than older adults—whether this is due to school and university closures remains unknown.[27–29] Although infected young people are often asymptomatic, they appear to shed similar amounts of virus as older people,[30,31] and might therefore circulate the infection to higher-risk demographics unknowingly. This also limits our ability to detect large outbreaks in educational facilities, which closed in nearly all countries before such detection was feasible (with exceptions[32]). As outbreaks detected in UK schools are rapidly increasing,[33] this topic merits further study.

Our study has several limitations. First, NPI effectiveness may depend on the context of implementation, such as the presence of other NPIs and country-specific factors. Our estimates must be interpreted as the average effectiveness over the contexts in our dataset,[5] and expert judgement is required to adjust them to local circumstances. Second, R may have been reduced by unobserved NPIs or spontaneous behaviour changes. To investigate whether these reductions could be falsely attributed to the observed NPIs, we perform several additional analyses and find that our results are stable to a range of unobserved effects (Appendix B.2). However, this cannot give final certainty. Investigating the role of unobserved effects is an important topic for future investigations. Third, our results cannot be used without qualification to predict the effect of *lifting* NPIs. For example, closing schools and universities seems to have greatly reduced transmission, but this does not mean that reopening them will cause infections to soar. Educational institutions can (and do) implement safety measures such as reduced class sizes. Further work is needed to analyse the effects of reopenings; our collected data may be instrumental. Fourth, while we included more

NPIs than previous work (Table F.4), several promising NPIs were excluded. For example, testing, tracing, and case isolation may be an important part of a cost-effective epidemic response,[34] but were not included because it is difficult to obtain comprehensive data. We discuss further limitations in Appendix E.

Currently, governments across the world are seeking to keep R below 1 while minimising the social and economic costs of their interventions. We hope that our results can guide policy decisions on which restrictions to lift, and which NPIs to implement in any potential second wave of infections. Additionally, our results show which areas of public life are most in need of restructuring, so that they can continue despite the pandemic. However, our results should not be seen as the final answer on NPI effectiveness, but rather as a contribution to a diverse body of evidence, alongside other retrospective studies, experimental trials, simulations, and clinical experience.

Acknowledgements

Jan Brauner was supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems [EP/S024050/1] and by Cancer Research UK. Sören Mindermann's funding for graduate studies was from Oxford University and DeepMind. Mrinank Sharma was supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems [EP/S024050/1]. Gavin Leech was supported by the UKRI Centre for Doctoral Training in Interactive Artificial Intelligence [EP/S022937/1].

The paid contractor work in the data collection and the development of the interactive website was funded by the Berkeley Existential Risk Initiative.

We thank Jacob Lagerros for operational support and for introducing some of the authors to each other. We thank Maksym Balatsko, Marek Pukaj, and Tomáš Witzany for developing the interactive website.

Declarations of interest

No conflicts of interests.

Authors' contributions

D Johnston, JM Brauner, J Kulveit, G Altman, AJ Norman, JT Monrad, G Leech, V Mikulik designed and conducted the NPI data collection. S Mindermann, M Sharma, JM Brauner, AB Stephenson, H Ge, YW Teh, Y Gal, J Kulveit, T Gavenciak, J Salvatier, MA Hartwick, L Chindelevitch designed the model and modelling experiments. M Sharma, AB Stephenson, T Gavenciak, J Salvatier performed and analysed the modelling experiments. J Kulveit, T Gavenciak, JM Brauner conceived the research. S Mindermann, T Besiroglu, J Kulveit, JM Brauner did the literature search. JM Brauner, S Mindermann, M Sharma, G Leech, T Besiroglu, V Mikulik wrote the manuscript. All authors read and gave feedback on the manuscript and approved the final manuscript. JM Brauner, S Mindermann, and M Sharma contributed equally. Y Gal and J Kulveit contributed equally to senior authorship.

Keywords

COVID-19, SARS-CoV-2, nonpharmaceutical intervention, countermeasure, Bayesian model

References

1 Seth Flaxman et al. "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe". In: *Nature* (2020), pp. 1–8.

2 Xiaohui Chen and Ziyi Qiu. "Scenario analysis of non-pharmaceutical interventions on global COVID-19 transmissions". https://arxiv.org/abs/2004.04529. Apr. 7, 2020.

3 Nicolas Banholzer et al. "Impact of non-pharmaceutical interventions on documented cases of COVID-19". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (Apr. 2020). DOI: 10.1101/2020.04.16.20062141. URL: https://www.medrxiv.org/content/10.1101/2020.04.16.20062141v3.

4 Solomon Hsiang et al. "The Effect of Large-Scale Anti-Contagion Policies on the Coronavirus (COVID-19) Pandemic". In: *medRxiv* (May 2020), p. 2020.03.22.20040642. DOI: 10.1101/2020.03.22.20040642.

5 Mrinank Sharma et al. "On the Robustness of Effectiveness Estimation of Nonpharmaceutical Interventions Against COVID-19 Transmission". In: *Arxiv* (2020).

6 Kristian Soltesz et al. "On the sensitivity of non-pharmaceutical intervention models for SARS-CoV-2 spread estimation". In: *medRxiv* (2020).

7 Cindy Cheng et al. "COVID-19 Government Response Event Dataset (CoronaNet v. 1.0)". In: *Nature Human Behaviour* (2020), pp. 1–13.

8 Oxford Covid Government Response Tracker. July 2020. URL: https://github.com/OxCGRT/covid-policy-tracker.

9 Johns Hopkins University Center for Systems Science and Engineering. *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. https://github.com/CSSEGISandData/COVID-19. 2020.

10 *Epidemic Forecasting Global NPI Database*. http://epidemicforecasting.org/datasets. 2020.

11 Thomas Hale et al. *Oxford COVID-19 Government Response Tracker*. Blavatnik School of Government. https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker. 2020.

12 #Mask4All. *What Countries Require Masks in Public or Recommend Masks?* https://masks4all.co/what-countries-require-masks-in-public/. (Accessed on 05/24/2020).

13 Suryakant Yadav and Pawan Kumar Yadav. "Basic Reproduction Rate and Case Fatality Rate of COVID-19: Application of Meta-analysis". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (May 2020). DOI: 10.1101/2020.05.13.20100750. URL: https://www.medrxiv.org/content/10.1101/2020.05.13.20100750v1.

14 J Wallinga and M Lipsitch. "How generation intervals shape the relationship between growth rates and reproductive numbers". In: *Proceedings of the Royal Society B: Biological Sciences* 274.1609 (Nov. 2006), pp. 599–604. DOI: 10.1098/rspb.2006.3754.

15 John M Griffin et al. "A rapid review of available evidence on the serial interval and generation time of COVID-19". In: *medRxiv* (2020).

16 D Cereda et al. "The early phase of the COVID-19 outbreak in Lombardy, Italy". In: (Mar. 20, 2020). arXiv: 2003.09320v1 [q-bio.PE]. URL: https://arxiv.org/abs/2003.09320.

17 Natalie M Linton et al. "Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data". In: *Journal of clinical medicine* 9.2 (2020), p. 538.

18 Qun Li et al. "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia". In: *New England Journal of Medicine* 382.13 (Mar. 2020), pp. 1199–1207. DOI: 10.1056/nejmoa2001316.

19 Qifang Bi et al. "Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (Mar. 2020). DOI: 10.1101/2020.03.03.20028423. URL: https://www.medrxiv.org/content/10.1101/2020.03.03.20028423v3.

20 Robert Verity et al. "Estimates of the severity of coronavirus disease 2019: a model-based analysis". In: *The Lancet Infectious Diseases* (Mar. 2020). DOI: 10.1016/s1473-3099(20)30243-7.

21 Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.47 (2014), pp. 1593–1623. URL: http://jmlr.org/papers/v15/hoffman14a.html.

22 Christopher Winship and Bruce Western. "Multicollinearity and model misspecification". In: *Sociological Science* 3.27 (2016), pp. 627–649.

23 Renyi Zhang et al. "Identifying airborne transmission as the dominant route for the spread of COVID-19". In: *Proceedings of the National Academy of Sciences* 117.26 (June 2020), pp. 14857–14863. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2009637117.

24 Derek K Chu et al. "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis". In: *The Lancet* (2020).

25 Graham P Martin, Esmée Hanna, and Robert Dingwall. "Urgency and uncertainty: covid-19, face masks, and evidence informed policy". In: *BMJ* 369 (2020).

26 Juanjuan Zhang et al. "Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China". In: *Science* (2020).

27 Nisha S Mehta et al. "SARS-CoV-2 (COVID-19): What do we know about children? A systematic review". In: *Clinical Infectious Diseases* (May 2020). DOI: 10.1093/cid/ciaa556.

28 Petra Zimmermann and Nigel Curtis. "Coronavirus Infections in Children Including COVID-19". In: *The Pediatric Infectious Disease Journal* 39.5 (May 2020), pp. 355–368. DOI: 10.1097/inf.0000000000002660.

29 *When Should a School Reopen? Final Report*. http://www.independentsage.org/wp-content/uploads/2020/05/Independent-Sage-Brief-Report-on-Schools-5.pdf. (Accessed on 05/28/2020). May 2020.

30 Terry C. Jones et al. “An analysis of SARS-CoV-2 viral load by patient age”. 2020.

31 Arnaud G L’Huillier et al. “Shedding of infectious SARS-CoV-2 in symptomatic neonates, children and adolescents”. In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (May 2020). DOI: 10.1101/2020.04.27.20076778.

32 Arnaud Fontanet et al. “Cluster of COVID-19 in northern France: A retrospective closed cohort study”. In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (Apr. 2020). DOI: 10.1101/2020.04.18.20071134.

33 *Weekly Coronavirus Disease 2019 (COVID-19) Surveillance Report - week 26*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/895356/Weekly_COVID19_Surveillance_Report_w26.pdf. 2020.

34 Tim Colbourn et al. *Modelling the Health and Economic Impacts of Population-Wide Testing, Contact Tracing and Isolation (PTTI) Strategies for COVID-19 in the UK*. ID 3627273. June 2020. DOI: 10.2139/ssrn.3627273. URL: https://papers.ssrn.com/abstract=3627273.

Appendix

Table of Contents

Appendix A. Modelling details

Appendix A.1. Detailed model description

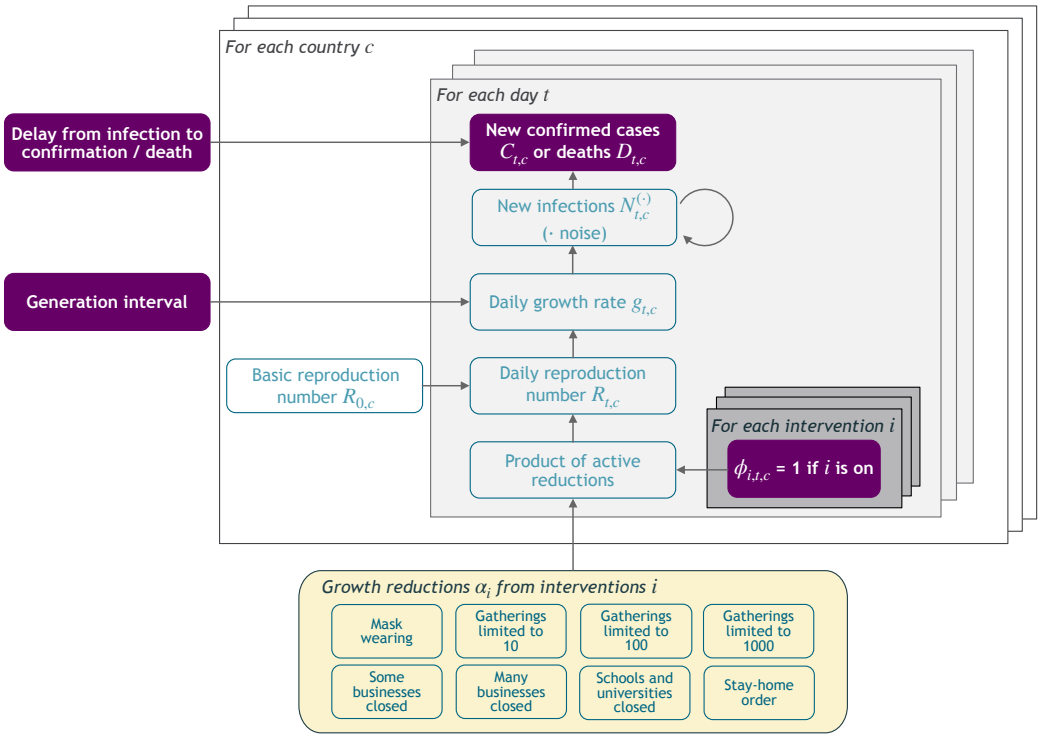

Figure A.6: Model Overview. Purple nodes are observed or have a fixed distribution. We describe the diagram from bottom to top: Each NPI's effectiveness is characterised by α_i, which is independent of the country. On each day, a country's daily reproduction number $R_{t,c}$ only depends on that country's base reproduction number $R_{0,c}$ and the active NPIs ($\Phi_{i,t,c}$). $R_{t,c}$ is transformed to the daily growth rate $g_{t,c}$, which is used to compute the new infections $N^{(C)}_{t,c}$ and $N^{(D)}_{t,c}$ that will turn into confirmed cases and deaths, respectively. Finally, the number of new confirmed cases $C_{t,c}$ and deaths $D_{t,c}$ is computed by convolution of $N^{(\cdot)}_{t,c}$ with the respective delay distributions. The same model structure is used for confirmed cases and deaths. The model combines both observations; it splits all nodes above the daily growth rate $g_{t,c}$ into separate branches for deaths and cases.

We construct a semi-mechanistic Bayesian hierarchical model, similar to Flaxman et al.[1] The main difference is that we model both confirmed cases *and* deaths, allowing us to leverage significantly more data. Furthermore, we do not assume a specific infection fatality rate (IFR) since we do not aim to infer the *total* number of COVID-19 infections. The end of this section details further adaptations which allow us to minimize assumptions about testing, reporting, and the IFR. A list of all technical details is given in Appendix A.2.

We describe the model in Figure A.6 from bottom to top. The epidemic's growth is determined by the time-and-country-specific (instantaneous) reproduction number $R_{t,c}$. It depends on: a) the basic reproduction number $R_{0,c}$ without any NPIs active and b) the active NPIs. We place a prior (and hyperprior) distribution over $R_{0,c}$, reflecting the wide disagreement of regional estimates of R_0.[2] We parameterize the effectiveness of NPI i, assumed to

be same across countries and time, with α_i. Each NPI is assumed to have an independent multiplicative as on $R_{t,c}$ as follows:

$$R_{t,c} = R_{0,c} \prod_{i=1}^{I} \exp\left(-\alpha_i \phi_{i,t,c}\right), \tag{A.1}$$

where $\phi_{i,c,t} = 1$ means NPI i is active in country c on day t ($\phi_{i,c,t} = 0$ otherwise), and I is the number of NPIs. We use a weakly informative symmetric prior $\alpha_i \sim \mathcal{N}(0, 0.2)$, allowing for both positive and negative effects, because we presently cannot rule out that some NPIs directly or indirectly increase transmission.

Growth rates. $N_{t,c}$ denotes the number of new infections at time t and country c. In the early phase of an epidemic, $N_{t,c}$ grows exponentially with a daily[a] growth rate $g_{t,c}$. During exponential growth, there is a well-known one-to-one correspondence between $g_{t,c}$ and $R_{t,c}$:[3]

$$R_{t,c} = \frac{1}{M(-\log(1 + g_{t,c}))}, \tag{A.2}$$

where $M(\cdot)$ is the moment-generating function of the distribution of the generation interval (the time between successive cases in a transmission chain). We assume that the generation interval distribution is given by a gamma distribution with mean 6.67 days and standard deviation 2.1. The mean is based on an Italian study,[4] which is deemed most relevant to European countries,[5] and the standard deviation stems from a international set of countries since in European countries it has only been estimated for the *serial* rather than generation interval to our knowledge.[6,7] Using (A.2), we can write $g_{t,c}$ as $g_{t,c}(R_{t,c})$ (see Appendix A.2).

Infection model. Rather than modelling the total number of new infections $N_{t,c}$, we model new infections that will either be subsequently a) confirmed positive, $N^{(C)}_{t,c}$, or b) lead to a reported death, $N^{(D)}_{t,c}$. These are backwards-inferred from the observation models for cases and deaths, shown further below. We assume that both grow at the same expected rate $g_{t,c}$:

$$N^{(C)}_{t,c} = N^{(C)}_{0,c} \prod_{\tau=1}^{t} \left[(1 + g_{\tau,c}) \cdot \exp\left(\epsilon^{(C)}_{\tau,c}\right)\right] \tag{A.3}$$

$$N^{(D)}_{t,c} = N^{(D)}_{0,c} \prod_{\tau=1}^{t} \left[(1 + g_{\tau,c}) \cdot \exp\left(\epsilon^{(D)}_{\tau,c}\right)\right] \tag{A.4}$$

[a]Many epidemiological models define growth rates as the exponent r in an exponential growth function. Here, we use daily growth rates instead for ease of exposition. These choices are mathematically equivalent. Note that we adapted equation (2.9) in Wallinga & Lipsitch[3] to account for our choice.

where $\epsilon^{(\cdot)}_{\tau,c} \sim \mathcal{N}(0, \sigma_N = 0.2)$ are separate, independent noise terms. Noise on the infection numbers is not used by Flaxman et al.[1] but has a history in epidemic modelling.[8] Empirically, we find that it leads to substantially more robust effectiveness estimates.[9]

We select σ_N by cross validation as no reference is available for it. We did not tune any other aspect of the model—instead, we use choices from Flaxman et al.[1] or the most relevant available sources. We evaluate five different values ($\sigma_N \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$), fitting the model on 35 countries each time and evaluating on a fixed, randomly chosen validation set of 6 countries. We select $\sigma_N = 0.2$ to maximize the log-likelihood on the validation set. Cross validation ensures a more calibrated model, less likely to produce overconfident and unstable estimates.[10] The final model with $\sigma_N = 0.2$ is then evaluated on data from 20 held-out days in all countries which were not used to adjust any aspect of the model (Figure B.7). We find that different values for σ_N produce very similar effectiveness estimates but larger values lead to greater uncertainty and robustness (Figure C.12 and Sharma et al.[9]).

We seed our model with unobserved initial values, $N^{(C)}_{0,c}$ and $N^{(D)}_{0,c}$, which have uninformative priors.[b]

Observation model for confirmed cases. The mean predicted number of new confirmed cases is a discrete convolution

$$\overline{C}_{t,c} = \sum_{\tau=1}^{t} N^{(C)}_{t-\tau,c} P_C(\text{delay} = \tau) \tag{A.5}$$

where $P_C(\text{delay})$ is the distribution of the delay from infection to confirmation. This delay distribution is the sum of two independent gamma distributions: the incubation period and the delay from onset of symptoms to confirmation. We use previously published and agreeing empirical distributions from China and Italy,[4,11–13] which sum up to a mean delay of 10.35 days. Finally, the observed cases $C_{t,c}$ follow a negative binomial noise distribution with mean $\overline{C}_{t,c}$ and an inferred dispersion parameter, following Flaxman et al.[1]

Observation model for deaths. The mean predicted number of new deaths is a discrete convolution

$$\overline{D}_{t,c} = \sum_{\tau=1}^{t} N^{(D)}_{t-\tau,c} P_D(\text{delay} = \tau),$$

where $P_D(\text{delay})$ is the distribution of the delay from infection to death. It is also the sum of two independent gamma distributions: the aforementioned incubation period and the delay from onset of symptoms to death,[11,14] which sum up to a mean delay of 22.9 days. Finally, the observed deaths $D_{t,c}$ also follow a negative binomial distribution with mean $\overline{D}_{t,c}$ and the same inferred dispersion parameter used for observed cases.

[b] Since we treat new infections as a continuous number, its initial value can (and often should) be between 0 and 1.

Single and combined models. To construct models which only use either confirmed cases or deaths as observations, we remove the variables corresponding to the disregarded observations.

Testing, reporting, and infection fatality rates. Scaling all values of a time series by a constant does not change its growth rates. The model is therefore invariant to the scale of the observations and consequently to country-level differences in the IFR and the ascertainment rate (the proportion of infected people who are subsequently reported positive). For example, assume countries A and B differ *only* in their ascertainment rates. Then, our model will infer a difference in $N_{t,c}^{(C)}$ (Eq. (A.5)) but *not* in the growth rates $g_{t,c}$ across A and B (Eq. (A.3)-(A.4)). Accordingly, the inferred NPI effectiveness will be identical.[c]

In reality, a country's ascertainment rate (and IFR) can also change *over time*. In principle, it is possible to distinguish changes in the ascertainment rate from the NPIs' effects: decreasing the ascertainment rate decreases future cases $C_{t,c}$ by a constant factor whereas the introduction of an NPI decreases them by a factor that grows exponentially over time.[d] The noise term, $\exp\left(\epsilon_{\tau,c}^{(C)}\right)$ (Eq. (A.3)), mimic changes in the ascertainment rate—noise at time τ affects all future cases—and allows for gradual, multiplicative changes in the ascertainment rate.

We infer the unobserved variables in our model using Hamiltonian Monte-Carlo[15,16] (HMC), a standard MCMC sampling algorithm.

Appendix A.2. Technical Model Description

Variables are indexed by intervention i, country c, and day t. All prior distributions are independent.

- **Data**
  1. **NPI Activations**: $\phi_{i,t,c} \in \{0, 1\}$.
  2. **Smoothed Observed Cases**: $C_{t,c}$.
  3. **Smoothed Observed Deaths**: $D_{t,c}$.
- **Prior Distributions**

[c] This is only approximately true. The negative binomial output distribution has a coefficient of variation diminishing with its mean; i.e., smaller observations are relatively more noisy and carry less weight. Furthermore, whilst the prior over $N_{0,c}^{(C)}$ could break scale invariance, the uninformative prior results in a negligible effect.

[d] However, our model may struggle when the ascertainment rate also changes exponentially over time. This could happen when a country reaches its testing capacity. See Appendix E.

1. **Country-specific R_0**

$$R_{0,c} = \texttt{Normal}(\bar{R}_0, \kappa) \tag{A.6}$$
$$\bar{R}_0 = 3.25, \text{ based on a meta analysis.}^{17} \tag{A.7}$$
$$\kappa \sim \texttt{Half Normal}(\mu = 0, \sigma = 0.5) \tag{A.8}$$

2. **NPI Effectiveness**

$$\alpha_i \sim \texttt{Normal}(\mu = 0, \sigma = \sqrt{0.2}) \tag{A.9}$$
$$\tag{A.10}$$

3. **Infection Initial Counts (uninformative priors)**

$$N_{0,c}^{(C)} = \exp(\zeta_c^{(C)}) \tag{A.11}$$
$$N_{0,c}^{(D)} = \exp(\zeta_c^{(D)}) \tag{A.12}$$
$$\zeta_c^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma = 50) \tag{A.13}$$
$$\zeta_c^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma = 50) \tag{A.14}$$
$$\tag{A.15}$$

4. **Observation Noise Dispersion Parameter**

$$\Psi \sim \texttt{Half Normal}(\mu = 0, \sigma = 5) \tag{A.16}$$

- **Hyperparameters**
  1. **Infection Noise Scale,** $\sigma_N = 0.2$ (selected by cross-validation).
- **Epidemiological parameters**
  1. **Generation Interval Parameters**. The generation interval is assumed to have a `Gamma` distribution with mean 6.67 and standard deviation 2.1 days.[4–6] This leads to a distribution Gamma($\alpha = 7.9, \beta = 1.2$).
  2. **Delay Distributions**. The time from infection to confirmation is assumed to be the sum of the incubation period and the time taken from symptom onset to laboratory confirmation. Therefore, the time taken from infection to confirmation, $\mathcal{T}^{(C)}$ is:[4,11–13]

$$\mathcal{T}^{(C)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Negative Binomial}(\mu = 5.25, \alpha = 1.57) \tag{A.17}$$

The time from infection to death is assumed to be the sum of the incubation period and the time taken from symptom onset to death. Therefore, the time taken from infection to death, $\mathcal{T}^{(D)}$ is:[1,11,14]

$$\mathcal{T}^{(D)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Gamma}(\mu = 17.8, \frac{\sigma}{\mu} = 0.45), \tag{A.18}$$

where α is known as the dispersion parameter. **Caution:** larger values of α correspond to a *smaller* variance, and less dispersion. With our parameterisation, the variance of the Negative Binomial distribution is $\mu + \frac{\mu^2}{\alpha}$.
For computational efficiency, we discretise this distribution using Monte Carlo sampling. We therefore form discrete arrays, $\pi_C[i]$ and $\pi_D[i]$ where the value of $\pi_C[i]$ corresponds to the probability of the delay being `i` days. We truncate π_C to a maximum delay of 31 days and π_D to a maximum delay of 63 days.

- **Infection Model**
  1. $R_{t,c} = R_{0,c} \cdot \exp\left(-\sum_{i=1}^{9} \alpha_i \ \phi_{i,t,c}\right)$.
  2. $g_{t,c} = \exp\left(\beta(R_{c,t}^{\frac{1}{\alpha}} - 1)\right) - 1$ where α and β are the parameters of the generation interval distribution. This is the exact conversion *under exponential growth*, following eq. (2.9) in Wallinga & Lipsitch.[3] (Note that we use daily growth rates.)
  3.

$$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(C)}\right], \tag{A.19}$$

$$N_{t,c}^{(D)} = N_{0,c}^{(D)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(D)})\right], \text{with noise} \tag{A.20}$$

$$\varepsilon_{\tau,c}^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma = \sigma_N), \tag{A.21}$$

$$\varepsilon_{\tau,c}^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma = \sigma_N) \tag{A.22}$$

$N_{t,c}^{(C)}$ represents the number of daily new infections at time t in country c who will eventually be tested positive ($N_{t,c}^{(D)}$ similar but for infections who will pass away).

- **Observation Model**: We use discrete convolutions to produce the expected number of new cases and deaths on a given day.

$$\bar{C}_{t,c} = \sum_{\tau=1}^{32} N_{t-\tau,c}^{(C)} \pi_C[\tau], \tag{A.23}$$

$$\bar{D}_{t,c} = \sum_{\tau=1}^{64} N_{t-\tau,c}^{(D)} \pi_D[\tau]. \tag{A.24}$$

Finally, the output distribution follows a Negative Binomial noise distribution as proposed by Flaxman et al.[1]

$$C_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{C}_{t,c}, \alpha = \Psi) \tag{A.25}$$

$$D_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{D}_{t,c}, \alpha = \Psi) \tag{A.26}$$

α is the dispersion parameter of the distribution. **Caution:** larger values of α correspond to a *smaller* variance, and less dispersion. With our parameterisation, the variance of the Negative Binomial distribution is $\mu + \frac{\mu^2}{\alpha}$, so that smaller observations are relatively more noisy.

This model was implemented in PyMC3[18] with the NUTS MCMC sampling algorithm.[16]

Appendix B. Validation

Appendix B.1. Unseen data

An important way to validate a Bayesian model is by checking how well it predicts unseen data, even if prediction is not the purpose of the model.[10,19] If an NPI effectiveness model is entirely unable to extrapolate to unseen countries, or to future unseen confirmed cases and deaths, we have strong reason to doubt its effectiveness estimates. However, we do not expect NPI effectiveness models to extrapolate perfectly. Almost always, there will be unobserved factors affecting the observed number of cases and deaths, such as changes in the ascertainment rate or IFR, spontaneous behaviour changes, and unobserved NPIs. Our models ought to treat these factors as noise and not attribute their effects on R to the observed NPIs.

We fit our model while holding out the last 20 days of cases and deaths for all countries, and then extrapolate to the last 20 days (Figure B.7, top left). A 20-day prediction is challenging; the longest attempted holdout period we found in data-driven NPI models was 3 days,[1] and most other related work does not validate predictions on unseen data at all.[9] The model is well-calibrated, with most points falling within the 95% credible intervals. The model predicts a higher number than reported exactly twice as often as predicting a lower number. This suggests that unobserved factors have reduced R below what would be predicted based on the active NPIs alone. We would indeed expect most countries to have fewer cases and deaths than predicted solely from the eight NPIs in our model. There are several other NPIs, as well as further unobserved behaviour changes, that we do not model but that likely reduce R on average. The result suggests that these factors are, at least to a certain extent, successfully treated as noise instead of confounding the effects of NPIs.

However, note that the model shown in Figure B.7 (top left) was fitted on 20 days less of data (per country) than the main model. The predictions can thus only serve for model validation insofar as we expect the model fitted on all days to have similar or better extrapolation to unseen data as the model fitted on all but the last 20 days. In further validation experiments, we analyse how the model extrapolates to individual countries left-out during fitting, and again find that it makes well-calibrated predictions (Appendix C.5).

Appendix B.2. Sensitivity Analysis

Sensitivity analysis reveals which results depend on uncertain parameters and modelling choices, and can diagnose model misspecification and excessive collinearity in the data.[20] We vary many of the components of our model and recompute the NPI effectiveness estimates, summarised here. Further analysis in Appendix C.

Sensitivity to epidemiological parameters. The epidemiological parameters in our model are the delay from infection to reporting, the delay from infection to death, and the generation interval. Furthermore, we specify a prior distribution over NPI effectiveness. In Fig-

ure B.7 (top right), we consider several alternative values for these parameters. Consistent with Flaxman et al.[1] and theoretical expectation,[3] we find that a shorter mean generation interval implies a smaller initial R_0 and therefore lower effectiveness estimates on average. However, the estimate for banning large gatherings increases, partly a consequence of including an unrealistically[13] short generation interval of 4 days. Restricting the prior to only allow NPIs to reduce, but not increase R, has no significant impact on the estimates ('Half Normal'). Using an uninformative prior (log-normal with $\mu = 1$ and $\sigma = 10$; 'Wide') amplifies differences between NPIs, suggesting that our default prior is informative. Using the prior of Flaxman et al. increases the differences between NPIs, an outcome which this prior encourages. We do not use this prior in our main analysis because it is designed to make realistic assumptions about the joint effect of all NPIs, not about their individual effects.

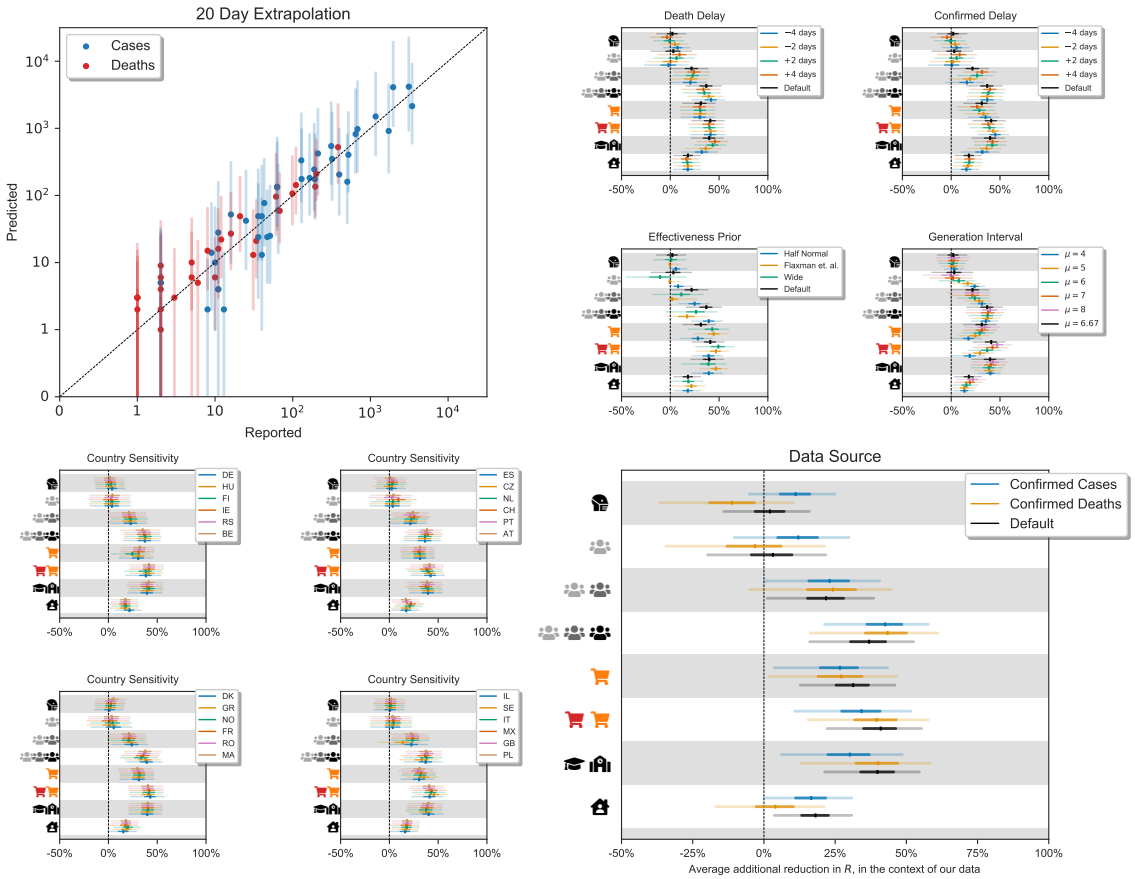

Figure B.7: Results validation. ***Top left:*** **We fit our model while holding out the last 20 days of cases and deaths for all countries. The figure shows the extrapolation to the last 20 days. Each dot represents predicted cases or deaths in one country. 95% sampled credible intervals shown. Observed cases and deaths are smoothed (see Methods).** ***Others:*** **NPI effectiveness estimates when epidemiological parameters or data are varied. Median, 50% and 95% credible intervals of the marginal posterior distribution of the effectiveness parameters are shown.** ***Top right:*** **Sensitivity to epidemiological parameter choices. Changes in the mean generation interval, delay distributions between infection and case confirmation/death, and the prior on NPI effectiveness.** ***Bottom left:*** **Sensitivity to removing one country at a time from the data.** ***Bottom right:*** **Sensitivity to different data sources: using only confirmed cases, only deaths, or both.**

Sensitivity to data. Figure B.7 (bottom right) shows the NPI effectiveness estimates from models that use only cases *or* deaths as observations, in contrast to our main model, which uses both. Reassuringly, the three models have similar results. This suggests that results are not biased by factors specific to deaths or confirmed cases, such as changes in the ascertainment rate, IFR, and model-specific time delays. Figure B.7 (bottom left) shows results if one country at a time is excluded from the data. As there is no strong justification for in- or excluding one particular country, results ought to be stable if a country is excluded. This is indeed the case. All countries are shown in Appendix C.

Sensitivity to structurally different models. A number of implicit structural assumptions are made in our model. We test sensitivity to these assumptions by evaluating NPI effectiveness estimates from alternative models, reproducing the structural sensitivity analysis from our concurrent work where these models are described in detail.[9]

As Figure B.8 shows, all models support the conclusions we draw in the Discussion. The models are:

1. *Different Effects Model*. The effectiveness of each NPI is allowed to vary across countries.
2. *Discrete Renewal Model*. Instead of converting R into a daily growth rate, a renewal process is used as the infection model, as in a number of earlier works.[1,8,21–23]
3. *Noisy-R Model*. The noise terms $\epsilon_{c,t}^{(\cdot)}$ affect R rather than the growth rate, as in Fraser[8]
4. *Additive Effects Model* Each NPI has an additive effect on R. The joint effectiveness of a set of NPIs is produced by summing, rather than multiplying, their individual effectiveness estimates.

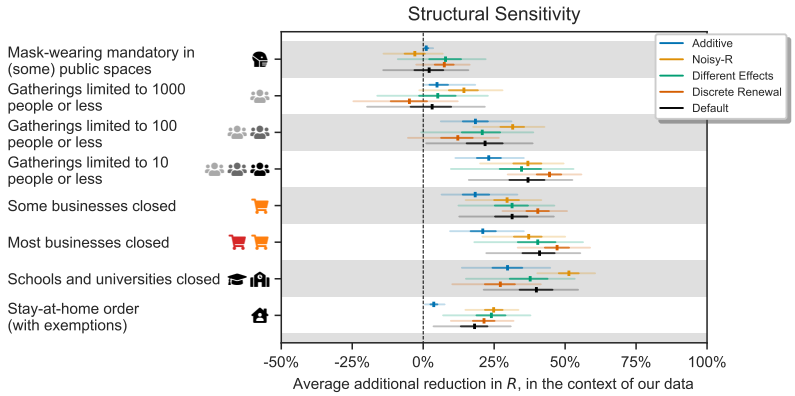

Figure B.8: Structural sensitivity analysis. Effectiveness estimates under different structural assumptions. Note that the additive model (blue) cannot be quantitatively compared to others (see text).

The results of the additive model (blue) cannot be directly compared to the other models since they are not expressed as percentage reductions in R, but in R_0. Its estimates are therefore smaller (but support the same conclusions).

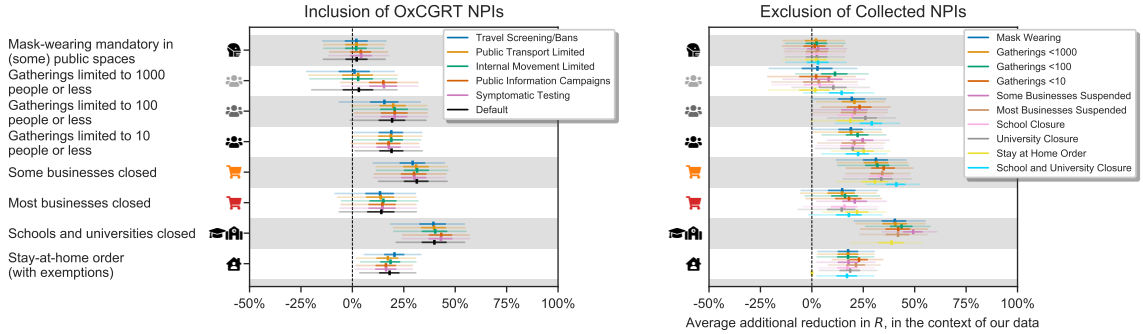

Figure B.9: Robustness to unobserved effects. *Left:* Results when controlling for previously unobserved NPIs. We include one additional NPI in turn and show the estimates for the NPIs in our study (the additional NPI is not shown). *Right:* Results when excluding previously observed NPIs. We exclude one of the NPIs in turn and show the estimates for the other NPIs. *Both:* Note that this figure shows the additional effect of each NPI. In other figures, we show the cumulative effects for gathering bans and businesses closures, denoted by showing multiple symbols (as explained in the caption to Figure 3). For example, Figure 3 displays the total effect of closing most nonessential businesses, while here we show the additional effect of closing most nonessential businesses *over* just closing some high-risk business. We show the additional effects here because the effect of a cumulative intervention would become undefined when part of it is excluded from the analysis.

Appendix B.3. Robustness to unobserved effects

Our data neither captures all NPIs implemented nor directly measures broader behavioural changes. Since these factors influence R, we must be wary of their effect being attributed to observed NPIs. We investigate this further by assessing how much effectiveness estimates change when previously unobserved factors are included and also when observed factors are excluded. This is best practice for assessing robustness to unobserved factors.[24,25] We also perform several additional investigations, outlined in Appendix C.1.

Unobserved factors can bias results if their timing is correlated with the timing of the observed NPIs.[26] The timing of our observed NPIs' implementation dates is indeed correlated, prompting the question how much excluding observed NPIs changes results. Figure B.9 (right) shows NPI effectiveness estimates when previously observed NPIs are excluded in turn. Estimates are robust, with all 50% credible intervals mutually overlapping. Considering that some of the excluded NPIs have strong estimated effects when included, and are correlated with other NPIs, this degree of robustness is surprisingly high. It suggests that unobserved factors will not significantly bias results as long as their effects and their correlations with the studied NPIs do not exceed those of the studied NPIs. We hypothesize that this robustness to unobserved factors is due to the noise on infection numbers in our model.[9]

In addition, Figure B.9 (left) shows NPI effectiveness estimates when we include (i.e. control for) additional NPIs, taken from the OxCGRT dataset.[27]

Our conclusions are not affected by controlling for these NPIs. This further suggests that unobserved factors are successfully treated as noise.

Appendix C. Additional sensitivity analyses and validation

Note: The individual effect of school and university closures cannot be meaningfully disentangled, as discussed in Appendix D.2. For completeness, we still show the *individual* estimates for school and university closures at the bottom of each graph in this section.

Appendix C.1. Role of NPI timing

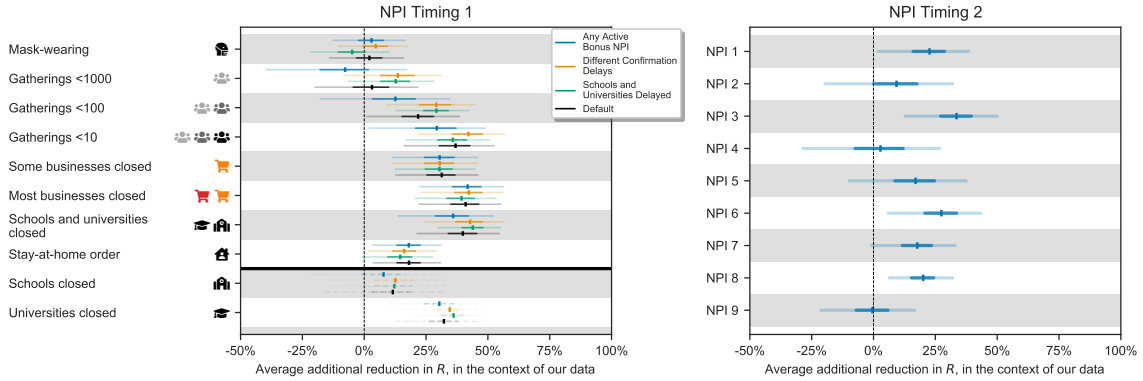

Figure C.10: Relationship between NPI effectiveness and timing. ***Left:*** **1) Effect of delaying school and university closures by 6 days in the data, 2) increasing confirmation delay by 2 days in countries without extensive testing and decreasing by 2 days in countries with testing, and 3) controlling for the start of government intervention (i.e. any NPI active).** ***Right:*** **Effectiveness for early and late NPIs. The model estimates the effect of all first, second, etc NPIs.**

There are several reasons to investigate the relationship between NPI timing and effectiveness. First, as previously discussed in Appendix B, unobserved factors such as behavior change may typically happen when the first NPIs are introduced and could confound their effects. This could lead to the surprisingly high estimates for school and university closures since these NPIs were often implemented early. Second, as discussed in the Results section, NPIs effectiveness may depend on the presence of other NPIs, and fewer NPIs will be present earlier which could reduce the additional effect of later NPIs.

In Figure C.10 (left), we show three additional experiments. First, following Flaxman et al.,[1] we control for the onset of government intervention by introducing an 'NPI' that is active from the day the first NPI is implemented. This test is intended to control for potential confounding from early unobserved NPIs and behavior changes. The result suggests that the estimate for school and university closures is not confounded or otherwise biased by the fact that these NPIs were often mandated early. However, including this covariate increases the uncertainty and somewhat decreases the mean estimate for bans of larger gatherings, which were also often the first NPIs implemented.

Second, we delay school and university closures by one mean generation interval (rounded to 6 days). This is motivated by the fact that children and adolescents are less likely to

show symptoms or die from COVID-19. Their infections may therefore show up with an additional delay in the case and death data, as they must first infect higher-risk demographics. Delaying school and university closures also causes them to be one of the later NPIs in most countries, a test if their high effectiveness may have been due to their relatively early appearance. As the inferred effect is stable, we can rule out these potential concerns (Figure C.10, left).

Third, we relax the assumption that the delay from infection to confirmation is equal between countries. It may be longer in countries that mostly test patients in hospitals and not in the community. Therefore, we increased the delay by 2 days in countries that did not offer testing to symptomatic people (using data from the Oxford COVID Government Response Tracker[27]), and decreased it by 2 days in countries that did. Although we do not have exact data for these different delays, 2 days is plausible based on hospital admission data.[28]

In Figure C.10 (right), we relabel the NPIs in each country as "1st NPI", "2nd NPI", etc. We then estimate the effect of these 'NPIs', which can represent various actual NPIs. The effectiveness ranking by order of implementation is $3, 6, 1, 8, 7, 5, 2, 4, 9$. The result shows that earlier NPIs are not necessarily estimated to be more effective, alleviating our concern that later NPIs are less effective due to interacting with earlier NPIs.

Appendix C.2. Sensitivity to data preprocessing

In the following sections, we show sensitivity to data preprocessing steps and additional parameters, and reprint sensitivity results from the previous section while additionally showing the *individual* estimates for school and university closures at the bottom of each graph.

We have smoothed case and death data with a moving average over a window of ± 2 days to incorporate the prior knowledge that large jumps in the data are due to inconsistent reporting. Here, we show results for smaller and larger smoothing windows, including no smoothing (Figure C.11, top right). There is no impact on our conclusions.

Furthermore, we exclude data in each country before 100 cases and 10 deaths are cumulatively reached to avoid biasing the model with foreign-imported cases. Here, we vary these thresholds. Interestingly, the death threshold has no clear effect on results, whereas the case threshold does (Figure C.11 (bottom left, bottom right). Raising the case threshold up to 500 removes a large portion of our data in March, making it difficult to determine R_0 (this is not the case for the death threshold). In contrast, making the case threshold too low likely introduces bias created by foreign-imported cases and early changes in testing regimes with lead to an overestimate of R_0. Nonetheless, all choices support our main conclusions (Discussion).

The other subfigures repeat sensitivity analyses from Appendix B with individual effects for school and universitiy closures shown at the bottom.

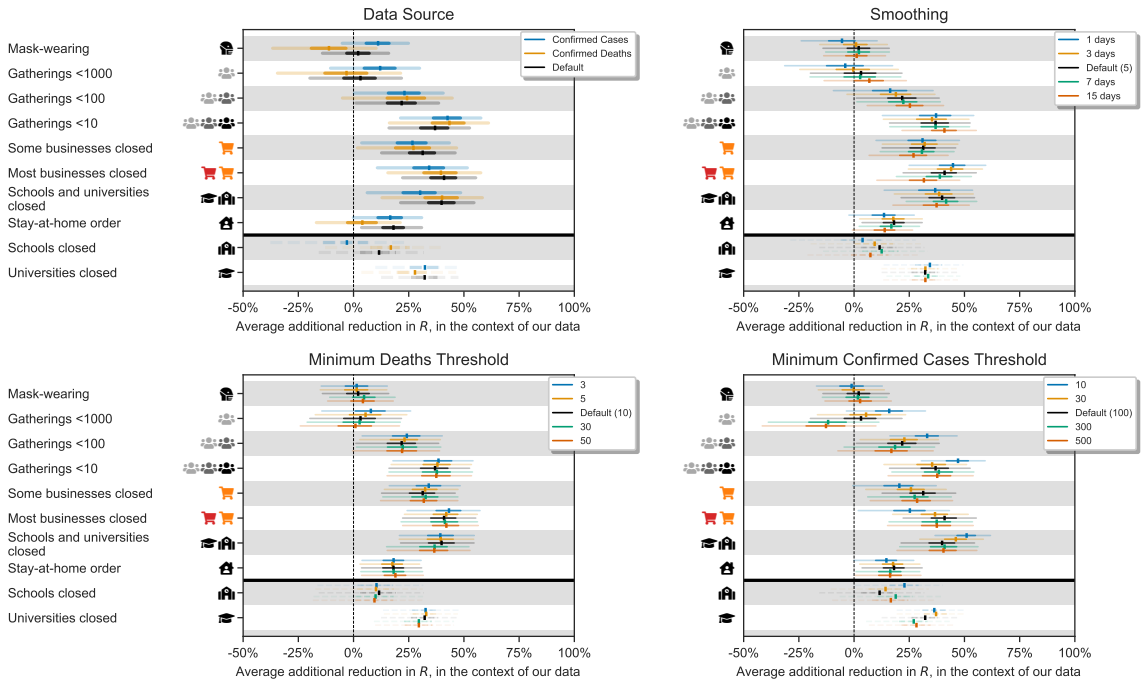

Figure C.11: Additional sensitivity analysis to data variations. ***Top right:*** **Sensitivity to smoothing case and death data over different windows (1 implies no smoothing).** ***Bottom left/right:*** **Sensitivity to the threshold for excluding case and death data.**

Appendix C.3. Sensitivity to additional epidemiological assumptions

For completeness, we test sensitivity to two further epidemiological assumptions. These are: 1) The prior on R_0—we vary its mean from a very small value (2.5) to the default value (3.25) and a large value (4.5). 2) The standard deviation of the noise σ_N on the growth rate. As previously discussed, this parameter is set to 0.2 using cross-validation. Other choices are not necessarily reasonable as they lead to miscalibrated predictions (the parameter could potentially be inferred from the data but this would be computationally challenging). We show these other choices for completeness. All choices support our conclusions and all credible intervals overlap, but higher noise, as expected, leads to higher uncertainty.

Results are shown in Figure C.12 (middle left, middle right).

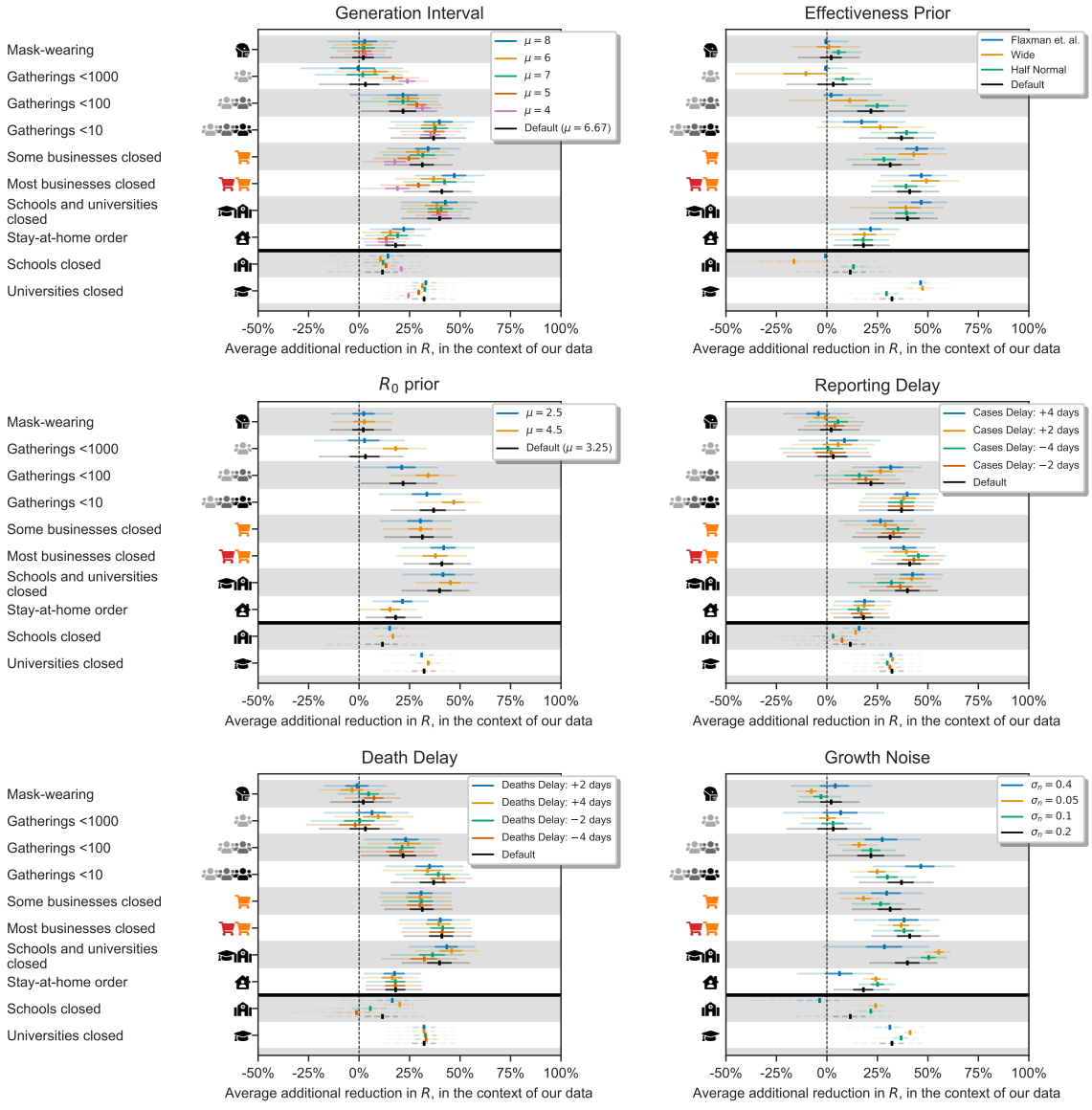

Figure C.12: Sensitivity to additional and previously shown epidemiological assumptions. ***Middle left:*** **Sensitivity to the prior mean on R_0.** ***Others:*** **Repetition of previously shown results with individual effect for school and university closures added.**

Appendix C.4. Additional country exclusions

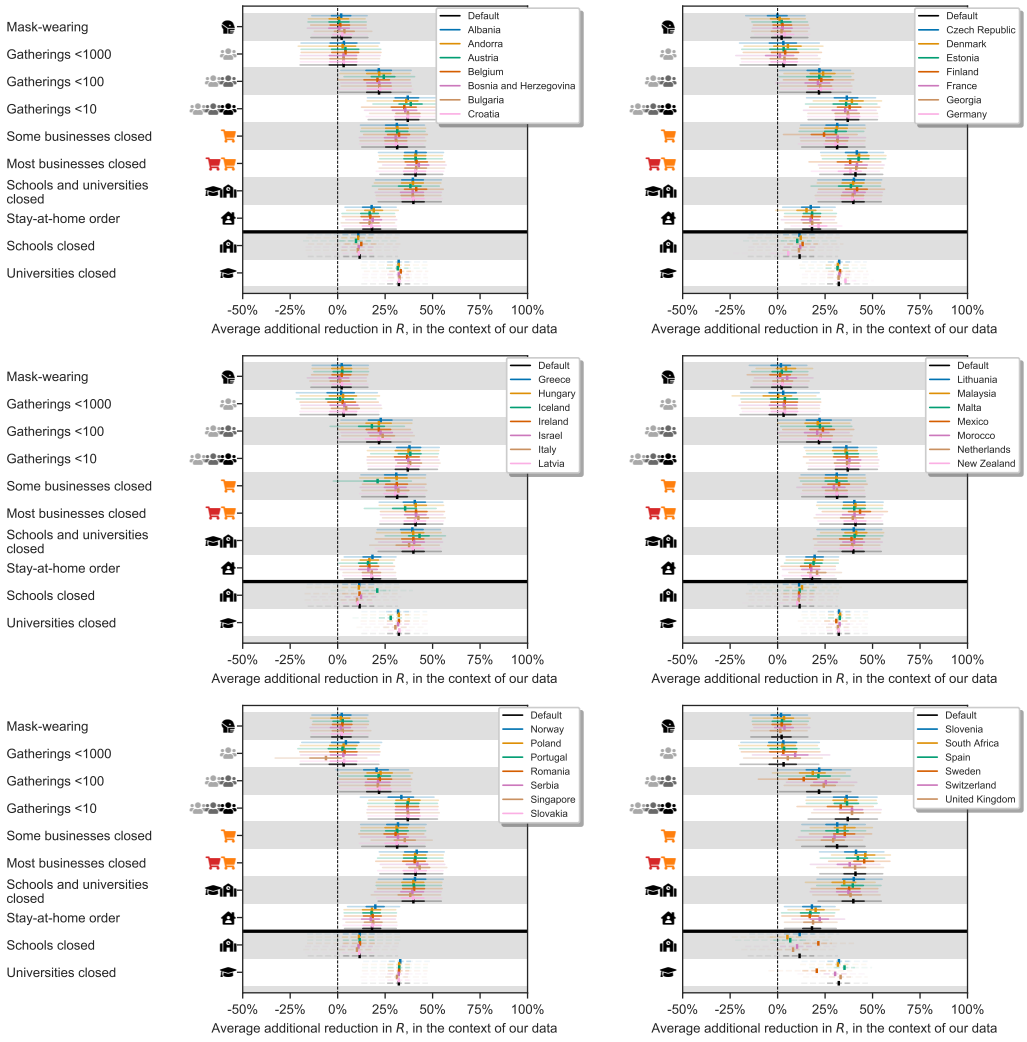

Figure C.13: Sensitivity to excluding one country at a time from the data.

Appendix C.5. Validation using predictions in excluded countries

Evaluating predictions on unseen data is an important model validation step, even if prediction is not the goal of the model.[29] We use 41-fold cross-validation: fitting the model on 40 countries and showing its predictions on the excluded country. We repeat this process for all 41 countries. In the excluded country, the first 14 days of death and case data are observed to allow roughly inferring R_0. These days are not used to infer NPI effectiveness. The activation dates of NPIs are also given, and the model uses the effectiveness estimates inferred from the 40 other countries.

Our model makes sensible, calibrated forecasts over long periods excluded in countries (Figures C.14 to C.17).

Explanation of the Figures: Vertical lines show the activation (and deactivation) date of NPIs. Shaded areas are 95% credible intervals. *Left:* The yellow and green lines ('Daily Infections - Later Reported/Later Fatal') show the estimates of daily new infections that will turn into confirmed cases ($N_t^{(C)}$) and deaths ($N_t^{(D)}$). Blue and red dots show the observed confirmed cases and deaths (smoothed), while blue and red lines show the median model estimates of cases (C_t) and deaths (D_t). Empty dots are not shown to the model. For each country, we show the full window of analysis (from the start of the epidemic until the first NPI was lifted, or 30th of May 2020, whichever was earlier (see Methods).

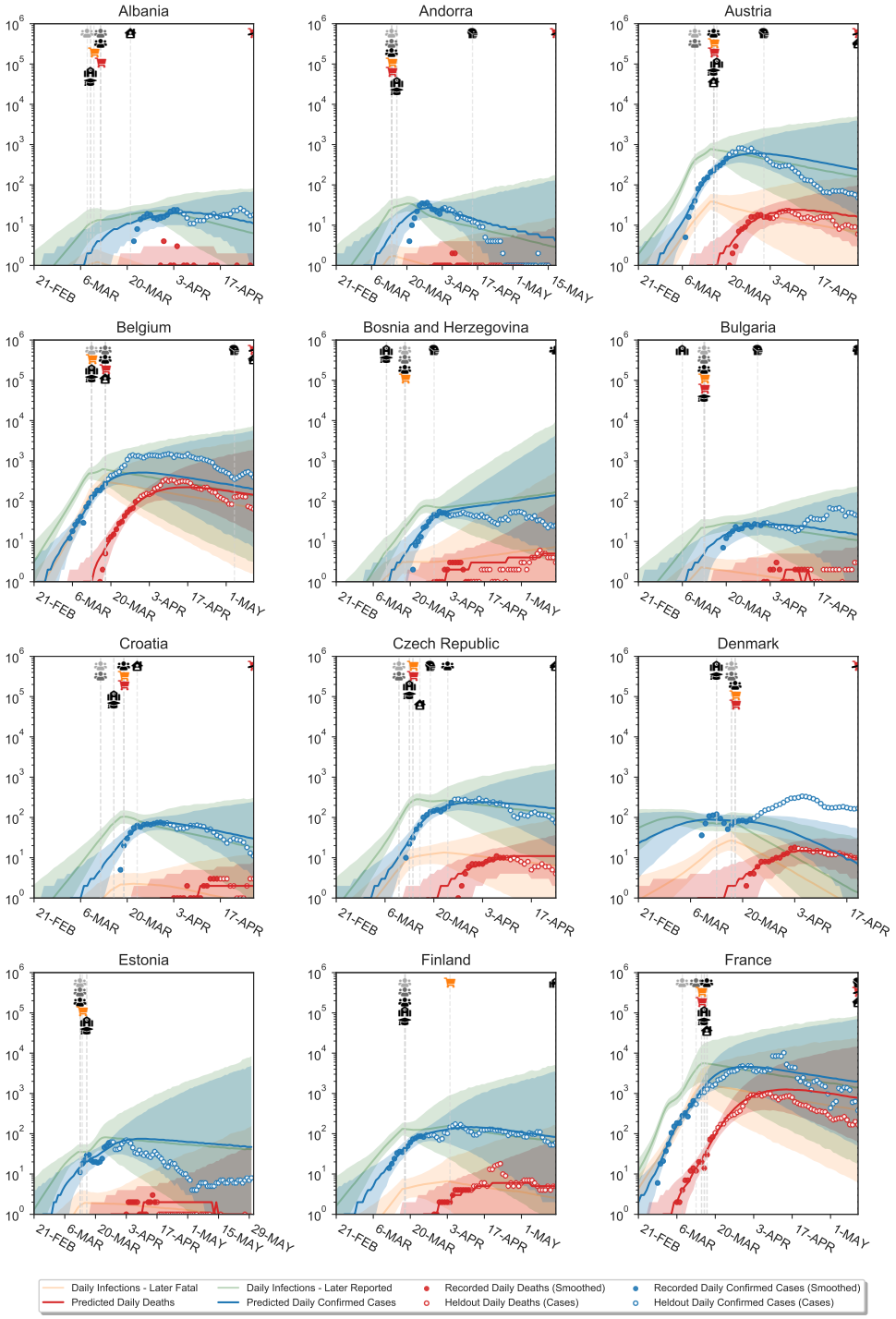

Figure C.14: Predictions on excluded countries.

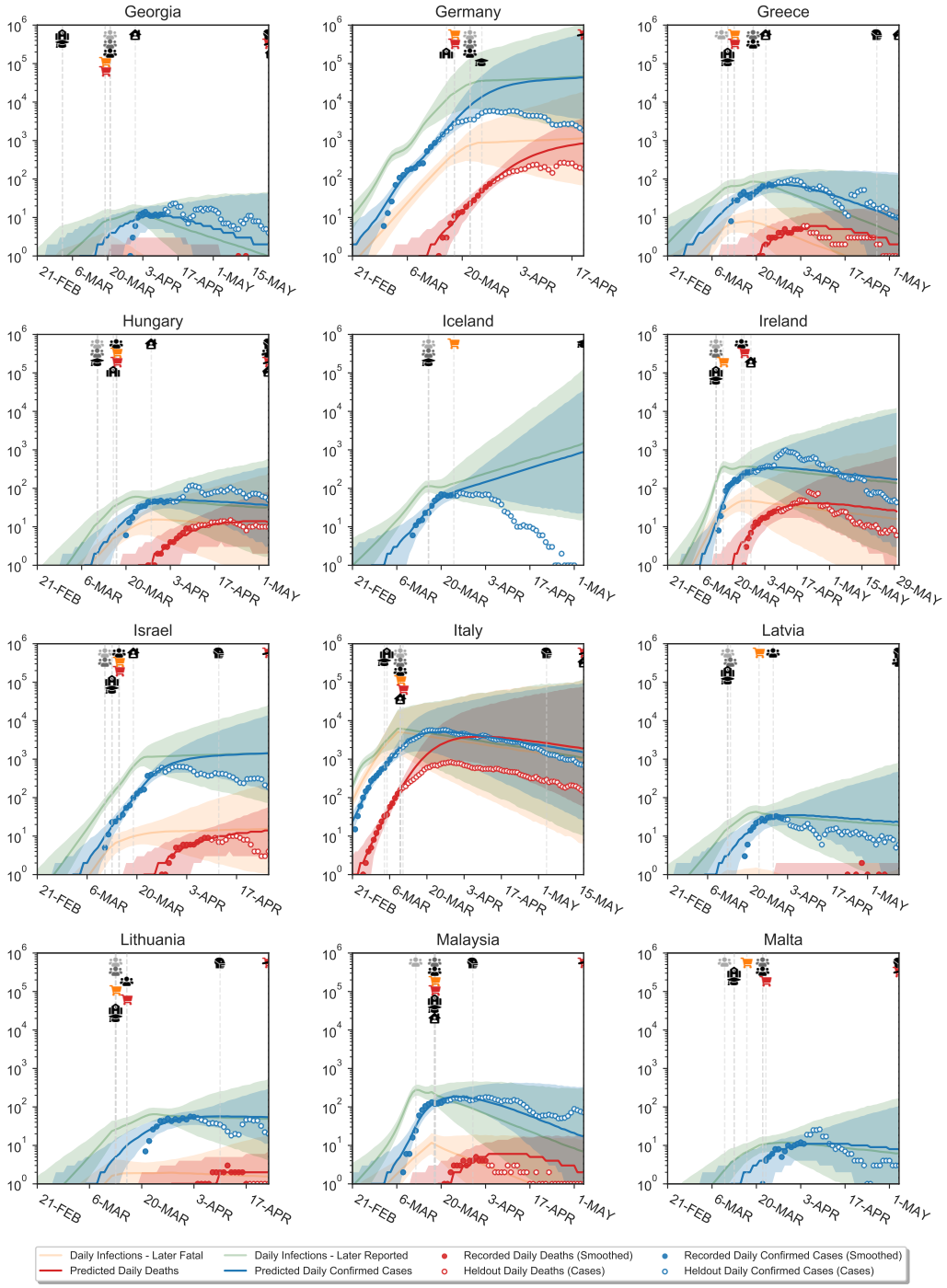

Figure C.15: Predictions on excluded countries.

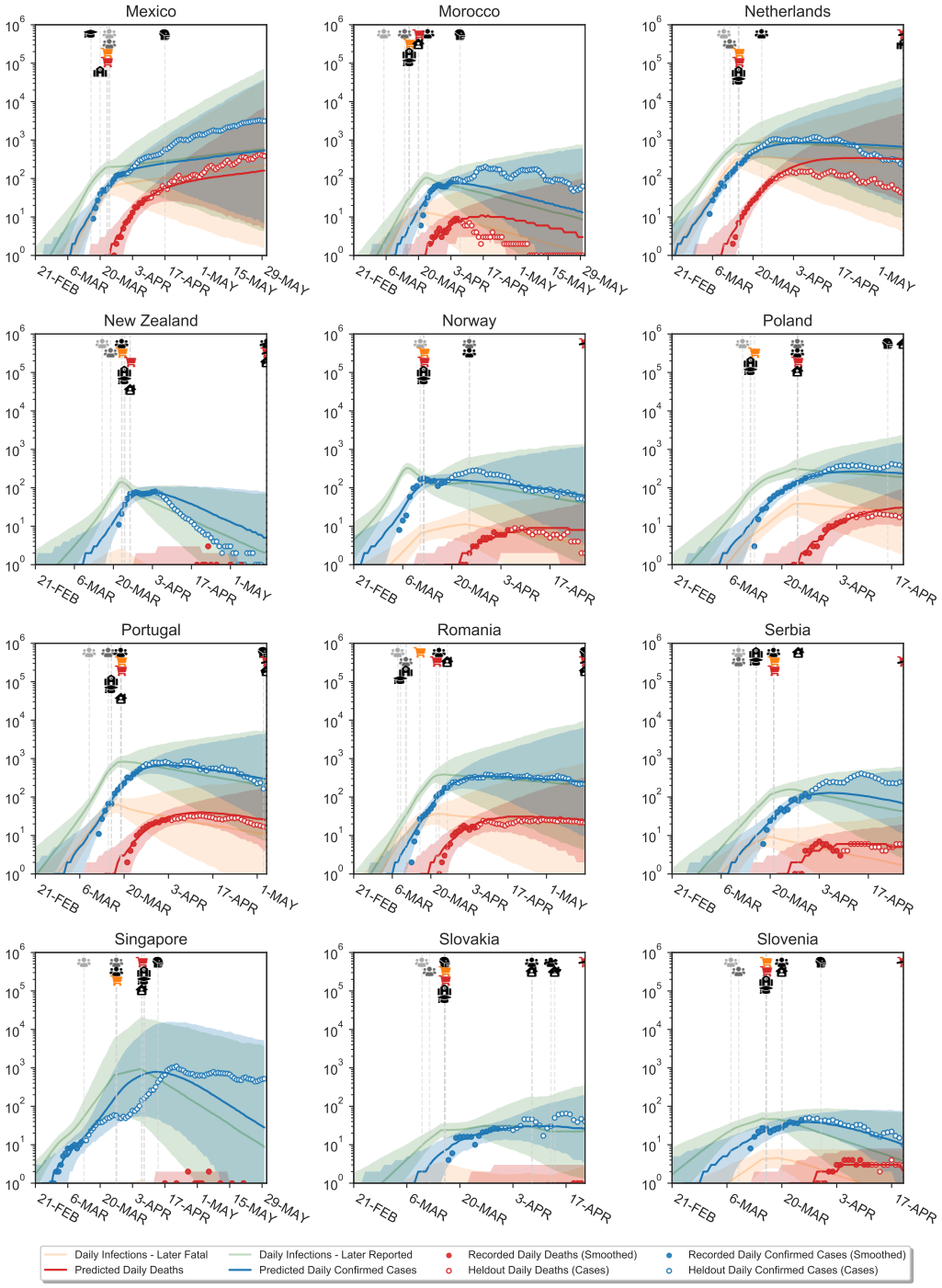

Figure C.16: Predictions on excluded countries.

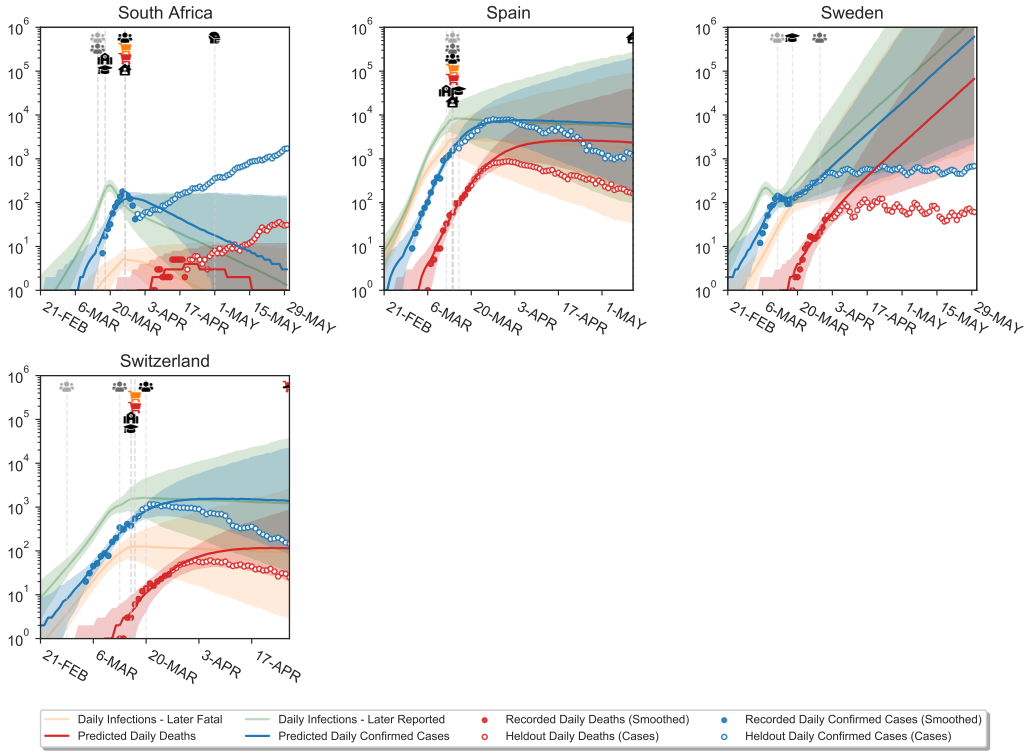

Figure C.17: Predictions on excluded countries. Empty dots are not shown to the model. 14 initial days are shown to the model, to enable inferring the basic R_0, but were not used to infer NPI effectiveness.

Appendix C.6. MCMC stability results

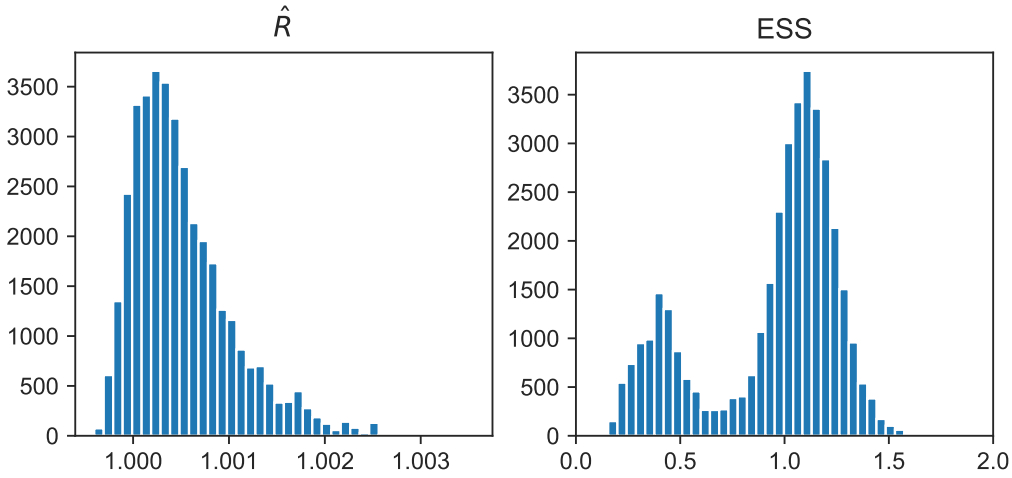

Figure C.18: MCMC stability results. ***Left:*** **R-hat statistic. Values are close to 1, indicating convergence.** ***Right:*** **Relative effective sample size. Values of 1 indicate perfect decorrelation between samples. Values over 1 indicate that the effective number of samples is higher than the actual number of samples (due to negative correlation), and vice versa.**

Appendix C.7. Posterior predictive distributions

The posterior predictive distribution (Figure C.19) shows the predicted true number of cases and deaths after observing the data. Although these curves can be called 'fits', the degree of fit to the data must be interpreted with great care. The fit is generally tight, but this is partly due to working with inferred latent noise variables: the noise terms $\epsilon_t^{(C)}$ and $\epsilon_t^{(D)}$ on the growth rates g_t. Inferring this latent noise allows the posterior predictive distribution to closely match the data without overfitting the effectiveness parameters to the data. Such behavior is common in Bayesian models, which often perfectly interpolate the data without overfitting.[30] The noise terms can account for periods where infections grew faster or slower than predicted based solely on the active NPIs. In such periods, the noise may account for changes in testing, reporting, and unobserved interventions.

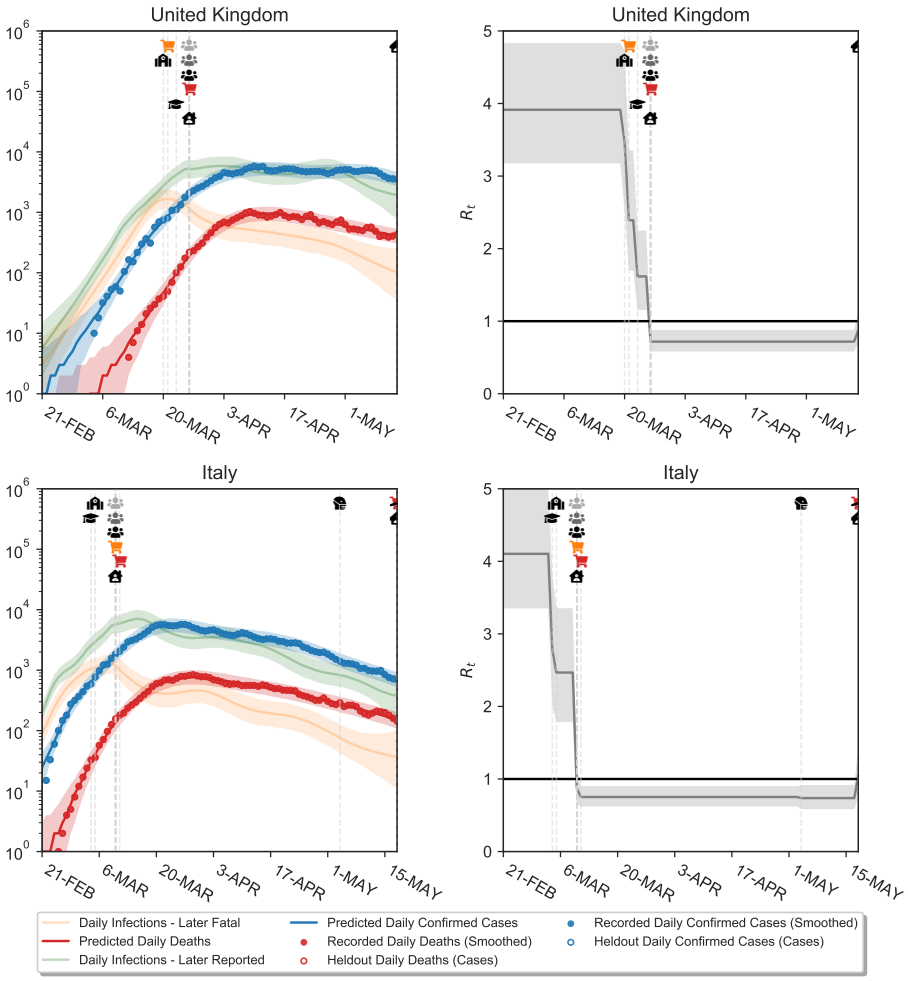

Figure C.19: ***Left:*** **Posterior predictive distributions for two exemplary countries. See text.** ***Right:*** **Inferred R_t over time.**

Appendix D. Additional results

Appendix D.1. Estimated R_0 by country

Table D.2: Estimated values for R_0, by country. The parenthesis give the 95% credible interval. The mean R_0 across countries is 3.8.

Country	Estimated R_0	Country	Estimated R_0
Albania	3.92 (2.85;5.25)	Lithuania	3.55 (2.55;4.78)
Andorra	2.81 (2.03;3.74)	Malaysia	3.52 (2.67;4.53)
Austria	3.59 (2.75;4.59)	Malta	3.64 (2.52;5)
Belgium	4.33 (3.44;5.39)	Mexico	4.7 (3.58;6.04)
Bosnia and Herzegovina	3.74 (2.73;4.93)	Morocco	4.43 (3.44;5.68)
Bulgaria	4.32 (3.26;5.63)	Netherlands	3.53 (2.77;4.41)
Croatia	3.93 (2.93;5.15)	New Zealand	2.65 (1.82;3.65)
Czech Republic	4.15 (3.06;5.46)	Norway	2.96 (2.21;3.89)
Denmark	2.92 (2.17;3.78)	Poland	4.87 (3.69;6.31)
Estonia	2.88 (2.11;3.78)	Portugal	4.4 (3.4;5.58)
Finland	2.96 (2.21;3.84)	Romania	4.74 (3.7;5.97)
France	3.91 (3.14;4.81)	Serbia	4.44 (3.35;5.74)
Georgia	4.3 (3.19;5.69)	Singapore	3.53 (2.72;4.45)
Germany	3.51 (2.73;4.43)	Slovakia	3.83 (2.64;5.29)
Greece	3.39 (2.56;4.39)	Slovenia	3.25 (2.32;4.4)
Hungary	4.74 (3.66;6.07)	South Africa	5.65 (4.46;7.13)
Iceland	1.59 (0.98;2.38)	Spain	4.71 (3.8;5.71)
Ireland	4.42 (3.48;5.56)	Sweden	2.25 (1.7;2.91)
Israel	4.39 (3.34;5.69)	Switzerland	3.07 (2.36;3.92)
Italy	4.12 (3.35;5.01)	United Kingdom	3.94 (3.17;4.83)
Latvia	3.27 (2.33;4.42)		

Appendix D.2. The individual effects of school and university closures

The dates of school and university closures coincide nearly perfectly for every country except Iceland and Sweden, which closed universities but not schools (Figure 1). As a consequence, the inferred individual effects depend strongly on the in- or exclusion of these countries in the dataset (Figure D.20). We conclude that we cannot meaningfully disentangle these two NPIs based on only two countries, and show their joint effect (Figure 3), for which there is much more data.

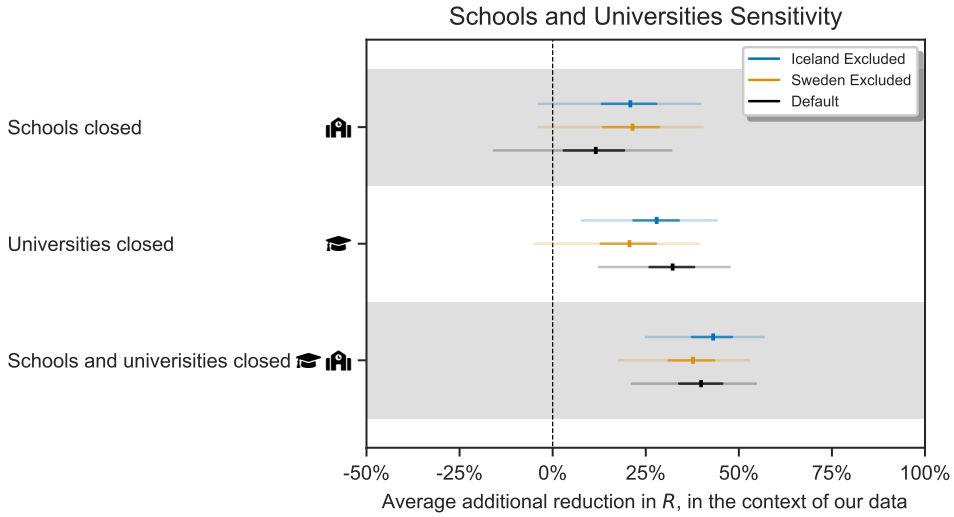

Figure D.20: The individual effectiveness of closing schools and of closing universities, as well as the joint effect of closings school *and* universities, estimated on all countries (default), all countries except Sweden, and all countries except Iceland. Median, 50% and 95% credible intervals are shown. The individual effect of school closures is highly sensitive to the in- or exclusion of one of the countries. If we include all countries, we would conclude that the joint effect is mostly driven by university closures. However, if we exclude Sweden, we would conclude that schools closures play a larger role. As there is no strong justification for in- or excluding one particular country, we conclude that we cannot meaningfully disentangle the effects of school and university closures. The combined effect is more stable.

Appendix D.3. Collinearity

Table D.3 shows the total number of days across all countries available to distinguish NPI effects. For every pair of NPIs (row - column), the entry shows the number of country-days on which only one of the NPIs was implemented (but not both or none). Note that we do not show the traditional collinearity statistics, variance inflation factors and data correlations, since their applicability to time series data is limited. In particular, the value of these statistics in our data increases as data for a longer time period becomes available, which would misleadingly suggest that we could address problems from collinearity by using less data.

Table D.3: Total number of days across all countries available to distinguish NPI effects. For every pair of NPIs (row - column), the entry shows the number of country-days on which exactly one of the NPIs was implemented. Abbreviations: G.: Gatherings; SBC: Some businesses closed; MBC: Most nonessential businesses closed; SaUC: Schools and universities closed; SaHO: Stay-at-home order.

	G. <1000	G. <100	G. <10	SBC	MBC	SaUC	SaHO
Mask-wearing	1829	1686	1472	1554	1315	1588	1038
Gatherings <1000		143	501	299	620	325	1173
Gatherings <100			358	240	515	284	1030
Gatherings <10				262	403	308	696
Some businesses closed					331	176	898
Most businesses closed						393	569
Schools and universities closed							940

Appendix D.4. Correlations between effectiveness estimates

The effectiveness parameters α_i are typically negatively correlated with each other for NPIs which are often used together, reflecting uncertainty about which NPI is reducing R. Excessive collinearity in the data would result in wide posterior credible intervals with strong correlations,[20] but we find weak posterior correlations between effectiveness estimates. The strongest correlation between any pair of NPIs is -0.41, between limiting gatherings to 100 attendants and 1000 attendants or less (Figure D.21). The weak correlations are one indicator that collinearity is manageable with our dataset.

To better understand posterior correlations, we visualize their effect in hosted video files. As we condition on different values for one NPI, we can see that the estimates of other NPIs change only slightly, always staying well within the credible intervals in Figure 3. The significance of posterior correlations is small enough that it is possible to calculate a reasonable approximation to the mean effect of a set of NPIs by simply combining the mean percentage reductions for each individual NPI (e.g. two 50% reductions lead to a 75% reduction). For example, this approximation leads to a joint effect of 82% for all NPIs together, which matches the exact mean joint effect (also 82%) up to rounding error.

Videos are available online here.

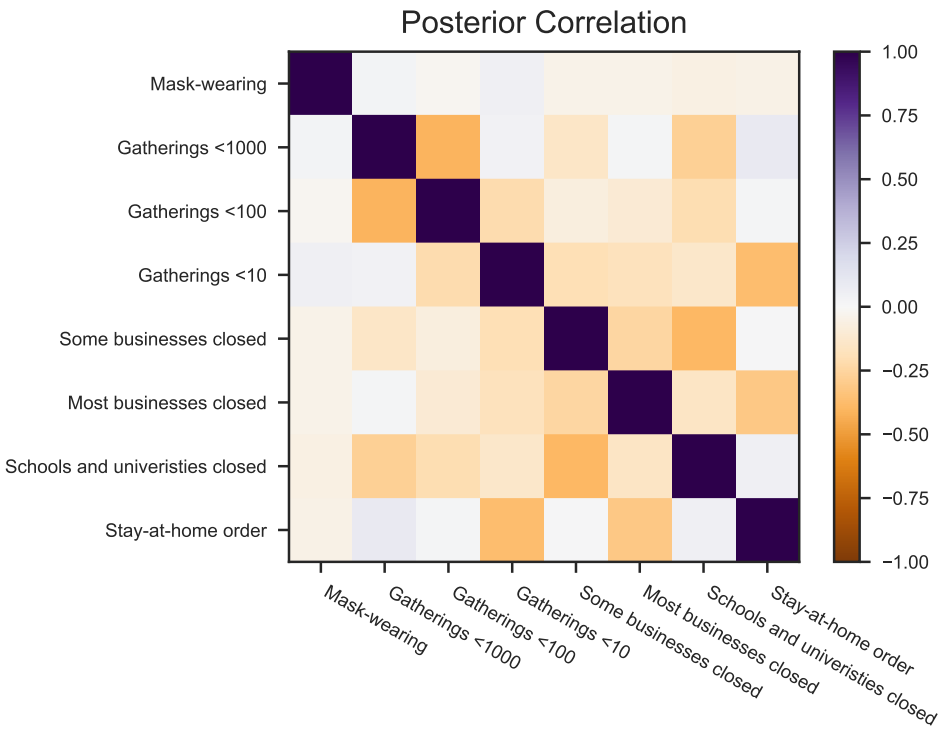

Figure D.21: Posterior correlations between effectiveness parameters α_i.

Appendix E. Additional discussion of assumptions and limitations

Appendix E.1. Limitations of the data

We only record NPIs if they are implemented in most of a country (if they affect more than three fourths of the population). We thus miss if NPIs were only implemented regionally. For example, a few regions in Germany implemented stay-at-home orders but most did not. Thus, Germany is listed as "no stay-at-home order" in our data. Additionally, our NPI definitions were not perfectly granular. For example, a gathering ban on gatherings of >15 people and a ban on gatherings of >60 people would both fall under the NPI "Gatherings limited to 100 people or less", despite likely having different effects on R. Finally, while we included more NPIs than previous work (Table F.4), there are many NPIs for which we were not able to collect enough high-quality data for our modeling, such as public cleaning or changes to public transportation.

Appendix E.2. Model limitations

Independence of country and time. We assume that the effect of NPIs on R is constant across countries and time. However, the exact implementation and adherence of each NPIs is likely to vary. Our uncertainty estimates in Figure 3 account for these problems only to a strictly limited degree. Additionally, different countries have different cultural norms and age profiles, affecting the degree to which a particular intervention is effective. For example, a country where a higher proportion of the population is in education will likely observe a larger effect from a government order to close schools and universities. Our estimates thus should be adjusted to local circumstances. To address differences between countries, our structural sensitivity analysis includes a model where each NPI can have a different effect per country (Appendix C). The average effectiveness estimates across countries in this model match the conclusions from our default model.

Testing, reporting, and the IFR. Our model can account for differences in testing (and IFR/reporting) between countries and over time, as discussed in Appendix A. However, we have not used additional data on testing to validate if it does so reliably. Our model may struggle to account for changes in the testing regime—for instance, when a country reaches its testing capacity so that the ascertainment rate declines exponentially. An exponential decline would have the same effect on observations as an unobserved NPI. Consequently, we cannot quantify its effect on our results (though the sensitivity analyses look promising).

Interaction between NPIs. As discussed in the Results section, our model reports the average additional effect each NPI had in the contexts where it was active in our data (in the sense mathematically shown by Sharma et al.[9]). Figure 3 (bottom left) summarises these contexts, aiding interpretation. The effectiveness of an NPI can only be extrapolated to other contexts if its effect does not depend on the context. For example, we may expect

that closing schools has a similar effectiveness whether or not businesses are also closed. But wearing masks in public may be less effective when a stay-at-home order limits public interactions.

Growth rates. The functional form of the relationship between the daily growth rate of the number of infections g and the reproductive number R holds exactly when the epidemic is in its exponential growth phase, but becomes less accurate as the number of susceptible people in a population decreases and/or control measures are implemented. However, we also reported results from a *renewal process* model[8] that lacks this assumption and finds similar effectiveness estimates.

Signalling effect of NPIs. As we explained in the Discussion for school closures, we do not distinguish between the direct effect of an NPI and its indirect effect as it signals the gravity of the situation to the public. Conversely, lifting interventions may also have a signalling effect.

Homogeneous effect of interventions. We work under the implicit assumption that NPIs affect different population groups equally. This could affect results in various ways. For example, suppose country A tests an older demographic than country B, and we are considering the effect of an NPI that mostly affects the older demographic (for example, isolating the elderly). Then the NPI will appear to have a greater effect on confirmed cases in country A, breaking the assumption that effects are constant across countries. Our previous discussion of interpreting results when this assumption is violated applies.

Appendix F. Overview of previous work

Table F.4: Existing data-driven studies of the effectiveness of observed (as opposed to hypothetical) NPIs in reducing the transmission of COVID-19.

Study	NPIs studied	Regions/countries studied	Method
Flaxman et al., 2020[1]	School or university closure, case-based isolation, ban on large public events, social distancing, lockdown	Austria, Belgium, Denmark, France, Germany, Italy, Norway, Spain, Sweden, Switzerland, UK	Semi-mechanistic Bayesian hierarchical model

Continued on next page

Table F.4 – *Continued from previous page*

Study	NPIs studied	Regions/countries studied	Method
Chen and Qiu, 2020[31]	Travel restriction, mask-wearing, lockdown, social distancing, school closure, centralized quarantine	Italy, Spain, Germany, France, UK, Singapore, South Korea, China, U.S.	Regression with delayed effect Susceptible-Infectious-Removed (SIR) model
Banholzer et al., 2020[32]	School closure, border closure, event ban, gathering ban, venue closure, lockdown, work ban	U.S., Canada, Australia, Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, UK, Norway, Switzerland	Semi-mechanistic Bayesian hierarchical model
Hsiang et al., 2020[33]	Restricting travel (5 subcategories), distancing (10 subcategories), quarantine and lockdown (2 subcategories), additional policies (2 subcategories)	China, South Korea, Italy, Iran, France, U.S.	Linear regression on estimated growth rates
Choma et al., 2020[34]	Single aggregated NPI	22 countries and 25 states	Regression with Susceptible-Infectious-Removed-Deceased (SIRD) model
Dehning et al., 2020[35]	Contact ban, restrictions on gatherings, schools, childcare, businesses	Germany	Bayesian inference of transmission rate
Siedner et al., 2020[36]	General social distancing	U.S.	Interrupted time-series
Kraemer et al., 2020[37]	Travel restrictions and cordon sanitaire	China	Regression
Kucharski et al., 2020[38]	Travel restrictions	Wuhan (China)	Various, including Susceptible-Exposed-Infectious-Removed (SEIR) model

Continued on next page

Table F.4 – *Continued from previous page*

Study	NPIs studied	Regions/countries studied	Method
Dandekar and Barbastathis, 2020[39]	General quarantine and isolation	Wuhan, Italy, South Korea, and U.S.	A mix of a mechanistic model and a data-driven neural network model
Maier and Brockmann, 2020[40]	General quarantine and isolation	Mainland China	Quantitative fits to empirical data
Sears et al., 2020[41]	Mobility changes as a proxy for stay-at-home mandates	U.S.	Difference-in-differences statistical model
Jarvis et al., 2020[42]	Physical (social) distancing measures	UK	Questionnaire data and compartmental epidemic model
Orea and Álvarez, 2020[43]	Lockdown	Spain	Spatial econometric analysis
Lorch et al., 2020[44]	Mobility restrictions, testing & tracing, social distancing, and business restrictions	Tübingen (Germany)	Authors' own spatiotemporal model of epidemics
Gatto et al., 2020[45]	Various restrictions to mobility and human-to-human interactions	Italy	Susceptible–Exposed–Infected–Recovered (SEIR)-like disease transmission model
Quilty et al., 2020[46]	Intercity travel restrictions	Beijing, Chongqing, Hangzhou, and Shenzhen (Mainland China)	Branching process transmission model

Appendix G. Handling edge cases in the data collection

In our data collection process, we relied on carefully worded definitions of 9 different NPIs (Table F.4), which allowed us to systematically determine the date on which a country imposed an NPI and, if applicable, the date the NPI was lifted.

In some cases, however, we faced ambiguities in how to interpret the start date of an NPI. One kind of challenge arose when descriptions of policy measures were less specific than our NPI definitions (e.g., a ban on "large gatherings" that does not specify the exact number of people that constitutes a "large gathering"). Another difficulty was due to NPI policies that made distinctions that we did not make in our own NPI definitions (e.g., an NPI policy that made a distinction between the number of people able to gather indoors vs outdoors).

To resolve these ambiguities in a consistent manner, our researchers developed a set of principles and guidelines that were followed during the data collection process. For each of the examples below, the relevant sources are available in the data table in the supplementary material.

Situation: Sometimes only public gatherings are banned, with no explicit ban on private gatherings

How we deal with it: We still counted this as a ban on gatherings.

Examples:

- Sweden: In Sweden, they banned all *public* gatherings of more than 50 people (demonstrations, religious meetings, theater performances, markets, and other events that relied on the constitutional freedom of assembly), however, the ban did not have a mandate to prohibit *private* gatherings (such as private parties). We counted this as a ban on gatherings.
- Finland: In Finland, they banned all public gatherings of more than 10 people on the 16th of March. Although formal restrictions did not apply to private gatherings, this policy met our definition of a ban on gatherings. (Note that this inclusion seems particularly valid in light of the fact that, according to Finnish police, the formal restrictions on public events were widely interpreted to apply to private gatherings as well, and there were very few reports of large private parties despite the absence of formal restrictions.)

Situation: The size limits on gatherings sometimes differ between indoor and outdoor gatherings.

How we deal with it: In these cases, we relied on the limitations on indoor events, as these events entail a greater risk of transmission.

Example:

- Spain: In Spain, a range of rules were employed as the country gradually eased restrictions on gatherings. In phase 1, cultural events were permitted with up to 30 people indoors and up to 200 outdoors. This was counted as "Gatherings limited to 100 people or less."

Situation: The size limit on gatherings sometimes differs between different types of gatherings.

How we deal with it: In this case, researchers would use their best judgment to infer whether the restriction would apply to *most* gatherings of a given size.

Example:

- Spain: In Spain, phase 1 of the reopening allowed for cultural events to have up to 30 participants indoors, while social gatherings were limited to 10 people. In this case, since "cultural events" is broad, we counted this as a case of "gatherings limited to 100 people or less." However, if for example all gatherings above 5 people had been banned with an exception for funerals, we would have counted this as "gatherings limited to 10 people or less," since the exemption only applied to a minority of gatherings.

Situation: Limitations on gathering sizes are not clearly given, yet a policy stating that "large events are banned" is in place.

How we deal with it: Our researchers used the relevant context to infer the most likely scope of the policy.

Example:

- Albania: on March 8 "authorities had also ordered cancellations of all large public gatherings including cultural events and were asking sporting federations to cancel scheduled matches". The events that are mentioned here are multi-thousand person gatherings, and so we took March 8th to be the start date of "Gatherings limited to 1000 people or less". However it was unclear whether gatherings of 100-1000 would also have been banned, so we did not yet say that "Gatherings limited to 100 people or less" was instantiated.

Situation: Only some schools were closed, or schools reopened gradually.

How we deal with it: Since our definition of the NPI is that "*Most* schools are closed," we did not count the closure of just a few schools or school years sufficient to meet this criteria. Similarly, if schools reopened for only a very limited number of year groups, for example for final year students sitting exams, we did not count this as a lifting of the "most schools closed" NPI.

Examples:

- Sweden: Sweden kept all schools through 9th grade open, but closed high schools (>16 year olds). In this case, we did not count this as "Most schools closed", since more than 75% of students are below 9th grade.
- Czech Republic: After closing all schools on the March 13, the Czech Republic allowed schools to reopen for teaching in some contexts from May 11 (specifically for students in their final year of primary school or high school preparing for exams). However, we still counted this as "Most schools closed" since the majority of students were not in school. We recorded the end date for school closure to be June 8, when all schools reopened.

Situation: In a country where most non-essential businesses were closed, the lifting of business closures is gradual, and businesses in different sectors are successively allowed to open.

How we deal with it: Countries reopen sectors in different, idiosyncratic ways and successions. Given the available data, it is not feasible to create a principle that can be applied unambiguously to every single case without some involvement of researcher judgment. The general guideline we used was: If only a few, low-risk businesses (e.g., bike stores, hardware stores, etc.) are additionally allowed to reopen, then we still counted this as "Most nonessential businesses closed." However if any *one* of the following criteria are met, then we counted "Most nonessential businesses closed" as having lifted, but the "Some businesses closed" NPI was still in place:

- All regular retail stores, with only a few exceptions e.g. size limitations, are open
- Contact-based services, such as hairdressers and tattoo parlors, are open
- Restaurants and bars are open and serving indoors

We decided that meeting any one of these criteria is a sufficient condition for taking a country from "Most nonessential businesses closed" to "Some businesses closed." This heuristic was partly based on the fact that the status of these categories appeared to be consistently correlated, meaning that, even in the absence of complete specifications as to what had reopened or not, it was typically possible to infer the overall level of reopening based on either of these categories. Meeting at least one of these criteria was considered a necessary condition for ending the "Most nonessential businesses closed" NPI.

Examples:

- Slovakia: On April 22, retail operations and services up to 300 m2 opened. Since this meets one of the sufficient conditions, we counted April 22 as the end date for "Most nonessential businesses closed"
- Ireland: On May 18, the following reopened: hardware stores, builders merchants and those providing essential supplies, retailers involved in the sale and repair of vehicles, certain office supply stores. Because this white list does not meet any of the

three criteria, Ireland's end date for "Most nonessential businesses closed" was not counted as May 18.

- Czech Republic: On April 20, several businesses reopened, including farmer's markets, marketplaces, locksmiths, bike shops, car dealers, electronics stores. At this point, none of the criteria were met, so we recorded the Czech Republic as still having "Most nonessential businesses closed". On May 11, a long list of businesses reopened, including barbers, hairdressers, museums, all establishments in sufficiently large shopping centers, shows with up to 100 participants, and restaurants with a window facing the street. Since contact-based services (hairdressers) and all retail establishments in sufficiently large spaces were allowed to reopen, we counted May 11 as the end date for the "Most nonessential businesses closed" NPI.
- Croatia: On April 27, all "trade activities" (except within shopping malls), service jobs that don't involve physical contacts, museums, libraries, galleries opened. Since the criteria regarding 'all retail stores being open' was met, we counted April 27 as the end date for "Most nonessential businesses closed".

Appendix H. References

References

1 Seth Flaxman et al. "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe". In: *Nature* (2020), pp. 1–8.

2 Suryakant Yadav and Pawan Kumar Yadav. "Basic Reproduction Rate and Case Fatality Rate of COVID-19: Application of Meta-analysis". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (May 2020). DOI: 10.1101/2020.05.13.20100750. URL: https://www.medrxiv.org/content/10.1101/2020.05.13.20100750v1.

3 J Wallinga and M Lipsitch. "How generation intervals shape the relationship between growth rates and reproductive numbers". In: *Proceedings of the Royal Society B: Biological Sciences* 274.1609 (Nov. 2006), pp. 599–604. DOI: 10.1098/rspb.2006.3754.

4 D Cereda et al. "The early phase of the COVID-19 outbreak in Lombardy, Italy". In: (Mar. 20, 2020). arXiv: 2003.09320v1 [q-bio.PE]. URL: https://arxiv.org/abs/2003.09320.

5 John M Griffin et al. "A rapid review of available evidence on the serial interval and generation time of COVID-19". In: *medRxiv* (2020).

6 Tapiwa Ganyani et al. "Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020". In: *Eurosurveillance* 25.17 (2020), p. 2000257.

7 Luca Ferretti et al. "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing". In: *Science* 368.6491 (2020).

8 Christophe Fraser. "Estimating individual and household reproduction numbers in an emerging epidemic". In: *PloS one* 2.8 (2007).

9 Mrinank Sharma et al. "On the Robustness of Effectiveness Estimation of Nonpharmaceutical Interventions Against COVID-19 Transmission". In: *Arxiv* (2020).

10 Andrew Gelman, Jessica Hwang, and Aki Vehtari. "Understanding predictive information criteria for Bayesian models". In: *Statistics and computing* 24.6 (2014), pp. 997–1016.

11 Natalie M Linton et al. "Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data". In: *Journal of clinical medicine* 9.2 (2020), p. 538.

12 Qun Li et al. "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia". In: *New England Journal of Medicine* 382.13 (Mar. 2020), pp. 1199–1207. DOI: 10.1056/nejmoa2001316.

13 Qifang Bi et al. "Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (Mar. 2020). DOI: 10.1101/2020.03.03.20028423. URL: https://www.medrxiv.org/content/10.1101/2020.03.03.20028423v3.

14 Robert Verity et al. "Estimates of the severity of coronavirus disease 2019: a model-based analysis". In: *The Lancet Infectious Diseases* (Mar. 2020). DOI: 10.1016/s1473-3099(20)30243-7.

15 Simon Duane et al. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (Sept. 1987), pp. 216–222. DOI: 10.1016/0370-2693(87)91197-x.

16 Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.47 (2014), pp. 1593–1623. URL: http://jmlr.org/papers/v15/hoffman14a.html.

17 Ying Liu et al. "The reproductive number of COVID-19 is higher compared to SARS coronavirus". In: *Journal of travel medicine* (2020).

18 John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (2016), e55.

19 A. Gelman et al. "Bayesian Data Analysis, Second Edition". In: Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2003. Chap. Model checking and improvement. ISBN: 9781420057294. URL: https://books.google.com.mx/books?id=TNYhnkXQSjAC.

20 Carsten F Dormann et al. "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance". In: *Ecography* 36.1 (2013), pp. 27–46.

21 Anne Cori et al. "A new framework and software to estimate time-varying reproduction numbers during epidemics". In: *American journal of epidemiology* 178.9 (2013), pp. 1505–1512.

22 Pierre Nouvellet et al. "A simple approach to measure transmissibility and forecast incidence". In: *Epidemics* 22 (2018), pp. 29–35.

23 Simon Cauchemez et al. "Estimating the impact of school closure on influenza transmission from Sentinel data". In: *Nature* 452.7188 (2008), pp. 750–754.

24 P. R. Rosenbaum and D. B. Rubin. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 45.2 (1983), pp. 212–218. DOI: 10.1111/j.2517-6161.1983.tb01242.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1983.tb01242.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1983.tb01242.x.

25 James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models". In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 1–94.

26 A Gelman and J Hill. "Causal inference using regression on the treatment variable". In: *Data Analysis Using Regression and Multilevel/Hierarchical Models* (2007).

27 Thomas Hale et al. *Oxford COVID-19 Government Response Tracker*. Blavatnik School of Government. https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker. 2020.

28 Eva S Fonfria et al. "Essential epidemiological parameters of COVID-19 for clinical and mathematical modeling purposes: a rapid review and meta-analysis". In: *medRxiv* (2020).

29 Andrew Gelman and Cosma Rohilla Shalizi. "Philosophy and the practice of Bayesian statistics". In: *British Journal of Mathematical and Statistical Psychology* 66.1 (Feb. 2012), pp. 8–38. DOI: 10.1111/j.2044-8317.2011.02037.x.

30 Carl Edward Rasmussen. "Gaussian processes in machine learning". In: *Summer School on Machine Learning*. Springer. 2003, pp. 63–71.

31 Xiaohui Chen and Ziyi Qiu. "Scenario analysis of non-pharmaceutical interventions on global COVID-19 transmissions". https://arxiv.org/abs/2004.04529. Apr. 7, 2020.

32 Nicolas Banholzer et al. "Impact of non-pharmaceutical interventions on documented cases of COVID-19". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (Apr. 2020). DOI: 10.1101/2020.04.16.20062141. URL: https://www.medrxiv.org/content/10.1101/2020.04.16.20062141v3.

33 Solomon Hsiang et al. "The Effect of Large-Scale Anti-Contagion Policies on the Coronavirus (COVID-19) Pandemic". In: *medRxiv* (May 2020), p. 2020.03.22.20040642. DOI: 10.1101/2020.03.22.20040642.

34 Jacques Naude et al. "Worldwide Effectiveness of Various Non-Pharmaceutical Intervention Control Strategies on the Global COVID-19 Pandemic: A Linearised Control Model". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (May 2020). DOI: 10.1101/2020.04.30.20085316. URL: https://www.medrxiv.org/content/early/2020/05/12/2020.04.30.20085316.

35 Jonas Dehning et al. "Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions". In: *Science* (2020).

36 Mark J Siedner et al. "Social distancing to slow the U.S. COVID-19 epidemic: an interrupted time-series analysis". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (Apr. 2020). DOI: 10.1101/2020.04.03.20052373. URL: https://www.medrxiv.org/content/10.1101/2020.04.03.20052373v2.

37 Moritz U. G. Kraemer et al. "The effect of human mobility and control measures on the COVID-19 epidemic in China". In: *Science* 368.6490 (Mar. 2020), pp. 493–497. DOI: 10.1126/science.abb4218.

38 Adam J Kucharski et al. "Early dynamics of transmission and control of COVID-19: a mathematical modelling study". In: *The Lancet Infectious Diseases* 20.5 (May 2020), pp. 553–558. DOI: 10.1016/s1473-3099(20)30144-4.

39 Raj Dandekar and George Barbastathis. "Neural Network aided quarantine control model estimation of global Covid-19 spread". In: (Apr. 2, 2020). arXiv: 2004.02752v1 [q-bio.PE]. URL: https://arxiv.org/abs/2004.02752.

40 Benjamin F. Maier and Dirk Brockmann. "Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China". In: *Science* 368.6492 (Apr. 2020), pp. 742–746. DOI: 10.1126/science.abb4557.

41 Sofia B Villas-Boas et al. *Are We #StayingHome to Flatten the Curve?* Tech. rep. UC Berkeley: Department of Agricultural and Resource Economics, 2020.

42 Christopher I. Jarvis et al. "Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK". In: *BMC Medicine* 18.1 (May 2020). DOI: 10.1186/s12916-020-01597-8.

43 Luis Orea and Inmaculada Álvarez. *How effective has been the Spanish lockdown to battle COVID-19? A spatial analysis of the coronavirus propagation across provinces*. Working Papers 2020-03. FEDEA, 2020. URL: http://documentos.fedea.net/pubs/dt/2020/dt2020-03.pdf.

44 Lars Lorch et al. "A Spatiotemporal Epidemic Model to Quantify the Effects of Contact Tracing, Testing, and Containment". In: (Apr. 15, 2020). arXiv: 2004.07641v2 [cs.LG]. URL: https://arxiv.org/abs/2004.07641.

45 Marino Gatto et al. "Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures". In: *Proceedings of the National Academy of Sciences* 117.19 (Apr. 2020), pp. 10484–10491. DOI: 10.1073/pnas.2004978117.

46 Billy J Quilty et al. "The effect of inter-city travel restrictions on geographical spread of COVID-19: Evidence from Wuhan, China". In: *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv* (2020). DOI: 10.1101/2020.04.16.20067504. eprint: https://www.medrxiv.org/content/early/2020/04/21/2020.04.16.20067504.full.pdf. URL: https://www.medrxiv.org/content/early/2020/04/21/2020.04.16.20067504.