The effectiveness and perceived burden of nonpharmaceutical interventions against COVID-19 transmission: a modelling study with 41 countries

Jan M. Brauner MD[@,*,a,b], Sören Mindermann[*,a], Mrinank Sharma[*,c], Anna B. Stephenson[d], Tomáš Gavenčiak PhD[e], David Johnston[f,g], John Salvatier[e], Gavin Leech[h], Tamay Besiroglu[i], George Altman MBChB[j], Hong Ge PhD[k], Vladimir Mikulik[e], Meghan Hartwick PhD[l], Prof Yee Whye Teh PhD[m], Prof Leonid Chindelevitch PhD[n], Prof Yarin Gal PhD[+,a], Jan Kulveit PhD[+,b]

[a]*OATML, Department of Computer Science, University of Oxford, Oxford, United Kingdom*
[b]*Future of Humanity Institute, University of Oxford, Oxford, United Kingdom*
[c]*Department of Computer Science, University of Oxford, Oxford, United Kingdom*
[d]*Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA*
[e]*Independent researcher*
[f]*College of Engineering and Computer Science, Australian National University, Australia*
[g]*Quantified Uncertainty Research Institute, California, USA*
[h]*School of Computer Science, University of Bristol, Bristol, United Kingdom*
[i]*Faculty of Economics, University of Cambridge, Cambridge, United Kingdom*
[j]*School of Medical Sciences, University of Manchester, Manchester, United Kingdom*
[k]*Engineering Department, University of Cambridge, Cambridge, United Kingdom*
[l]*Tufts Initiative for the Forecasting and Modeling of Infectious Diseases, Tufts University, Boston, USA*
[m]*Department of Statistics, University of Oxford, Oxford, United Kingdom*
[n]*Computational Epidemiology Lab, School of Computing Science, Simon Fraser University, Burnaby, Canada*

Abstract

Background. Existing analyses of nonpharmaceutical interventions (NPIs) against COVID-19 transmission have concentrated on the joint effectiveness of large-scale NPIs. With increasing data, we can move beyond estimating joint effects towards disentangling individual effects. In addition to effectiveness, policy decisions ought to account for the burden placed by different NPIs on the population.

Methods. To our knowledge, this is the largest data-driven study of NPI effectiveness to date. We collected chronological data on 9 NPIs in 41 countries between January and April 2020, using extensive fact-checking to ensure high data quality. We infer NPI effectiveness with a novel semi-mechanistic Bayesian hierarchical model, modelling both confirmed cases and deaths to increase the signal from which NPI effects can be inferred. Finally, we study

[@]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

how much perceived burden different NPIs impose on the population with an online survey of preferences using the MaxDiff method.

Results. Eight NPIs have a >95% posterior probability of being effective: closing schools (mean reduction in R: 50%; 95% credible interval: 39%–59%), closing nonessential businesses (34%; 16%–49%), closing high-risk businesses (26%; 8%–42%), and limiting gatherings to 10 people or less (28%; 8%–45%), to 100 people or less (17%; -3%–35%), to 1000 people or less (16%; -2%–31%), issuing stay-at-home orders (14%; -2%–29%), and testing patients with respiratory symptoms (13%; -1%–26%). As validation is crucial for NPI models, we performed 15 sensitivity analyses and evaluated predictions on unseen data, finding strong support for our results. We combine the effectiveness and preference results to estimate effectiveness-to-burden ratios.

Conclusions. Our results suggest a surprisingly large role for schools in COVID-19 transmission, a contribution to the ongoing debate about the relevance of asymptomatic carriers in disease spread. We identify additional interventions with good effectiveness-burden tradeoffs, namely testing symptomatic individuals, closing high-risk businesses, and limiting gathering size. Closing most nonessential businesses and issuing stay-at-home orders impose a high burden while having limited additional effect.

---

1. Introduction

The governments of the world have mobilized vast resources to fight the COVID-19 pandemic. A wide range[1] of nonpharmaceutical interventions (NPIs) has been deployed, including drastic measures like national lockdowns and the closure of all non-essential businesses. Recent analyses show that these large-scale NPIs appear to be jointly effective at reducing the virus' effective reproduction number.[2] As time progresses, more data becomes available from different countries that have implemented various NPIs (Figure 1). We can thus move beyond estimating the aggregate effect of a bundle of NPIs and understand the effect of individual NPIs.

But, selecting the right policy depends on more than estimates of effectiveness. Drastic NPIs, such as society-wide social distancing, cause widespread disruption to many aspects of social life, including quality of life, economic prospects,[3] and, potentially, the mental health of the entire population.[4] When selecting policies, it is thus important to consider the burden they impose.

This paper's aim is to estimate the effectiveness of various NPIs at reducing the spread of COVID-19 and their associated burden on the population.

To disentangle individual NPI effects, we need to leverage data from multiple regions with diverse bundles of NPIs. With some exceptions (Flaxman et al.[2], Chen and Qiu[5], and Banholzer et al.[6]), previous data-driven studies focus on single NPIs and/or single regions (Table 1). In contrast, we evaluate the impact of 9 NPIs on the growth of the epidemic in 34 European and 7 non-European countries. To our knowledge, this is the largest data-driven model of NPI effects on COVID-19 transmission to date. Additionally, the focus of previous work has largely been on costly NPIs (Table 1). In line with our aim of identifying effective interventions with little burden, we additionally analyse the effects of several less disruptive NPIs (Table 2).

Before collecting data, we experimented with two public datasets on NPIs, finding that they contained some incorrect dates and were not complete enough for our modelling.[a] By focusing on a smaller set of countries and NPIs than is present in these datasets, we were able to implement strong quality controls in our data collection. We make this high-quality dataset public, as well as the Epidemic Forecasting Global NPI database, a much larger but less rigorously verified dataset.

[a] We evaluated the following datasets:

- Oxford COVID-19 Government Response Tracker (OxCGRT)[7]
- #COVID19 Government Measures Dataset[8]

Note that these datasets are under continuous development. Many of the mistakes we found will already have been corrected. Also, we know from our own experience that data collection can be very challenging. We have the fullest respect for the work of the people behind these datasets. In this paper, we focus on a much 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.

To estimate NPI effectiveness, we design a novel semi-mechanistic Bayesian hierarchical model with a time-delayed effect for each NPI. A key assumption of our model is that the effect of each NPI on the reproduction number is stable across different countries and over time. This assumption is present in all closely related works. Our model can be seen as an extension to that of Flaxman et al.,[2] using both confirmed cases and deaths as observations to increase the amount of signal available for inferring NPI effects.

Constructing an NPI model is a perilous task since its conclusions can be sensitive to the assumptions and data. Therefore, it is crucial to validate it. However, such validation is often incomplete or absent from previous work. We perform what is, to our knowledge, by far the most extensive validation of any NPI model for COVID-19 to date—evaluating predictions for countries and time periods not seen during training (Figures 4 & 5), evaluating different models that use different observations (deaths and confirmed cases; Figure 6), testing robustness to unobserved NPIs (Figure D.10), and analyzing sensitivity to many perturbations (Appendix D). Nonetheless, our model comes with important limitations and uncertainties, which we discuss in Appendix H.

Finally, to study how burdensome people perceive different NPIs to be, we collected preference data using a best-worst scaling[9] discrete choice online survey instrument. As community surveys are often successfully used in public health settings to estimate the preferences over various treatments and interventions,[10] we believe this data can provide valuable input when evaluating NPIs. While there are many other ways to estimate NPI cost, for example by modelling economic impacts, these are often dominated by long-term effects. For example, a large part of the economic impact of closing schools could consist in human-capital loss.[11] These long-term effects are currently hard to predict and are codetermined by economic policy responses and many other effects beyond the scope of this study.

Summary of contributions:

- High-quality data on the largest number of countries and NPIs studied to date, including several less costly NPIs
- A novel combined model utilising both confirmed cases and deaths
- Extensive model validation
- Estimation of population preferences over NPIs and analysis of effectiveness-burden tradeoffs

Table 1: 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 (ICL report #13)[2]	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
Chen and Qiu, 2020[5]	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[6]	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, and Switzerland	Semi-mechanistic Bayesian hierarchical model
Choma et al., 2020[12]	Single aggregated NPI	22 countries and 25 states	Regression with Susceptible-Infectious-Removed-Deceased (SIRD) model
Siedner et al. 2020[13]	General social distancing	U.S.	Interrupted time-series
Kraemer et al., 2020[14]	Travel restrictions and cordon sanitaire	China	Regression
Kucharski et al., 2020[15]	Travel restrictions	Wuhan (China)	Various, including Susceptible-Exposed-Infectious-Removed (SEIR) model
Dandekar and Barbastathis, 2020[16]	General quarantine and isolation	Wuhan, Italy, South Korea, and U.S.	A mix of a mechanistic model and a data-driven neural network model

Continued on next page

Table 1 – *Continued from previous page*

Study	NPIs studied	Regions/countries studied	Method
Maier and Brockmann, 2020[17]	General quarantine and isolation	Mainland China	Quantitative fits to empirical data
Sears et al., 2020[18]	Mobility changes as a proxy for stay-at-home mandates	U.S.	Difference-in-differences statistical model
Jarvis et al., 2020[19]	Physical (social) distancing measures	UK	Questionnaire data and compartmental epidemic model
Orea and Álvarez, 2020[20]	Lockdown	Spain	Spatial econometric analysis
Lorch et al., 2020[21]	Mobility restrictions, testing & tracing, social distancing, and business restrictions	Tübingen (Germany)	Authors' own spatiotemporal model of epidemics
Gatto et al., 2020[22]	Various restrictions to mobility and human-to-human interactions	Italy	Susceptible–Exposed–Infected–Recovered (SEIR)-like disease transmission model
Quilty et al., 2020[23]	Intercity travel restrictions	Beijing, Chongqing, Hangzhou, and Shenzhen (Mainland China)	Branching process transmission model

2. Methods

2.1. Dataset

We collected a large database from 67 countries, which we call the Epidemic Forecasting Global NPI (EFGNPI) database. The database contains more than 1700 events, tagged with 194 keywords, which are distilled into 24 classes of NPIs. Details of the EFGNPI database are given in Appendix B.

As described in the introduction, we found that public datasets on NPIs contained frequent incorrect entries. We expect the same to be true for the full EFGNPI database. For the smaller set of NPIs and countries used in this study, we implemented further steps to ensure data quality (see below). The data used in this study, including sources, can be found at https://github.com/robust-npis/covid-19-npis.

Table 2: NPIs included in the modelling dataset

NPI	Description
Mask-wearing	One or both of: • a country has implemented a policy of requiring mask usage among the general public, sometimes limited to certain domains like a duty to wear masks in public transportation and supermarkets • survey reports indicate that over 60% of people were wearing masks in public.
Symptomatic testing[d]	Testing is available to anyone showing COVID-19 symptoms (as defined by the country). In a few countries, testing is even available to people without symptoms.
Gatherings limited to 1000 people or less	A country has set a size limit on gatherings. The size 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 size limit is at most 100 people (often less) and gatherings above the maximum size are disallowed.
Gatherings limited to 10 people or less	A country has set a size limit on gatherings. The size limit is at most 10 people (often less) and gatherings above the maximum size are disallowed.
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, 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 many or all schools. (Note that this was accompanied by closing universities in more than 75% of cases in our data.)
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.

[d]Feature taken from the Oxford COVID-19 Government Response Tracker[7]

We analyse 41 countries[c] (see Figure 1) and 9 NPIs (Table 2). We only recorded when NPIs were implemented in most of a country. The window of analysis spans the period from 22nd January to 25th April 2020[d], inclusive. Data on confirmed COVID-19 cases and deaths were taken from the John Hopkins Center CSSE COVID-19 Dataset[24,25].

Data collection

Gathering bans, school closure, business closure, stay-at-home order

For each NPI and each country, one to three contractors independently collected data on the start date of the NPI, including sources. Each country was then extensively researched by one of the authors, using media articles, government sources, and Wikipedia articles. The researcher finalised the data based on their research, the data in the EFGNPI dataset, the data provided by the contractors, and, if available, data from the Oxford COVID-19 Government Response Tracker.[7]

Mask-wearing

To estimate the local prevalence of mask-wearing, we conducted surveys of n=908 participants from most of the countries studied. Respondents were asked about the number of people they had seen wearing masks (details in Appendix C). We also used Wikipedia and the masks4all dataset[26] to ascertain when countries mandated mask-wearing in (some) public places. In all countries in which the government mandated mask-wearing, our survey results indicate that more than 60% of people started wearing masks around the time when the mandate was implemented.

Testing

The Oxford COVID-19 Government Response Tracker[7] has complete data on testing policies implemented in different countries. To check its accuracy, we compared the data with the number of tests per confirmed case[27] and found that activation of the testing feature was correlated with a substantial increase in the number of tests per confirmed case. We did not do further verification. As of version 5.0 of the dataset, our "symptomatic testing" feature corresponds to the following feature in the OxCGRT dataset: ID H2, levels 2-3.

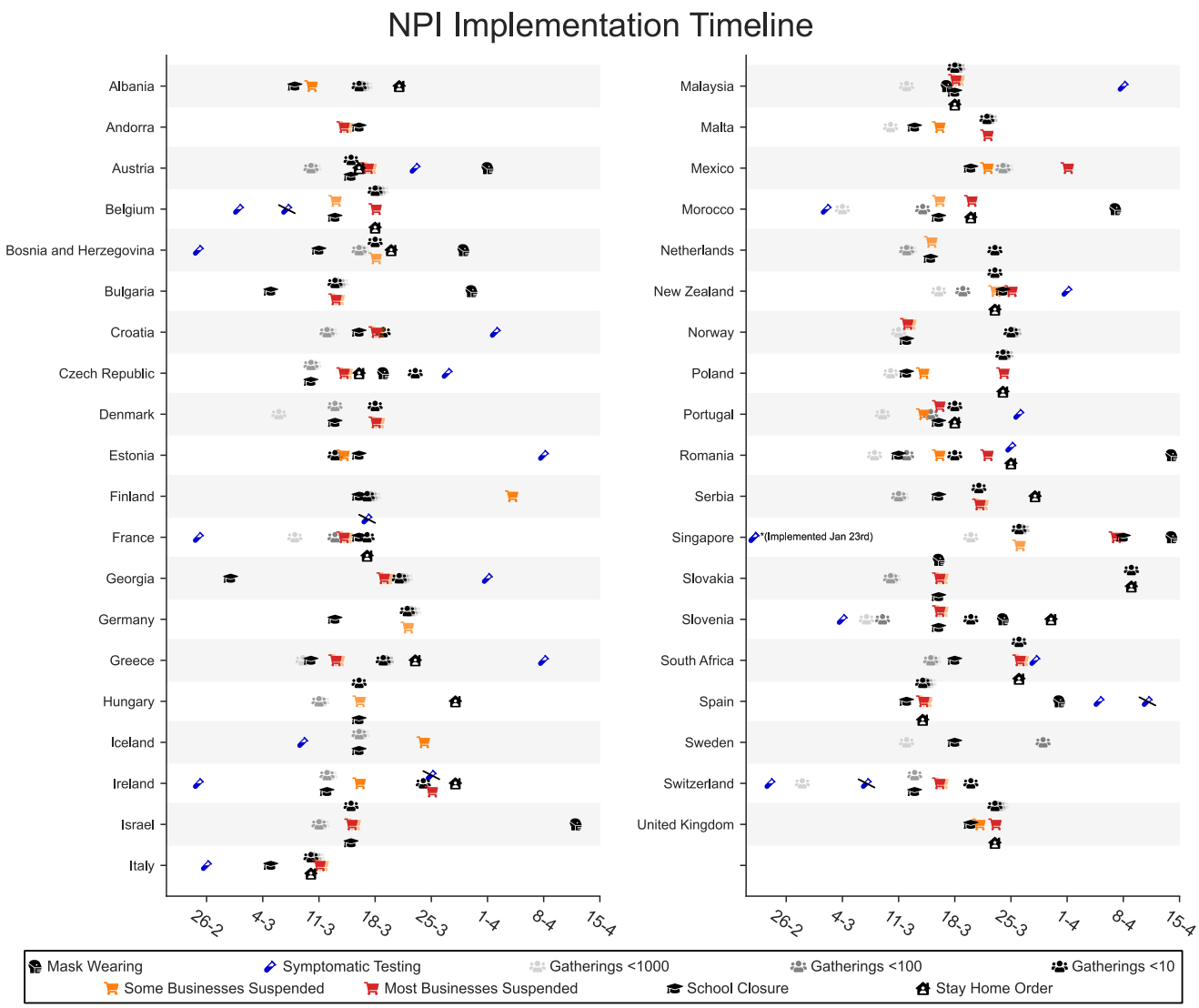

Figure 1: Timing of NPI implementations. Crossed-out symbols signify when an NPI was lifted.

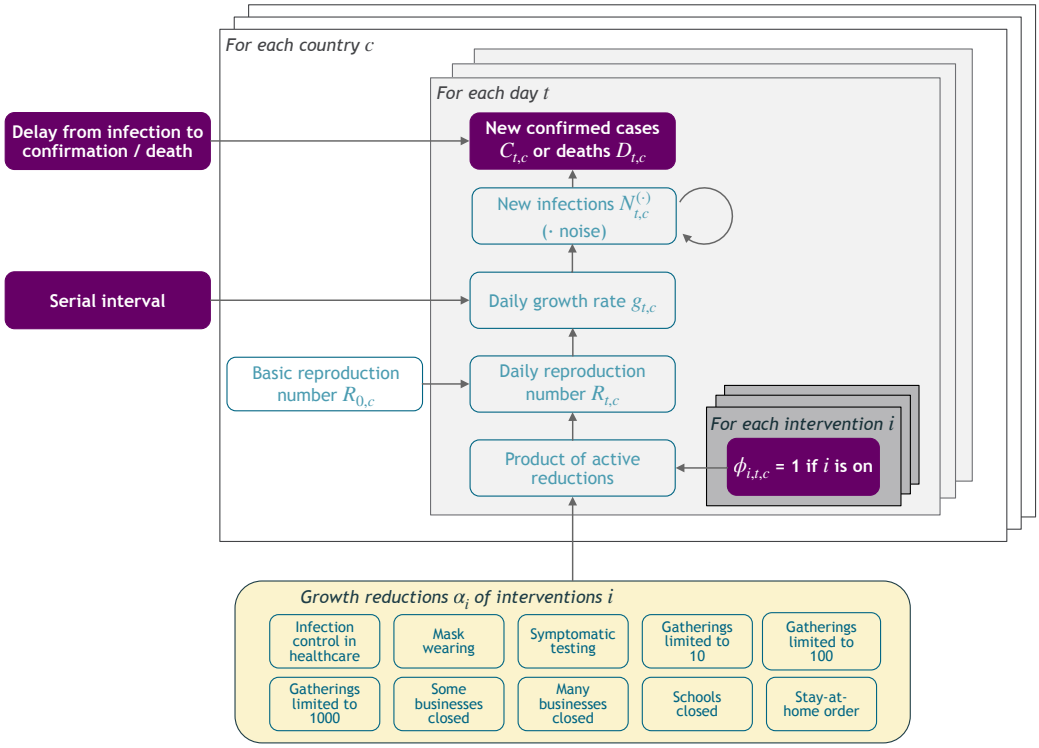

Figure 2: Model Overview. Purple nodes are observed or have a fixed distribution. The same structure is used for both deaths and confirmed cases. Our primary model combines both observations; it splits all nodes above the daily growth rate $g_{t,c}$ into separate branches for deaths and cases.

2.2. Model

We construct a semi-mechanistic Bayesian hierarchical model, similar to Flaxman et al.[2] 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 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 make minimal assumptions about testing, reporting, and the infection fatality rate (IFR). Please see Appendix F for further details.

We describe the model in Figure 2 from bottom to top. The growth of the epidemic is determined by the time-and-country-specific 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

[c]The countries were selected by a case threshold (at the time of modelling), the availability of reliable data on NPIs, and how trustworthy we estimated the reporting of deaths from this country to be. Some particular countries were excluded for specific reasons. For example, we excluded South Korea because the country made heavy use of contact tracing which we don't model (because data on contact tracing is very hard to get).

[d]22nd January - 17th April for confirmed cases

estimates of R_0.[28] We parameterize the effectiveness of NPI i, assumed to be similar across countries and time, with α_i. The effect of each NPI on $R_{t,c}$ is assumed to be multiplicative (and therefore independent) as follows:

$$R_{t,c} = R_{0,c} \exp\left(-\sum_{i=1}^{9} \alpha_i \phi_{i,t,c}\right),$$

where $\phi_{i,c,t} = 1$ means NPI i is active in country c on day t ($\phi_{i,c,t} = 0$ otherwise). In Section 3, we discuss this interaction between NPIs. There is a symmetric prior (and hyperprior) over α_i, allowing for both positive and negative effects.

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[e] growth rate $g_{t,c}$. During exponential growth, there is a well-known one-to-one correspondence between $g_{t,c}$ and $R_{t,c}$:[29]

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

where $M(\cdot)$ is the moment-generating function of the distribution of the serial interval (the time between successive cases in a chain of transmission). We assume that the serial interval distribution is given by a Gamma(5.18, 0.96)[f] distribution[30]. Using (1), we can write $g_{t,c}$ as $g_{t,c}(R_{t,c})$ (see Appendix F).

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

$$N^{(C)}_{t,c} = N^{(C)}_{0,c} \prod_{\tau=1}^{t} \left[(1 + g_{\tau,c}) \cdot \exp\left(\epsilon^{(C)}_{\tau,c}\right)\right] \tag{2}$$

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

where $\epsilon^{(\cdot)}_{\tau,c} \sim \mathcal{N}(0, \sigma_N = 0.2)$ are separate, independent noise terms. We seed our model with unobserved initial values, $N^{(C)}_{0,c}$ and $N^{(D)}_{0,c}$, which have uninformative priors.[g]

[e] 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[29] to account for our choice.

[f] The two parameters are the shape and rate. The mean is 5.1 days.

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

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

$$\overline{C}_{t,c} = \sum_{\tau=1}^{t} N^{(C)}_{t-\tau,c} P_C(\text{delay} = \tau) \tag{4}$$

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 consistent empirical distributions from China and Italy,[31–34] 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.[2]

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[31,35], which sum up to a mean delay of 23.9 days. Finally, the observed deaths $D_{t,c}$ also follow a negative binomial distribution with mean $\overline{D}_{t,c}$ and an inferred dispersion parameter.

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 the infected cases 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 $N^{(C)}_{t,c}$ (Eq. (4)) but *not* in the growth rates $g_{t,c}$ across A and B (Eq. (2)-(3)). Accordingly, the inferred NPI effectiveness will be identical.[h]

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 effects of NPIs: decreasing the ascertainment rate decreases future cases $C_{t,c}$ by a constant factor whereas

[h] 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 $N^{(C)}_{0,c}$ could break scale invariance, the uninformative prior results in a negligible effect.

the introduction of an NPI decreases them by a factor that grows exponentially over time.[i] The noise terms, $\exp\left(\epsilon_{\tau,c}^{(C)}\right)$ (Eq. (2)), mimic changes in the ascertainment rate—noise at time τ affects all future cases—and allow for gradual, multiplicative changes in the ascertainment rate.

We infer the unobserved variables in our model using Hamiltonian Monte-Carlo[36,37] (HMC), a standard MCMC sampling algorithm.

The model code can be found at https://github.com/robust-npis/covid-19-npis.

2.3. Preference elicitation

We collected preference data to study the direct impact of NPIs on people's lives. We used a best-worst scaling discrete choice survey instrument, specifically MaxDiff,[9] and surveyed $N = 474$ US residents recruited on Amazon's Mechanical Turk platform. The platform typically yields participants with greater demographic diversity than typical internet samples.[38] Note that this survey was entirely separate from the survey used for studying mask-wearing described above.

Each respondent was given a short description of all studied NPIs (Appendix G) and then presented with 12 MaxDiff questions with 6 options, where each option consisted of a type of NPI and a duration (1 week, 2 weeks, 1 month, 3 months, 6 months, 1 year). Participants were asked to select the two options that they perceived as overall least and most burdensome (example question in Appendix G).

Before analysis, 140 responses with inconsistent answers were discarded; we considered answers erroneous when they preferred a longer duration of an intervention (often this happened for participants who responded quickly). To extract utility scores, we used the analytical estimation for the multinomial logit model,[39] as implemented in the bwsTools package[40] in R.

2.4. Effectiveness-Burden-Ratio

To analyse how the effectiveness of NPIs compares to their social impact, we can use the utility scores derived from the survey responses. However, utility scores are on an interval scale, because the survey only asks for relative comparisons between options.[41] While respondents presumably dislike all choices, we cannot say that, for example, a stay-at-home order is three times worse than school closure.

[i] 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 H.

To estimate the effectiveness-burden-ratio, we need to estimate a measure for the intervention burden on a ratio scale, which we call "perceived intervention costs". These can be derived from the utility scores with additional assumptions, which are well justified by the empirical data (Figure 7, details in Appendix G).

With these, the effectiveness-burden-ratio EBR_i of intervention i can be defined as:[j]

$$EBR_i = -\frac{\ln(m_i)}{c_i}$$

where m_i is the multiplicative factor on R (e.g., for a 20% reduction in R, $m_i = 0.8$), and c_i is the cost of intervention i. To determine the error of EBR_i, we used error propagation:[42]

$$V(EBR_i) = \frac{V(m_i)}{(m_i \cdot c_i)^2} + \frac{\ln(m_i)^2 \cdot V(c_i)}{c_i^4}$$

where $V(\cdot)$ is the variance.

2.5. Ethics

The online survey experiments were approved by the Medical Sciences Interdivisional Research Ethics Committee at the University of Oxford (Ethics Approval Reference: R69410/RE001)

2.6. Role of the funding source

The funding source did not influence any aspect of study design, execution, or reporting.

3. Results

3.1. International timeline of NPI implementation

We aim to estimate the effectiveness of individual NPIs. If all countries implemented the same set of NPIs, on the same day, the individual effect of each NPI would be unidentifiable. However, many countries implemented different sets of NPIs, at different times, in different orders (Figure 1).

[j] This particular functional form is chosen because it is a simple expression that satisfies three desirable properties:

i) repeated application of an intervention x times that has effectiveness factor m and a constant unit cost c has equal effectiveness-burden-ratio each time it is applied. Formally: for any $c \in \mathbb{R}^+$ and any $m \in (0,1)$, we have that $f(m^x, xc) = f(m,c)$ for any $x \in \mathbb{Z}$.

ii) it is increasing in m

iii) it is decreasing in c

3.2. Model fits

The model fits the observations well in 3 randomly selected countries (Figure 3, left). The fits for all other countries can be found in Appendix E. Plotting posterior values of the noise terms $\epsilon_t^{(C)}$ and $\epsilon_t^{(D)}$ shows periods where infections grew faster or slower than predicted based on the active NPIs, illustrating where the model might account for unobserved interventions or changes in reporting.

3.3. Held-out data experiments

An important way to validate a Bayesian model is by checking its predictions on held-out data.[43] Our model makes sensible, calibrated forecasts over long periods in countries whose data was not used to infer the effectiveness of NPIs (Figure 4, see Appendix E for other countries).

We additionally validate our model's predictions by holding out the last 20 days of both new cases and deaths for *all* countries. These are challenging predictions; the longest attempted period we found in related work was 3 days.[2] The accurate forecasts in Figure 5 provide strong empirical evidence that our estimates of R are plausible.

3.4. Effectiveness per NPI

The estimates of NPI effectiveness are our main result. To interpret them correctly, we need to keep in mind that our model assumes no interaction between different NPIs. In our model, each NPI reduces R by a multiplicative factor, independent of the *context*, i.e., the presence of other NPIs. This independence assumption is present in all multi-NPI studies we are aware of and seems reasonable for many NPIs. For instance, the effectiveness of closing businesses is likely to be similar whether or not schools are closed. However, in some situations, the effectiveness of an NPI might depend on its context. For example, if a stay-at-home order is in place, a larger fraction of the remaining transmission might occur in private spaces, and wearing masks in public spaces might be less effective.

Given this discussion, the effectiveness estimates should not be interpreted as the average effectiveness across all possible contexts, but rather as the (additional) **effectiveness averaged across the contexts in which the NPI was present in our data.** This result, which is equally important for the interpretation of other related studies, is derived for a simplified model in Appendix F.3. Figure 6 (bottom left) visualises the contexts of each NPI in our data, aiding interpretation.

Figure 6 shows the estimates of NPI effectiveness. Reassuringly, our three models have similar results. This suggests that results are not biased by factors that are specific to the deaths or cases model, such as changes in the ascertainment rate, reporting, and model-specific time delays. All NPIs except mask-wearing had a >95% posterior probability of being effective.

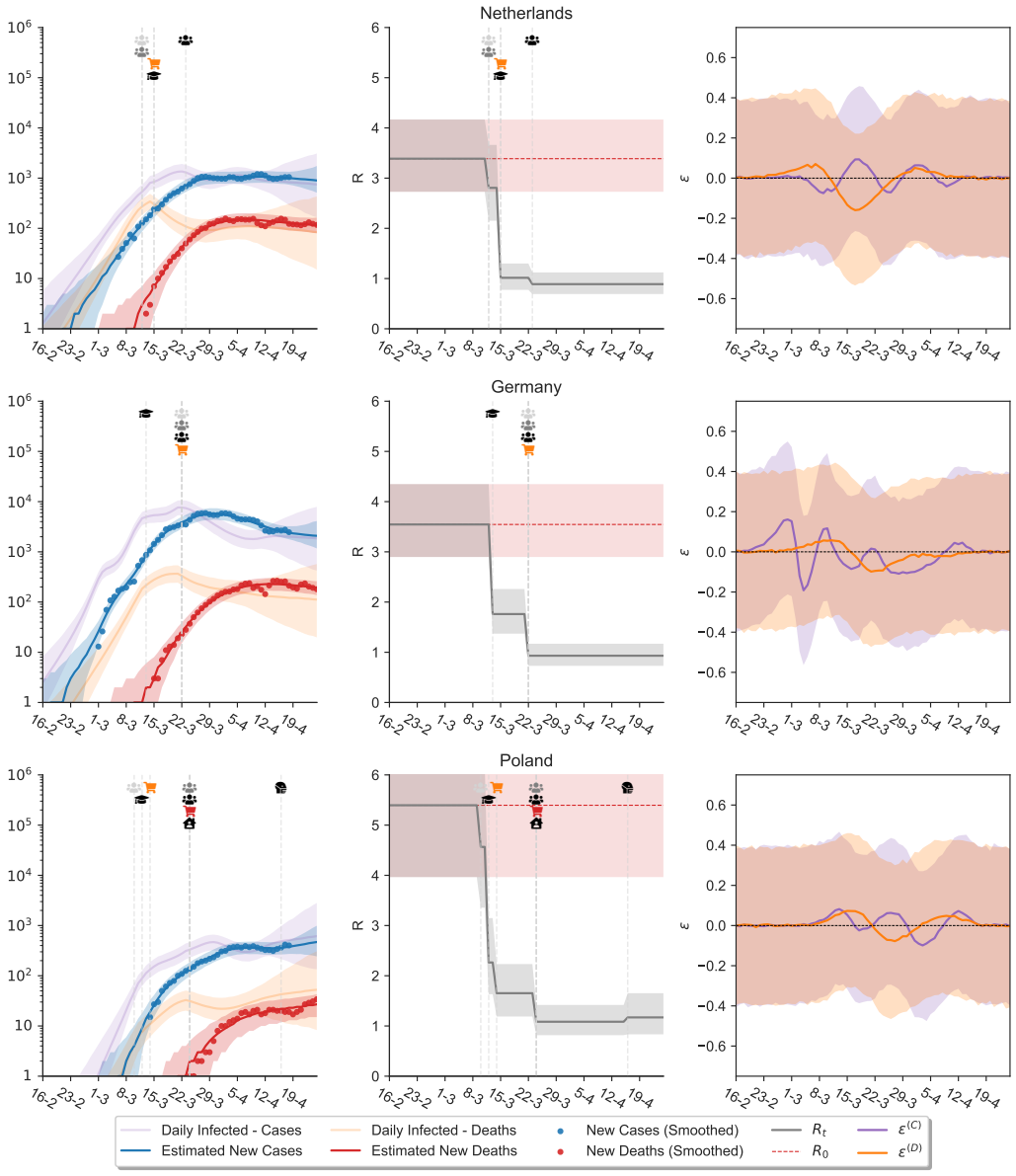

Figure 3: Model fits for 3 randomly selected countries. Vertical lines show the activation of NPIs. Shaded areas are 95% credible intervals. *Left:* Country-level estimates of daily new infections $N_t^{(C)}$ and $N_t^{(D)}$, smoothed confirmed cases C_t, and deaths D_t. Note that the curves show the fit to data, and not epidemiological forecasts. *Middle:* Estimates of reproduction numbers. *Right:* Inferred noise $\epsilon_t^{(C)}$ and $\epsilon_t^{(D)}$ on new infections. Values above zero indicate that infections grew faster than predicted solely based on the active NPIs.

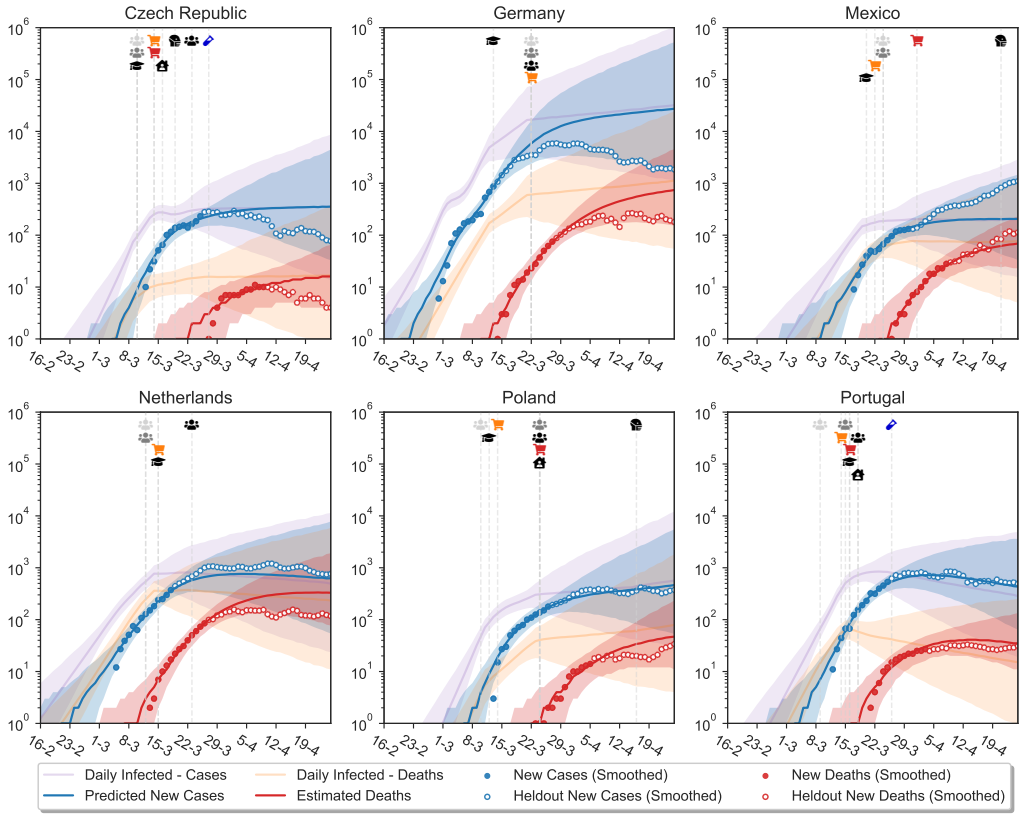

Figure 4: Predictions for held-out countries. We randomly selected 6 countries with >100 deaths. Empty dots are not shown to the model. 14 initial days are shown to the model, to enable inferring the country-specific R_0.

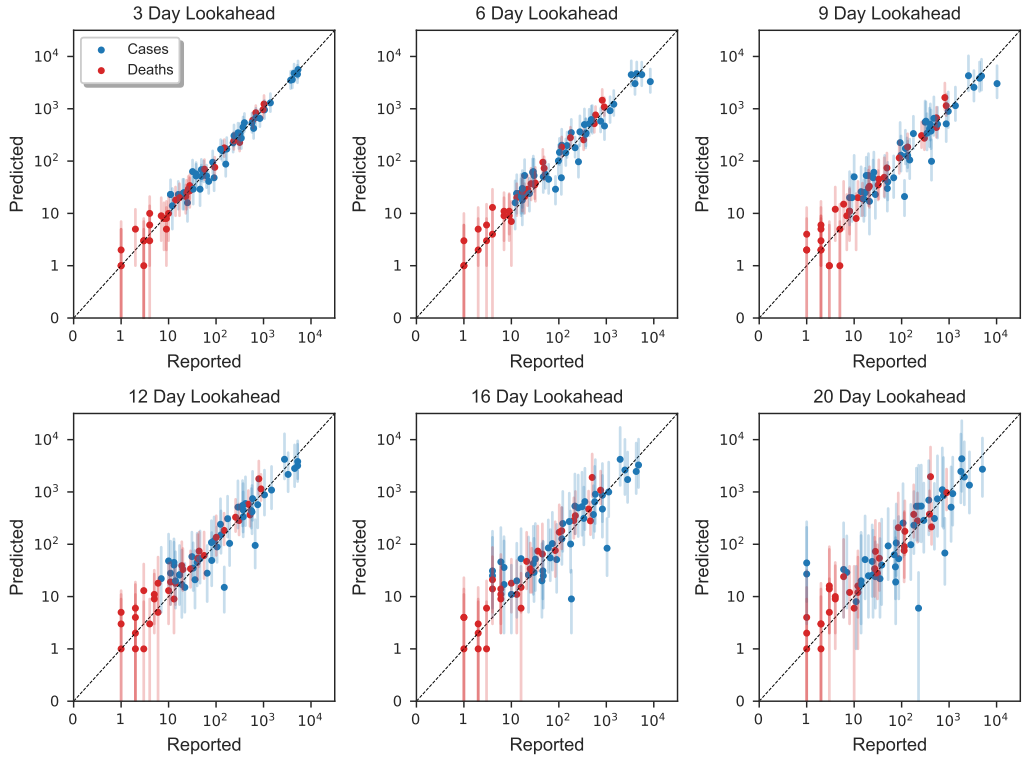

Figure 5: Predictions for held-out last 20 days. These results were obtained by holding out the last 20 consecutive days for all countries and predicting them (these days were not available to infer NPI effectiveness). Each point represents a country. The plot shows 95% sampled posterior credible intervals and the median predicted values.

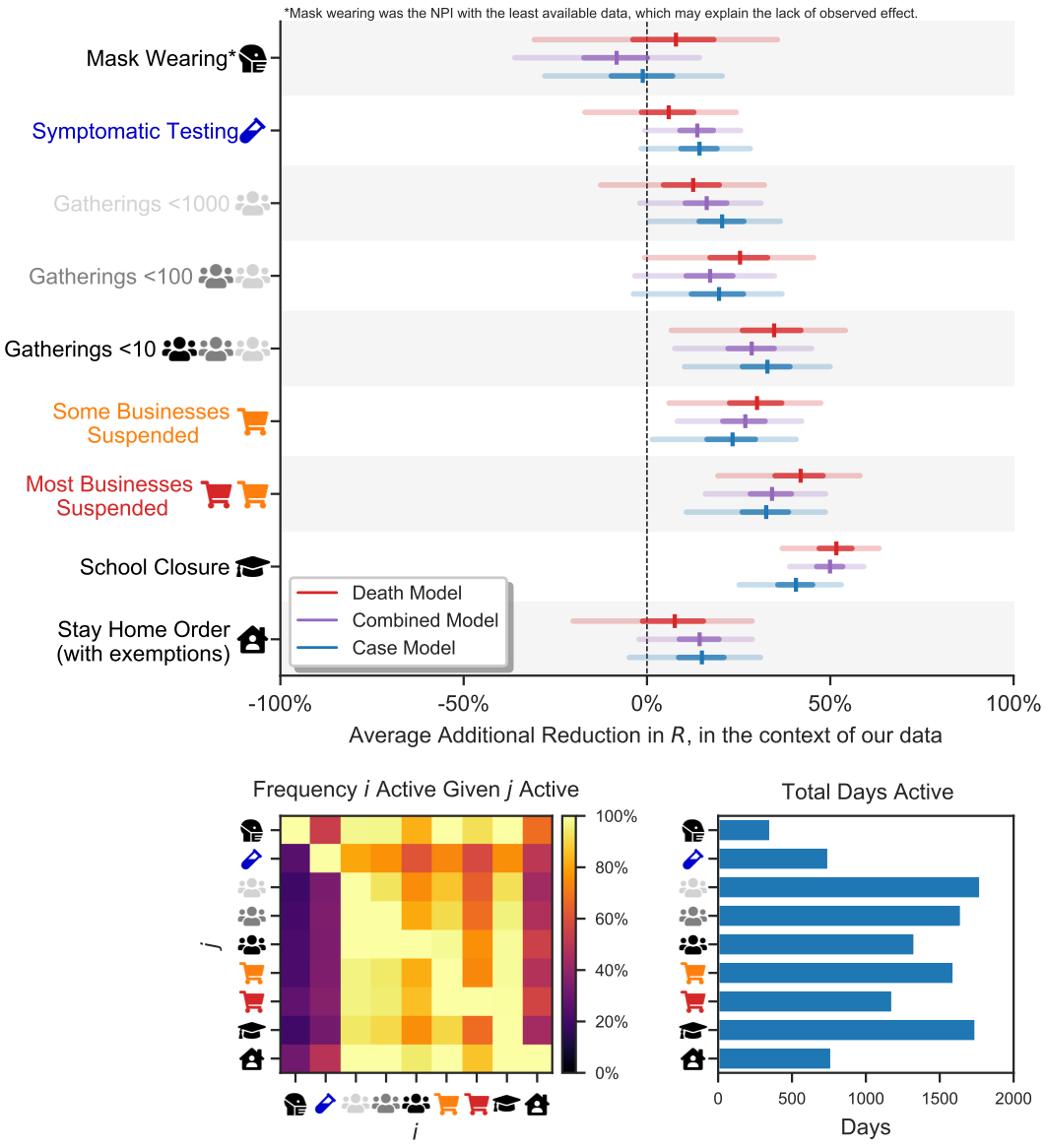

Figure 6: ***Top:*** **Posterior reduction in R for each NPI. The plot shows 50% and 95% credible intervals. A negative 1% reduction refers to a 1% increase in R. The following NPIs are hierarchical: gathering bans and business closures. For example, the result for *Most Businesses Suspended* shows the cumulative effect of two NPIs with separate parameters and symbols: suspending some (high-risk) businesses, and suspending most remaining (non-high-risk, but nonessential) businesses. The exact numbers are given in Appendix A.** ***Bottom Left:*** **The conditional activation matrix shows the situations encountered in our data. Cell values indicate the *frequency* that NPI i (x-axis) is active given that NPI j was active(y-axis) e.g., schools were closed whenever a stay-home-order had been issued (bottom row, second column from the right), but not vice versa.** ***Bottom Right:*** **Total number of days each NPI was active across countries.**

We confirmed the quality of the MCMC inference with the Gelman-Rubin convergence statistic[44] (Appendix E).

3.5. Sensitivity experiments

We ran a wide range of sensitivity experiments on our combined model. Appendix D shows effectiveness-per-NPI plots for the many conditions we tested. Table 3 summarizes the results qualitatively. We diagnosed 'low - moderate' sensitivity when, for every NPI, all 95% credible intervals, but not all 50% intervals, overlap. 'Low' sensitivity means all 50% intervals overlap.

Results were stable, not affecting our conclusions.

Table 3: Sensitivity of effectiveness estimates. Summary of the results in Appendix D. 'Low - moderate' sensitivity means that all 95% credible intervals overlap for all NPIs. 'Low' means all 50% intervals overlap.

Sensitivity to	**Sensitivity**
Mobility data as 'NPI'	Mobility explains only the effect of business closures and stay-home-orders
Unobserved NPIs	Low, but moderate for removal of most effective NPI
Left-out countries	Low
Delay to confirmation	Low
Delay to death	Low
Standard deviation of the noise $\epsilon_{t,c}^{(\cdot)}$ on infections	Low - moderate
Dispersion of observation noise (deaths)	Low
Dispersion of observation (cases)	Low
Serial interval mean	Low - moderate
Minimum cumulative cases before which data is masked	Low
Prior over effectiveness	Low - moderate
Hyperprior over $R_{0,c}$	Low
Schools open/closed in Sweden (data ambiguity)	Low

Robustness to unobserved effects. The model assumes that there are no unobserved factors changing R (i.e., *unobserved confounders* such as spontaneous social distancing). But this is not necessarily true in practice. We test robustness to unobserved factors by computing NPI effectiveness whilst removing the observation of each NPI in turn. The sensitivity is low, supporting the claim that the model successfully unobserved factors.

Furthermore, we investigated robustness to unobserved confounding factors by including mobility data[45] as an 'NPI' that serves as a proxy for behaviour changes. We find that the mobility data explains the effect of business closures and stay-home-orders, which is expected as the effect of these NPIs is mediated through retail and recreation mobility. The inferred effectiveness of other NPIs is unchanged.

We do not report sensitivity to:

- The prior over the initial outbreak size $N_{0,c}$ (because it is already extremely wide, having a negligible effect)
- Alternative models of infection and NPI interaction

3.6. Preference elicitation

We surveyed 474 US residents recruited on Amazon's Mechanical Turk platform about their preferences regarding various NPIs using a best-worst scaling survey. 140 responses were filtered for internally inconsistent answers, and 334 were used for subsequent analysis (demographics in Appendix G). The NPI *Symptomatic testing* was not included in the preference elicitation because the mere option to get tested for Covid-19 when having symptoms does not impose any burden on people.

The ranking of the NPIs is largely independent of the duration (Figure 7). The duration-dependence of preferences is largest for mask-wearing, which is more preferable if required only briefly, and the most stringent interventions, stay-at-home orders and the closure of most nonessential businesses, which are perceived as particularly bad if implemented for unrealistically long durations. Table A.4 displays the aggregate utility scores across all durations.

3.7. Effectiveness-Burden-Tradeoff

Figure 8 compares the effectiveness of different NPIs to survey participants' preferences. With some further assumptions (see Section 2.4), we can convert the utility scores to a ratio-scaled measure of intervention burden and calculate an effectiveness-burden-ratio for every NPI (Figure 9).

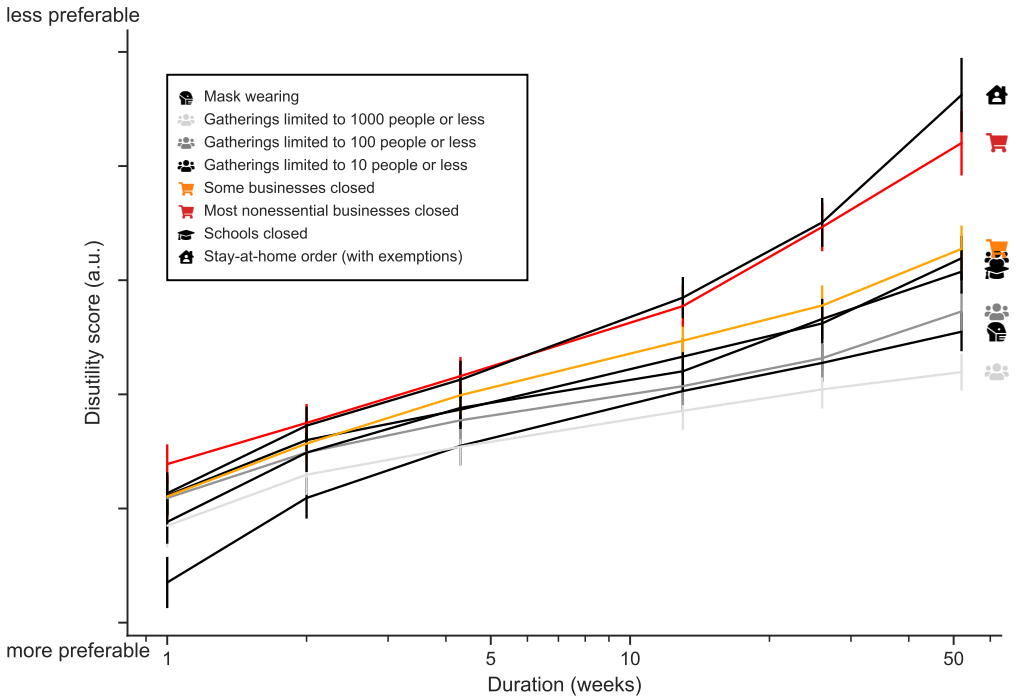

Figure 7: NPI disutility scores in dependence of the duration of the intervention. Lower disutility implies higher utility and a stronger preference. Utilities are on an interval scale, the absolute values have no significance, only differences between utilities carry meaning. The error bars indicate the 95% confidence interval.

4. Discussion

We find evidence for the effectiveness of several NPIs. The conclusions discussed here were robust across 15 sensitivity analyses.

Combining effectiveness estimates with results from preference surveys, we can draw interesting conclusions:

- Closing high-risk businesses, such as bars and restaurants, appears only slightly less effective than closing most nonessential businesses, while imposing a substantially smaller burden.
- There is no obvious best choice for gathering-size restrictions: though stricter limits are more effective, they are more burdensome, giving a similar effectiveness-burden ratio.

We now discuss some of the main or more surprising results in detail.

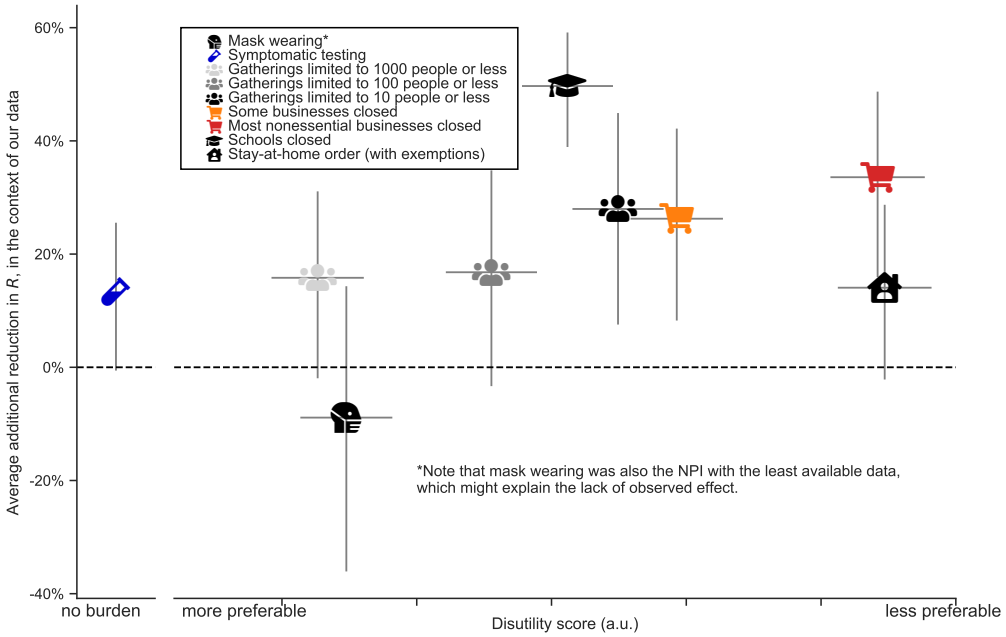

Figure 8: Effectiveness of NPIs compared to survey participants' preferences. The combined confirmed cases + deaths model was used for the effectiveness estimates. The dashed line represents no effect. Error bars indicate 95% credible/confidence intervals. The NPI *Symptomatic testing* was not included in the preference survey because the mere option to get tested for Covid-19 does not impose any burden on people. It is thus shown here on a separate x-axis.

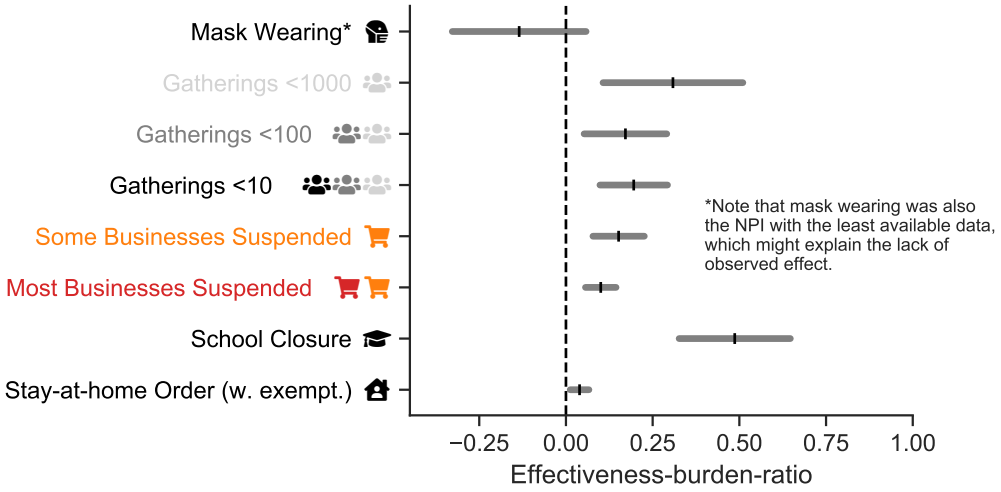

Figure 9: Effectiveness-burden-ratio of NPIs. The combined active cases + deaths model was used to generate the effectiveness estimates. The dashed line represents no effect. The error bars display the standard deviation. The definition of the effectiveness-burden-ratio is given in Section 2.4.

Testing. With no direct negative effects on the population and a demonstrable effect on transmission, testing of patients with respiratory symptoms looks very promising from an effectiveness-burden perspective.[k] Of course, the main negative effect of testing is the cost of purchasing and conducting tests. However, a recent economic analysis concluded that even testing asymptomatic people is vastly more cost-effective than indiscriminate measures.[46].

Stay-at-home-orders. We estimate a comparatively small effect for stay-at-home orders. The 'stay-at-home order (with exemptions)' NPI (Table 2) should be interpreted literally: a mandatory order to generally stay at home, except for exemptions. When countries introduced stay-at-home orders, they nearly always also banned gatherings and closed nonessential businesses and schools if they had not done so already (Figure 6). Accounting for the effect of these NPIs, it is not surprising that the additional effect of ordering citizens to stay at home is small-to-moderate. Accordingly, it may be acceptable to lift burdensome stay-home-orders, provided other NPIs stay active. Our result agrees with Banholzer et al.[6] (they call this NPI 'lockdown'), and we have not seen contradictory results in related work. In particular, the 'lockdown' NPI in Flaxman et al.[2] includes several other NPIs. Chen & Qui[5] found a significant effect, but without defining 'lockdown'.

Mask-wearing. Mask-wearing was often introduced towards the end of our analysis period (Figure 1), meaning that it is, by far, the NPI with the least data (Figure 6). We conclude that we have insufficient data to make claims about the effectiveness of mask-wearing, and indeed, in most of our sensitivity analyses, the result for mask-wearing was the least robust one (Appendix D). In particular, we do *not* conclude that mask-wearing is likely harmful. Additionally, mask-wearing might have a reduced effect in the context of the particular countries we studied. People started wearing masks when interactions in public spaces were already limited by other NPIs. When relatively more transmission occurs in private spaces, wearing masks *in public* is expected to be less effective. This might explain the difference to Chen & Qui,[5] who found a small significant effect of mask-wearing based on data from two countries (China and South Korea), as mask-wearing was common in South Korea before other NPIs were implemented.

School closures. All our models find a very large effect for school closures. This result is surprising, even when accounting for the fact that school closure usually coincided with university closure. However, the large effect was remarkably robust across our sensitivity analysis, different structural assumptions (e.g., about infection and NPI interaction - not reported) we implemented during our model checking process[47], and across a long process of collecting data for additional countries and NPIs. By inspecting the data and the in-

[k] Note that we did not directly measure the burden of testing because this is not possible in the framework of our preference analysis (Section 3.6)

ferred infections, it is easy to see why the effect is so large: school closures are consistently followed by a clear reduction in growth (after the appropriate delay).

It is possible that our model confuses the effect of closing schools and unobserved behaviour changes. However, our sensitivity analysis showed that results are fairly robust to unobserved NPIs, suggesting they are robust to unobserved factors. Furthermore, we directly modelled unobserved factors by introducing mobility data 'NPIs' as a proxy for them. Again, the effect of school closures was unchanged. While these techniques closely mirror well-established sensitivity checks for unobserved causal effects,[48,49] they, too, rely on assumptions.

A further concern is that school closures have a delayed effect on deaths and confirmed cases, since children are less likely to die or show symptoms than adults. However, the result is not sensitive to the mean delay we assume (Appendix D).

Additionally, since the closure of schools was often the first major NPI introduced (Figure 1), it may have caused public concern to increase, causing behaviour changes. We do not distinguish this indirect *signalling* effect from the direct effect (for any NPI). Conversely, reopening schools could also have a signalling effect.

Previous evidence relevant to school closures is mixed. Flaxman et al.[2] and Banholzer et al.[6] did not find a significant non-zero effect with their data (Banholzer et al. focused on primary schools). Limited data suggests that children are equally susceptible to infection but have a lower observed case rate than adults[50–52]—whether this is due to school closures remains unknown. There is insufficient data about transmission from children. However, viral shedding appears to be comparable across age groups.[53,54] Little is known about the attack rate in schools (since they are closed); the best-documented case found that 38.3% to 59.3% were infected in one French high school.[55] As our results suggest a large role of schools (and universities) in Covid-19 transmission, this topic deserves further study.

Our study is not without assumptions and limitations, which are discussed in greater detail in Appendix H. To highlight some important points: NPI effectiveness may vary across countries and time; we cannot quantify the influence of unobserved factors on our results; regional differences within countries complicate the analysis. Therefore, a high degree of uncertainty remains. Our results should not be seen as the final answer on NPI effectiveness and burdens, but rather as a contribution to a diverse body of evidence, next to other retrospective studies, experimental trials and clinical experience.

5. Acknowledgements

Survey participant compensation was funded by a grant from the Berkeley Existential Risk Initiative. 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. 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].

6. Declarations of interest

No conflicts of interests.

7. Authors' contributions

D Johnston, JM Brauner, J Kulveit, G Altman, G Leech designed and conducted the NPI data collection S Mindermann, M Sharma, JM Brauner, A Stephenson, H Ge, YW Teh, Y Gal, J Kulveit, T Gavenciak, J Salvatier, M Hartwick, L Chindelevitch designed the model and modelling experiments. M Sharma, A Stephenson, T Gavenciak, J Salvatier performed and analysed the modelling experiments. J Kulveit, JM Brauner designed and conducted preference survey 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, G Leech, T Besiroglu, M Sharma, 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.

8. Keywords

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

Appendix A. Main results table

See next page. The effectiveness estimates are computed with the combined cases + deaths model. For the disutility scores, lower disutility implies higher utility and a stronger preference. Utilities are on an interval scale, the absolute values have no significance, only differences between utilities carry meaning. The zero point has no particular meaning. We can, e.g., say that the preference for *Some businesses closed* over *Stay-at-home order* was equally strong as the preference for *Gatherings limited to 100 people or less* over *Some businesses closed* (ca. 0.3 a.u., arbitrary units)

Table A.4: Main results of the paper. Mean ± standard deviation and 95% credible/confidence interval.

NPI	Percentage reduction in R	Disutility score	Perceived intervention cost
Mask-wearing	$-8.89 \pm 12.98(-36.07; 14.32)$	$0.12 \pm 0.03(0.05; 0.18)$	$0.63 \pm 0.19(0.27; 1.00)$
Symptomatic testing	$13.42 \pm 6.67(-0.58; 25.54)$	n.a.	n.a.
Gatherings limited to 1000 people or less	$15.81 \pm 8.48(-1.94; 31.08)$	$0.07 \pm 0.03(0.00; 0.14)$	$0.56 \pm 0.16(0.24; 0.88)$
Gatherings limited to 100 people or less	$16.79 \pm 9.66(-3.33; 34.79)$	$0.33 \pm 0.03(0.26; 0.40)$	$1.07 \pm 0.31(0.46; 1.68)$
Gatherings limited to 10 people or less	$27.97 \pm 9.53(7.55; 44.92)$	$0.52 \pm 0.03(0.45; 0.59)$	$1.68 \pm 0.49(0.71; 2.64)$
Some businesses closed	$26.25 \pm 8.69(8.26; 42.17)$	$0.61 \pm 0.03(0.54; 0.67)$	$2.01 \pm 0.60(0.84; 3.17)$
Most nonessential businesses closed	$33.58 \pm 8.40(15.95; 48.70)$	$0.90 \pm 0.04(0.83; 0.97)$	$4.08 \pm 1.27(1.59; 6.57)$
Schools closed	$49.69 \pm 5.20(38.92; 59.15)$	$0.44 \pm 0.03(0.38; 0.51)$	$1.41 \pm 0.41(0.60; 2.22)$
Stay-at-home order with exemptions	$14.06 \pm 7.96(-2.16; 28.71)$	$0.91 \pm 0.04(0.84; 0.98)$	$3.86 \pm 1.22(1.48; 6.24)$

Appendix B. The Epidemic Forecasting Global NPI database

Appendix B.1. Overview

Up-to-date information on the Epidemic Forecasting Global NPI (EFGNPI) database can be found at `http://epidemicforecasting.org/containment`.

The full database (DB) is a daily representation of the response of each of 97 countries. It aims at collecting as broad a range of NPIs as possible. However, data on minor NPIs is often hard to find. As a result, the absence of an entry does not necessarily mean that this NPI was not implemented by a country.

A smaller dataset, the EFGNPI Features dataset (FD), is derived from the full DB. The FD data aggregates many tags in the main database to produce a dataset easier to use in machine learning applications. The tags are also used to determine a stringency score for each feature. (Please note that details of how the FD data is produced from the main database may change slightly over time.)

Table B.5: Metadata for the two datasets

Dataset Name	Number of countries covered	Format	Number of indicators/ keywords	Indication of strength	Number of rows	Other information
Epidemic Forecasting Global NPI database (DB)	97*	1 row per country per day, comma separated keywords indicate measures introduced on that day. If no measures were introduced in a country on a given day, no row is recorded	194	For some tags; most are binary	1703	State, City/County, Target Country & State for travel restrictions, Plain text descriptions, Sources
Epidemic Forecasting Global NPI features dataset (FD)	67	1 row per country per day, active indicators in columns (integers or floats)	24	Yes	7370	NA

* The database contains data on 97 countries, but only 67 of these are complete at time of writing.

Appendix B.2. Collection

The underlying data was gathered by a team of volunteers. The database integrates many sources. Wikipedia entries were taken as a starting point for the set of NPIs implemented by each country. These were then refined by reference to national centres for disease control.

The full database is recorded as a dataset of tags. We began without a predefined list of attributes to record, so collection proceeded with a dynamic set of keyword tags as data on national responses was collected. After the data had been collected, a method for aggregating tags was created. The resulting database includes a 'Source' field for most rows.

Please note that the full EFGNPI database, in contrast to the data used in this study, has not been subject to extensive fact-checking.

Appendix B.3. Comparison to other datasets

Table B.6: Comparison to other datasets. *The database contains data on 97 countries, but only 67 of these are complete at time of writing.

Dataset Name	Number of countries covered	Data Type	Number of indicators/ keywords	Indication of strength
Epidemic forecasting global NPI database (DB)	97*	keywords, comma separated	194	For some tags, most are binary
Epidemic forecasting global NPI features dataset (FD)	67	Indicators	24	Yes
Oxford COVID-19 Government Response Tracker (CGRT)[7]	190	Indicators	17	Yes
ACAPS COVID-19: Government Measures Dataset[8]	191	Categories/ subcategories	6 categories, 35 subcategories	For some categories

It is important that researchers select the dataset appropriate for their use-case. We think that a particular strength of the EFGNPI database is that it tracks a vast array of NPIs, but possibly at the cost of completeness. For the features that are contained in it, it seems

likely that the Oxford COVID-19 Government Response Tracker dataset will have the highest quality, given the large team behind this dataset. However, as we have stated in various sections of this paper: Given our experience with several public datasets and our own data collection, we encourage fellow Covid-19 researchers to independently verify the quality of public data they use, if feasible.

Appendix C. Mask prevalence survey

Volunteers and Amazon Mechanical Turk (AMT) workers were asked to fill out an online survey between 25th March and 7th April 2020. The first-round volunteers were recruited via Facebook posts and private emails, with a request to both complete the survey and share it with their contacts, especially overseas contacts. Owing to a lack of geographical coverage in the first round, a second round, surveying users of country-specific forums on Reddit, was conducted and completed on 28th April.

The survey features three sets of questions, regarding:

1. the requirements or recommendations to wear masks in the participant's home country. (This question was added in the second round.)
2. the percentage of mask-wearers they saw in public at weekly intervals between the end of February and the beginning of April
3. the number of people in indoor public areas as a percentage of the usual number of people seen in these areas at weekly intervals between the end of February and the beginning of April

Both strategies (private word of mouth and public internet sampling) are likely to yield non-representative samples owing to self-selection. This could yield poor results if mask usage varies a lot within countries, for instance in large countries such as India and the United States. However, we found a good deal of consistency in responses within countries on specified days. The average standard deviation of "percentage of population wearing masks" within country-days was 18.6, while the same measure, between countries but within days, was 28.6. Given this, we expect the inclusion of countries with even a single response to give a better indication of mask-wearing behaviour in that country than assuming such countries to have average levels.

Appendix C.1. Data transformation and combination with government orders

We computed a binary feature of mask-wearing, attributed to the middle day of each week in the survey, by thresholding the average survey response for that week at 60%.

To create the mask-wearing feature used in our modelling, we combined the data from the surveys with data on government orders requiring the the wearing of masks in public places in the following way:

Table C.7: Total survey responses by country after data cleaning

Country	Responses
Albania	15
Andorra	2
Austria	22
Belgium	5
Bosnia and Herzegovina	11
Bulgaria	1
Croatia	6
Czechia	49
Denmark	9
Estonia	102
Finland	25
France	9
Georgia	1
Germany	10
Greece	1
Hungary	18
Iceland	132
Ireland	16
Israel	3
Italy	1
Latvia	10

Country	Responses
Lithuania	1
Malaysia	35
Malta	14
Mexico	19
Morocco	0
Netherlands (the)	192
New Zealand	0
Norway	7
Poland	11
Portugal	17
Romania	20
Serbia	0
Singapore	4
Slovakia	64
Slovenia	5
South Africa	16
Spain	17
Sweden	25
Switzerland	2
United Kingdom of Great Britain and Northern Ireland (the)	11

- We only considered survey results for countries with at least 5 responses
- If there was either a government order or a mask-wearing start date according to the survey results (but not both), we accepted that date
- If there was both a government order and a mask-wearing start date according to survey results, we accepted whichever was earlier. An exception were cases where the start date according to surveys was less than 3 days before the government order. In these cases we accepted the date of the government order (because the temporal resolution of the survey results was $+/-$ 3.5 days)

Mask data in detail (sheet “combined”): LINK

Appendix C.2. Data calibration

If we assume that, for country days with over 15 responses, the true number of people wearing masks is given by the mean of the survey responses, we can estimate the misclassification rate for different numbers of responses by randomly sampling responses for that country day and comparing them with the sample mean excluding the selected responses. Table C.8 represents the average from 100 iterations of this procedure.

Table C.8: Results of bootstrap simulation ($n = 100$) of misclassification rates

	1 response per country	4 responses per country
False positives	4.2	1.8
True positives	22.6	23.4
False negatives	6.0	5.4
True negatives	87.25	89.5

Appendix D. Sensitivity results

We replicate the posterior of the effectiveness of NPIs, showing its sensitivity to variations of the assumptions and the data.

Recall that we show *cumulative* effects for two sets of NPIs: gatherings and business closures. This means that, e.g., a high sensitivity for closing some businesses will show up a second time as a high sensitivity for closing most businesses. This overstates the number of individual parameters α_i which are sensitive. To illustrate this duplication, we have also plotted the first sensitivity with cumulative effects (Figure D.10) and without (Figure D.11). All other figures are cumulative.

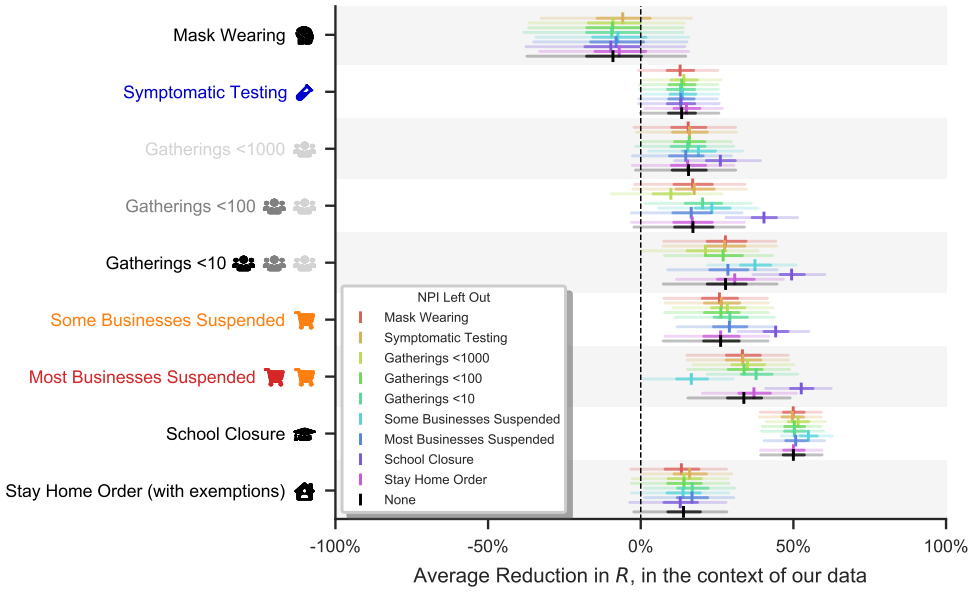

Figure D.10: Robustness to left-out / unobserved NPIs. Replications of the posterior in Figure EF for the combined model while hiding each of the NPIs once. Note that we display *cumulative* effects for gathering bans and business closures, so that any sensitivity of these NPIs is also cumulative, showing up multiple times on the graph. The figures thus overstate the number of parameters α_i which are sensitive. Figure D.11 shows sensitivity without this accumulation.

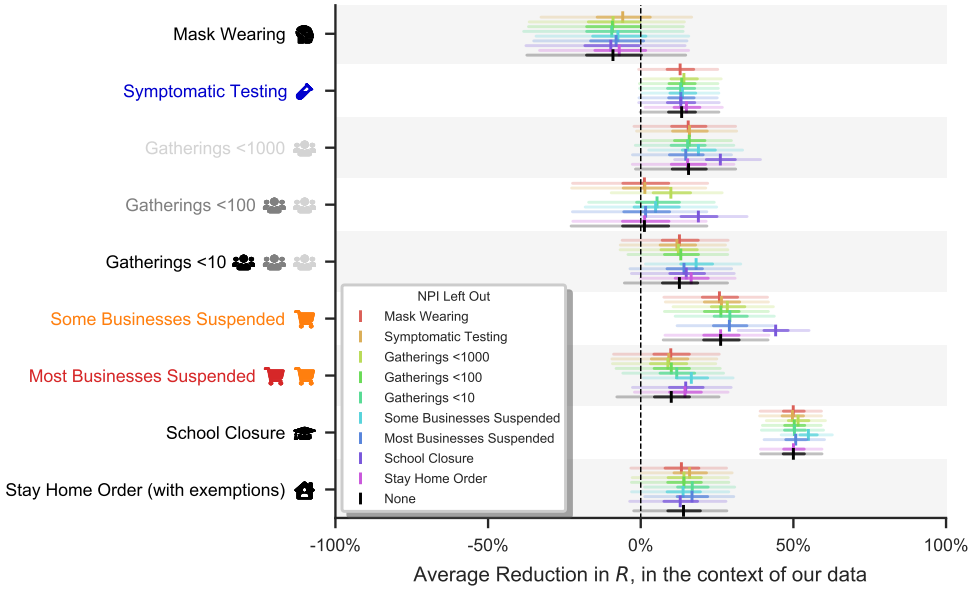

Figure D.11: Robustness to left-out / unobserved NPIs - with marginal / non-cumulative effects.

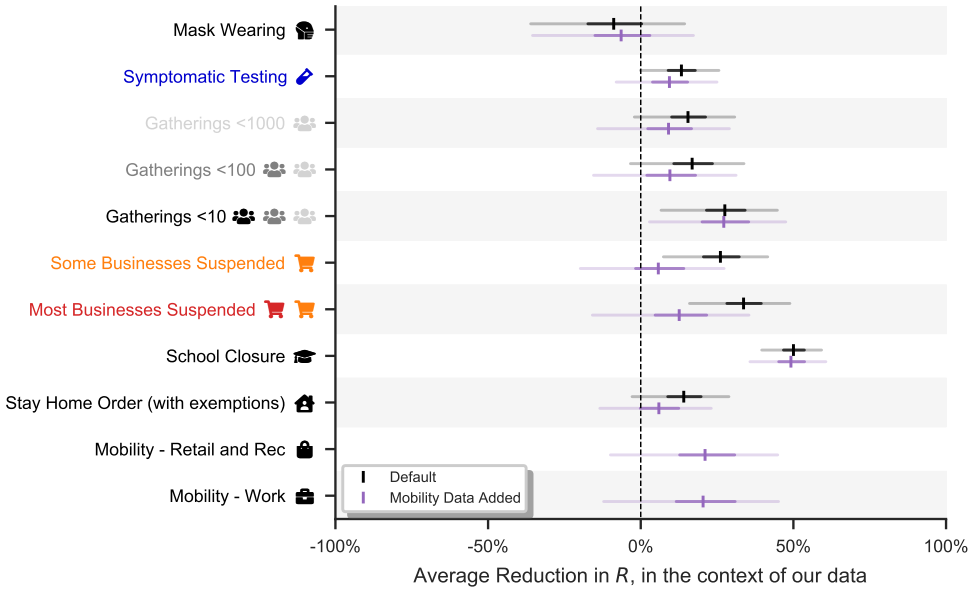

Figure D.12: Sensitivity to including mobility data as additional 'NPI'. Mobility data serves as a proxy for unobserved behavior changes. Mobility data explains most of the effect of business closures and stay-home-orders, which is expected as the effect of these NPIs is mediated through retail, recreation, and workplace mobility. Results were nearly identical when excluding workplace mobility (not shown). We did not experiment with other mobility categories such as groceries and pharmacy.

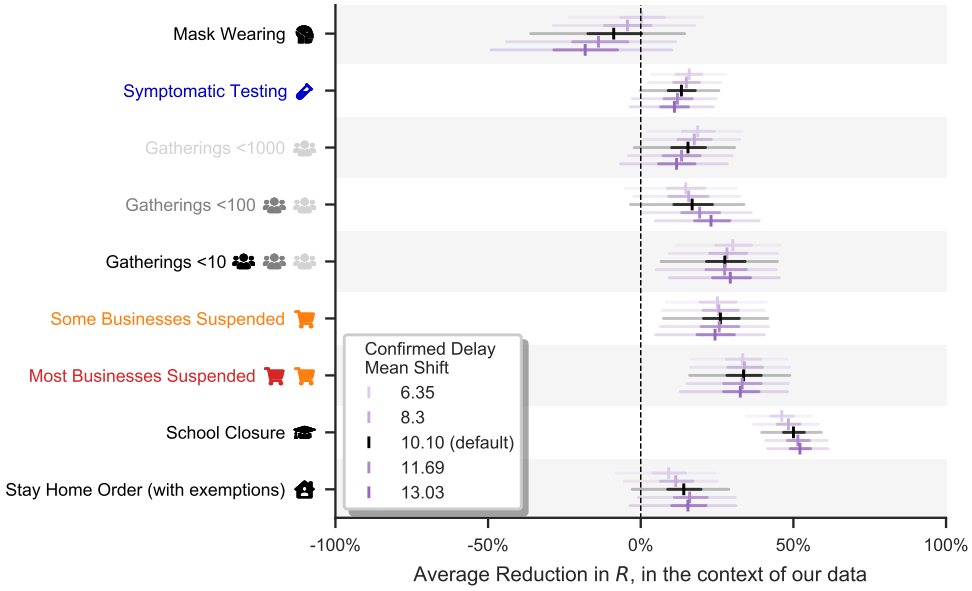

Figure D.13: Sensitivity to mean delay from infection to confirmation (combined model). The default mean is 10.1 days (including the incubation period); it is shifted over a window of 8 days in this figure.

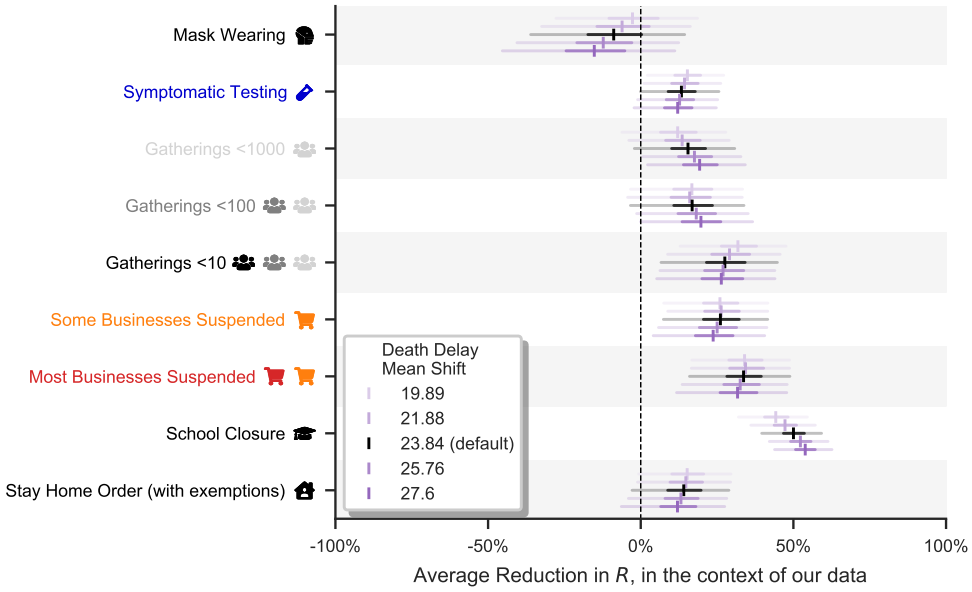

Figure D.14: Sensitivity to mean delay from infection to death (combined model). The default mean is 23.84 days (including the incubation period); it is shifted over a window of 8 days in this figure.

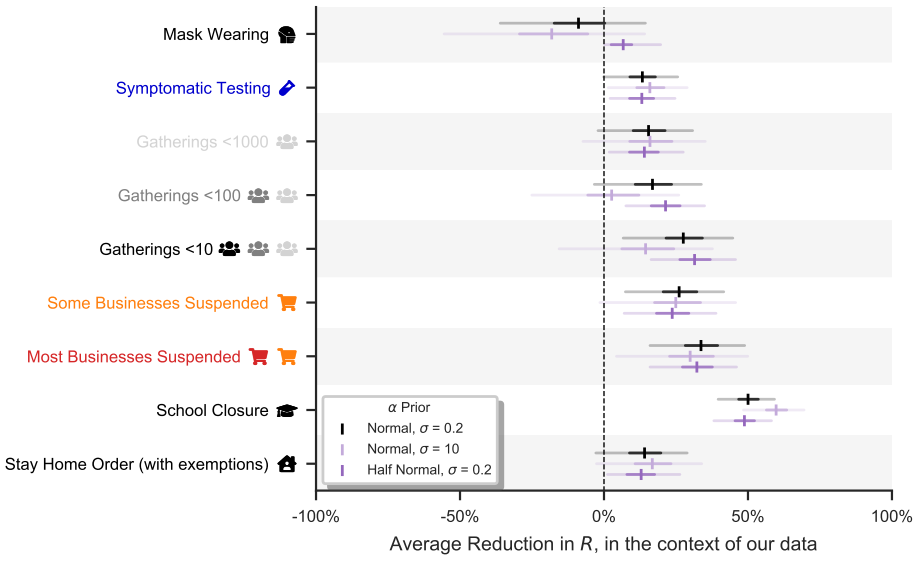

Figure D.15: Sensitivity to prior on the effectiveness parameters α_i (combined model). The default prior has α_i normally distributed with mean 0 standard deviation $\sigma = 0.2$ (Appendix F). The alternative priors we tested are 1) a very wide prior, with $\sigma = 10$ and a 2) Half-Normal prior that only allows for positive effectiveness.

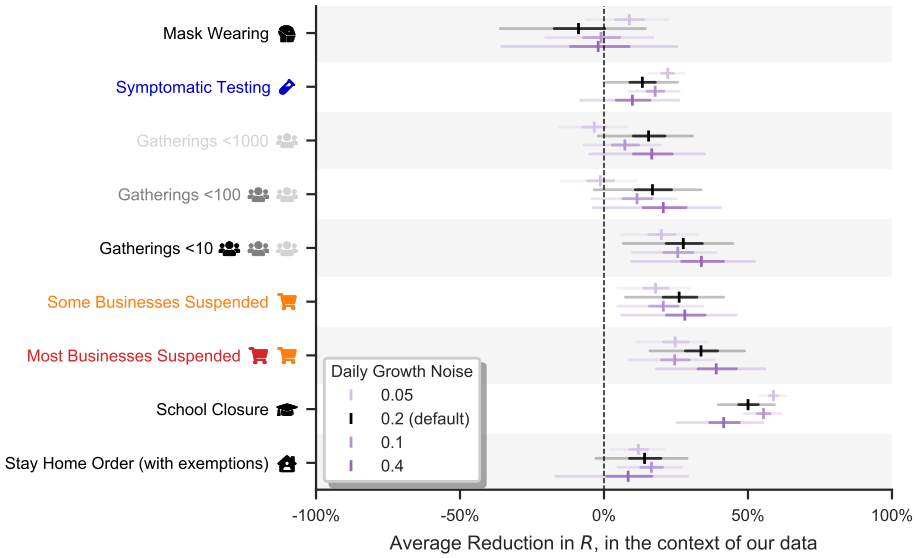

Figure D.16: Sensitivity to the standard deviation of multiplicative noise $\epsilon_{t,c}^{(\cdot)} \sim \text{Lognormal}(0, \sigma_N)$ on new infections (combined model). We vary σ_N. Deaths and cases have independent noise terms, with the same standard deviation σ_N. Note that a larger noise scale implies that the rates of ascertainment (testing) and fatality are allowed to change more rapidly. Predictably, results are less confident given more noise. Our default value was chosen by cross-validation (with the validation log-likelihood).

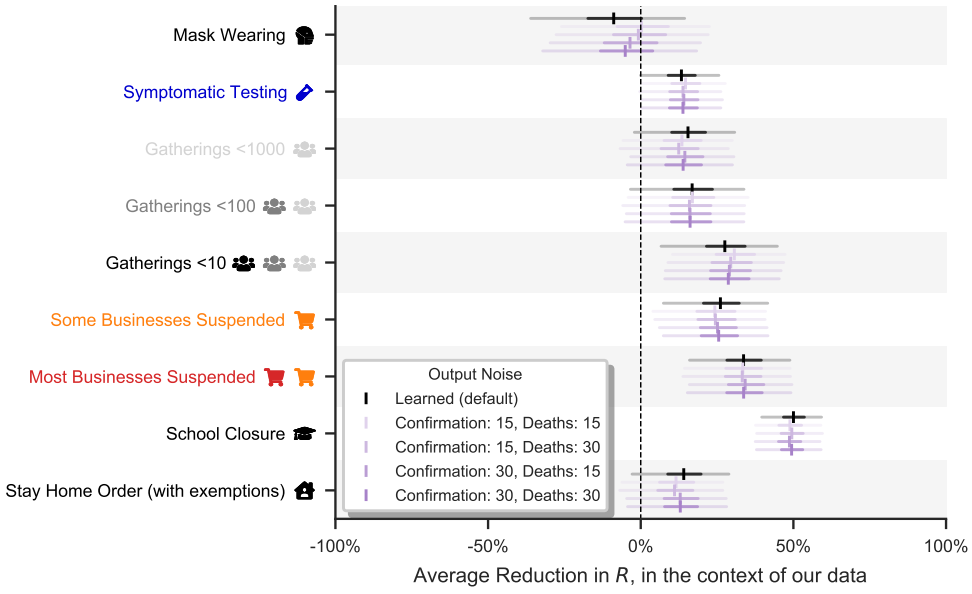

Figure D.17: Sensitivity to the dispersion of the output noise on deaths and confirmed cases (combined model). We vary the parameter ψ, given in Appendix Appendix F. In our main model, we learned this parameter.

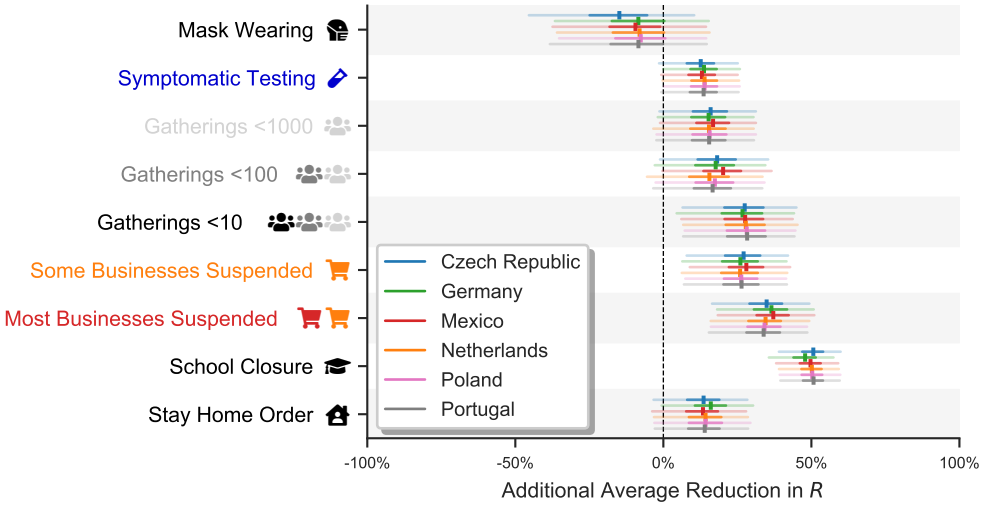

Figure D.18: Sensitivity to 6 randomly selected left out countries with >100 deaths. Note the Czech Republic is one of the countries implementing mask-wearing before April, explaining the higher sensitivity.

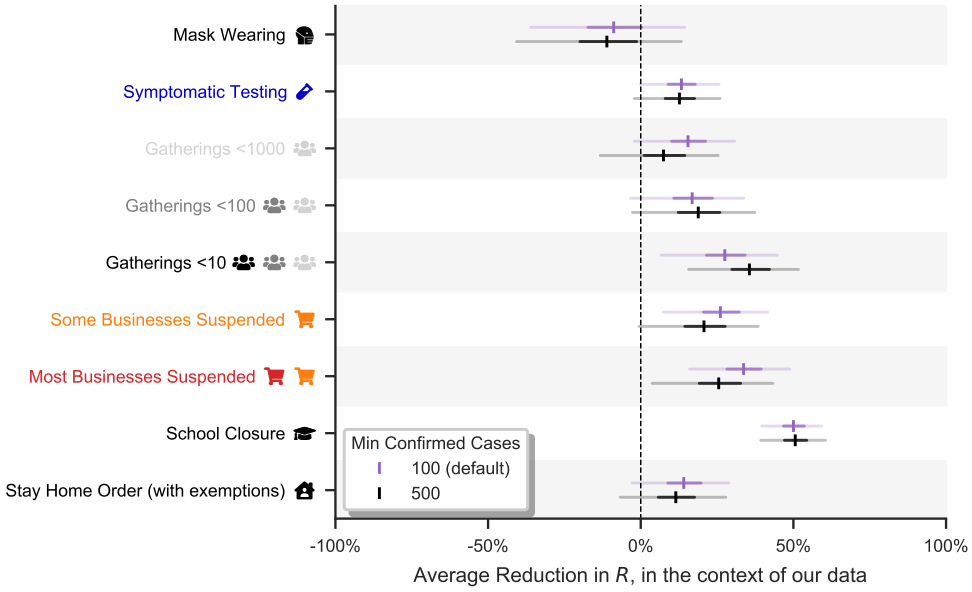

Figure D.19: Sensitivity to excluding days with few cumulative cases. By default, we mask days in each country before there were <100 cumulative cases, because imported cases could bias the numbers. Changing to <500 cumulative cases removes a substantial fraction of our data.

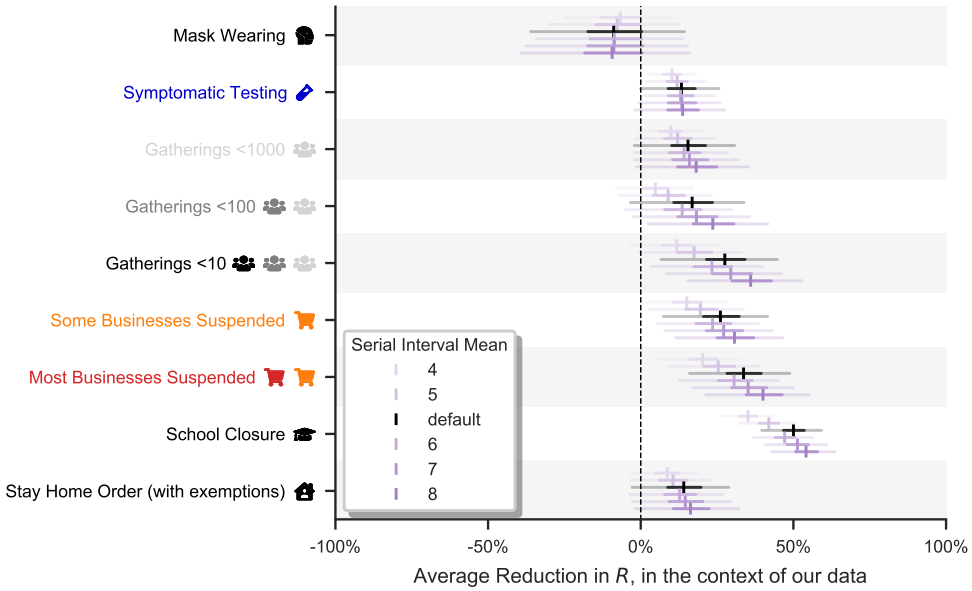

Figure D.20: Sensitivity to shifting the serial interval distribution. A shorter serial interval implies a lower value of *R0*, so it is expected that the reductions in *R* will be smaller (since *R0* will be small to begin with). Indeed, smaller reductions are sufficient given a smaller *R0*.

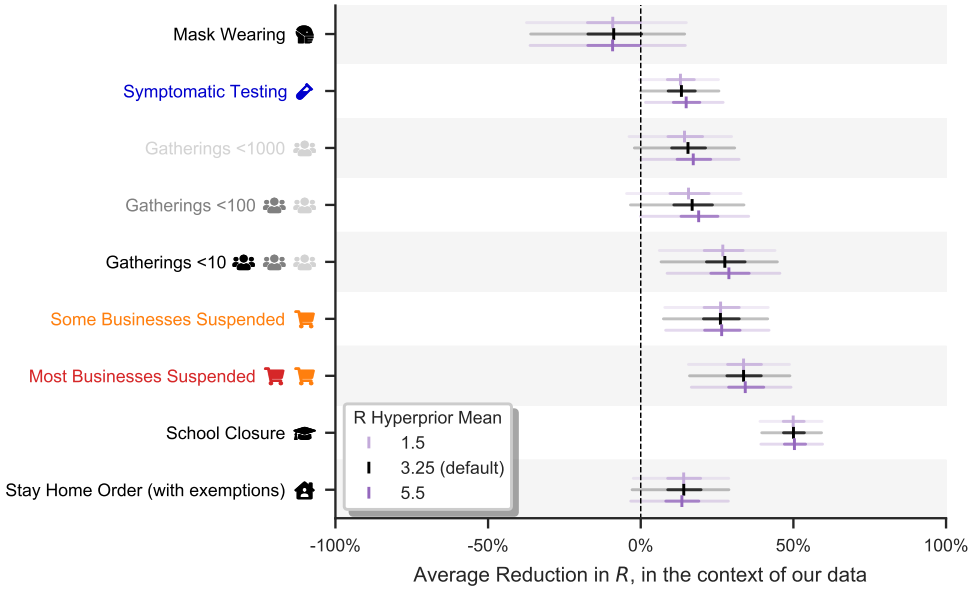

Figure D.21: Sensitivity to the mean of the hyperprior on $R_{0,c}$

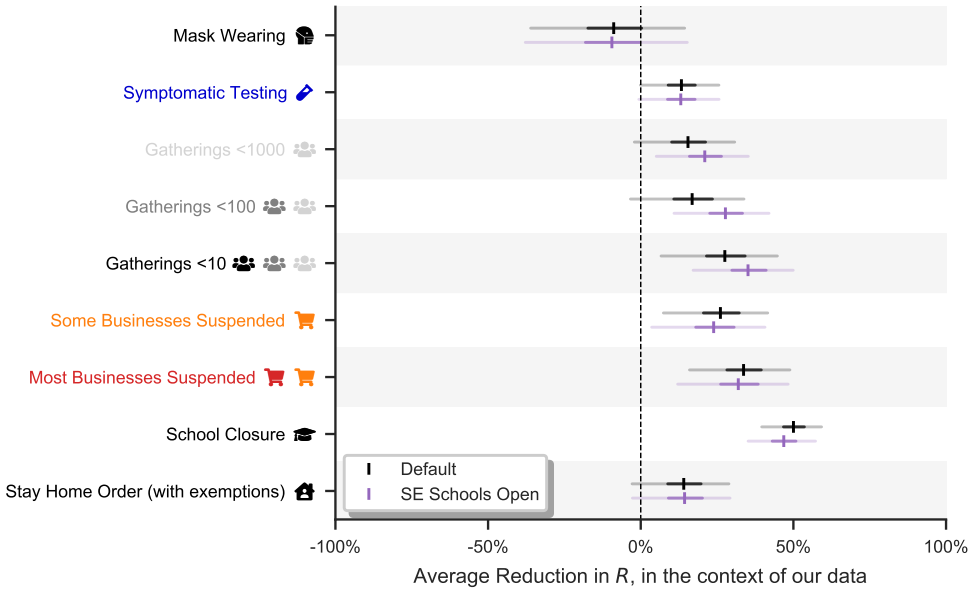

Figure D.22: Sensitivity to counting schools as open/closed in Sweden. Sweden closed high schools and universities on the 18th of March, but not elementary schools. We and Flaxman et al. counted this as "schools closed", but Banholzer et al. counted this as "schools open". This was the largest difference between our data on schools and Banholzer et al.

Appendix E. Additional Results

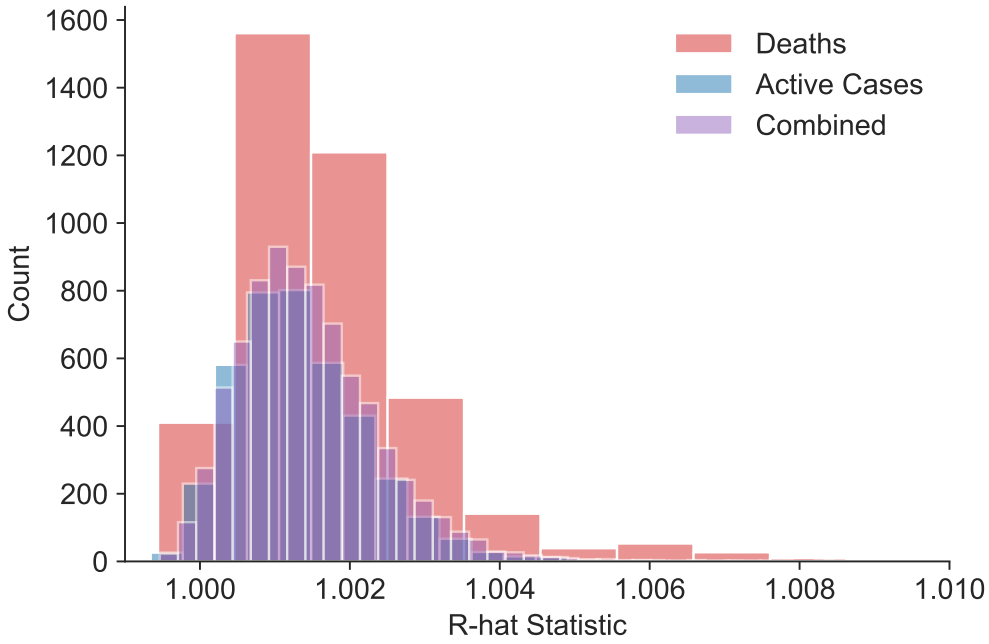

Figure E.23: MCMC stability results. Values are close to 1, indicating convergence.

Appendix E.1. Posterior Correlation

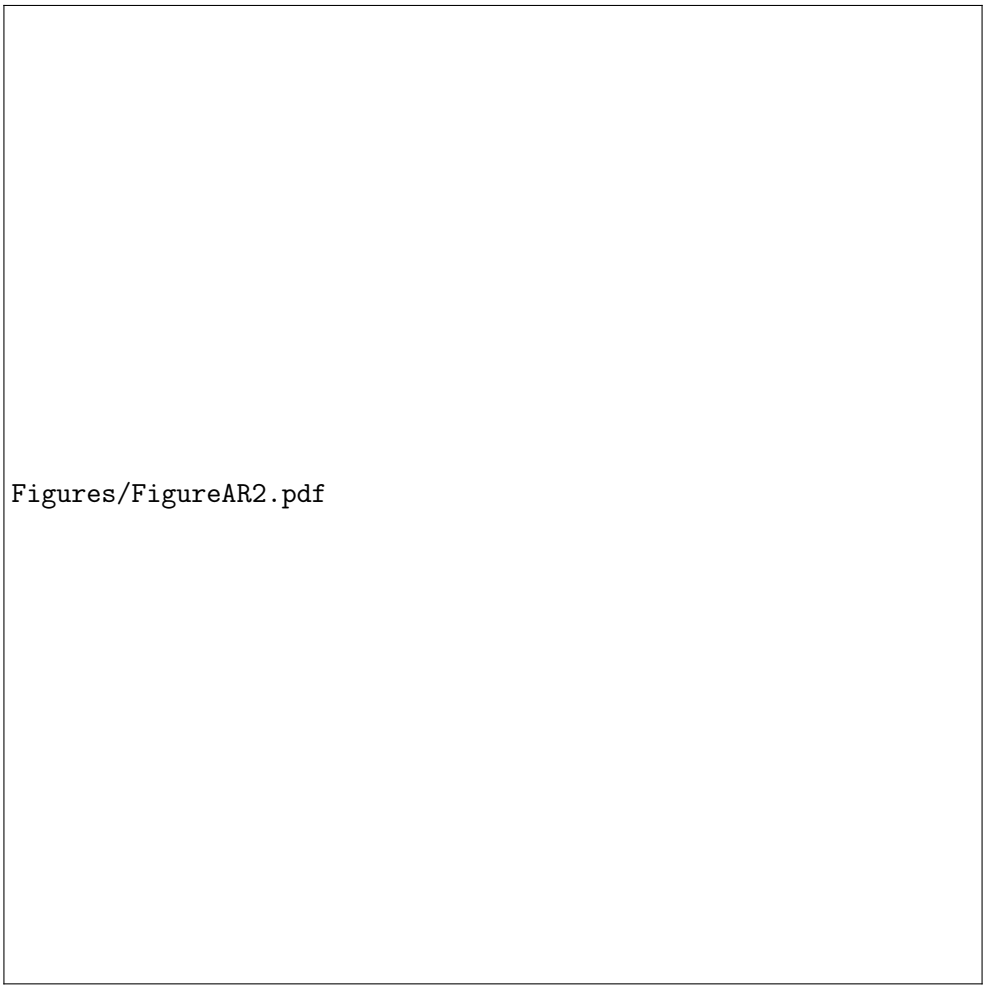

Figure E.24: Combined model posterior correlations. The parameters α_i are typically negatively correlated for NPIs which are often used together, such as stay-home-orders and suspending most businesses, reflecting uncertainty about which NPI is reducing *R*. The effectiveness of the *combination* of two negatively correlated NPIs may have narrower uncertainty estimates than the individual effects we plotted in the main text and Appendix Appendix D.

Appendix E.2. Additional Country Holdouts

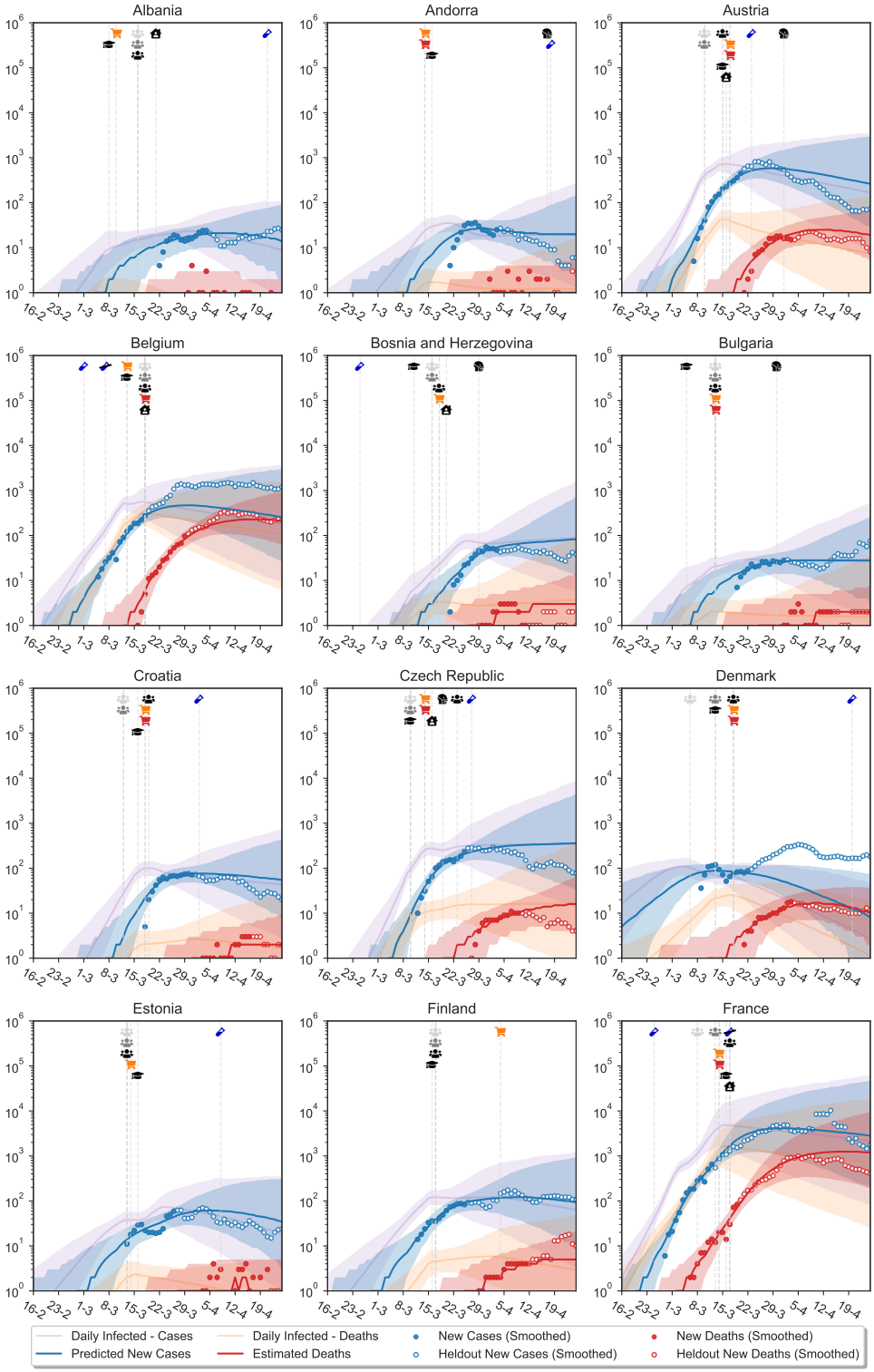

Figure E.25

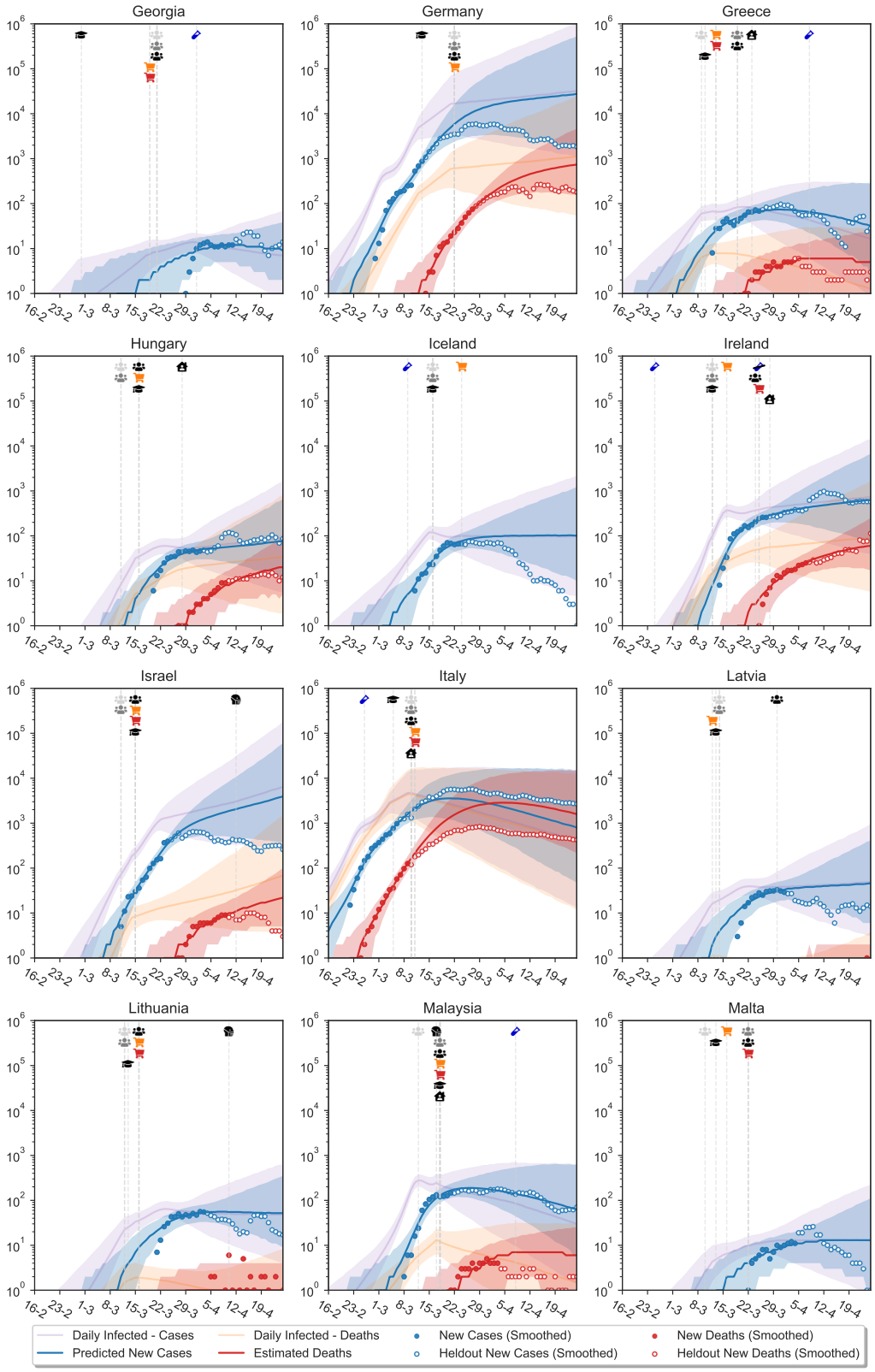
Georgia
Germany
Greece
Hungary
Iceland
Ireland
Israel
Italy
Latvia
Lithuania
Malaysia
Malta
16-2 23-2 1-3 8-3 15-3 22-3 29-3 5-4 12-4 19-4
Daily Infected - Cases
Predicted New Cases
Daily Infected - Deaths
Estimated Deaths
New Cases (Smoothed)
Heldout New Cases (Smoothed)
New Deaths (Smoothed)
Heldout New Deaths (Smoothed)

Figure E.26

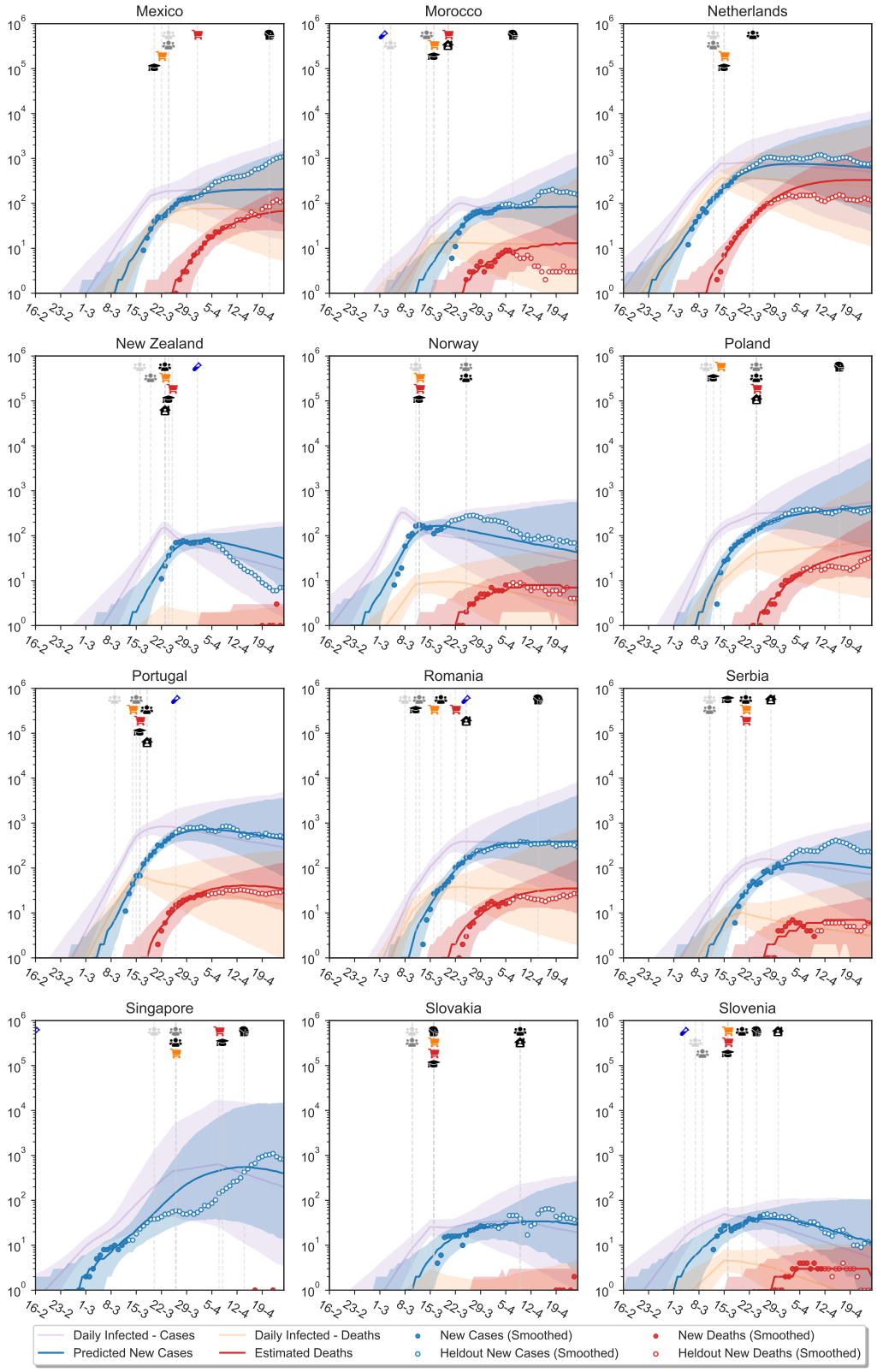
Mexico
Morocco
Netherlands
New Zealand
Norway
Poland
Portugal
Romania
Serbia
Singapore
Slovakia
Slovenia
16-2 23-2 1-3 8-3 15-3 22-3 29-3 5-4 12-4 19-4
Daily Infected - Cases
Predicted New Cases
Daily Infected - Deaths
Estimated Deaths
New Cases (Smoothed)
Heldout New Cases (Smoothed)
New Deaths (Smoothed)
Heldout New Deaths (Smoothed)

Figure E.27

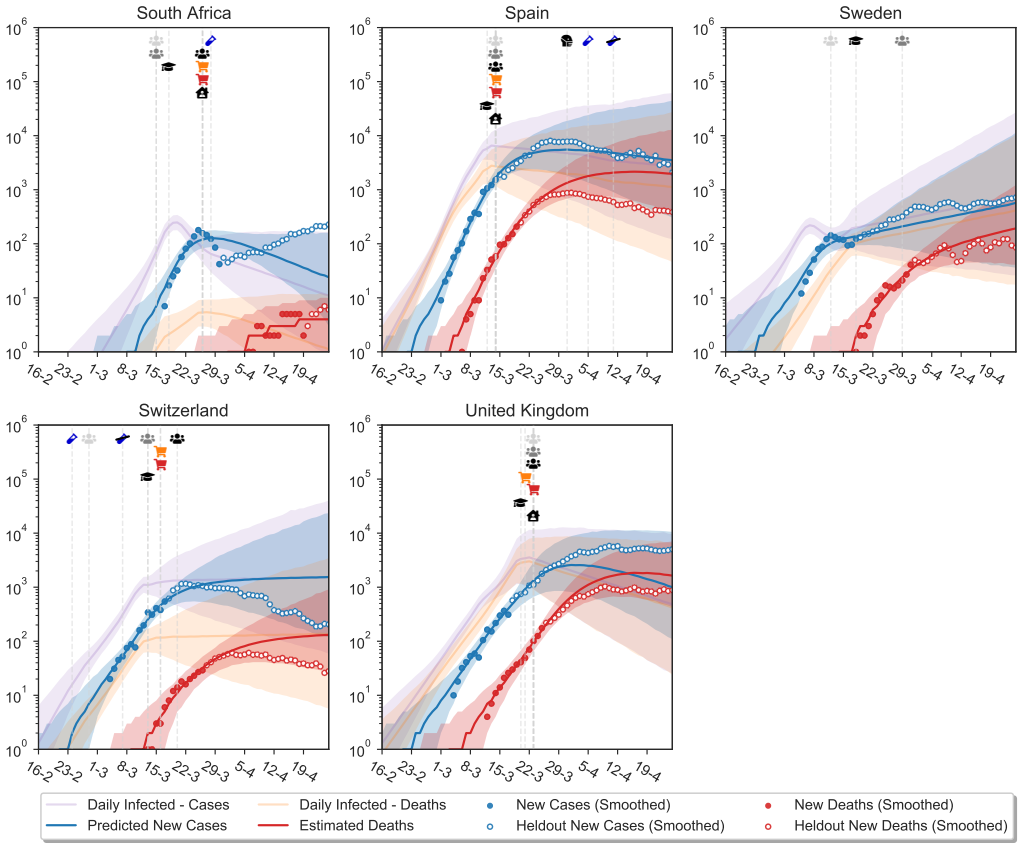

Figure E.28

Figure E.29: Holdout predictions of deaths and cases for all 41 countries (combined model). Empty dots are not shown to the model. 14 initial days are shown to the model, to enable inferring the basic R_0. The results show that our model makes sensible and well-calibrated forecasts over long time periods. There are no predicted deaths in some regions because there were no recorded deaths yet in the first 14 days with data.

Appendix E.3. Additional model fits

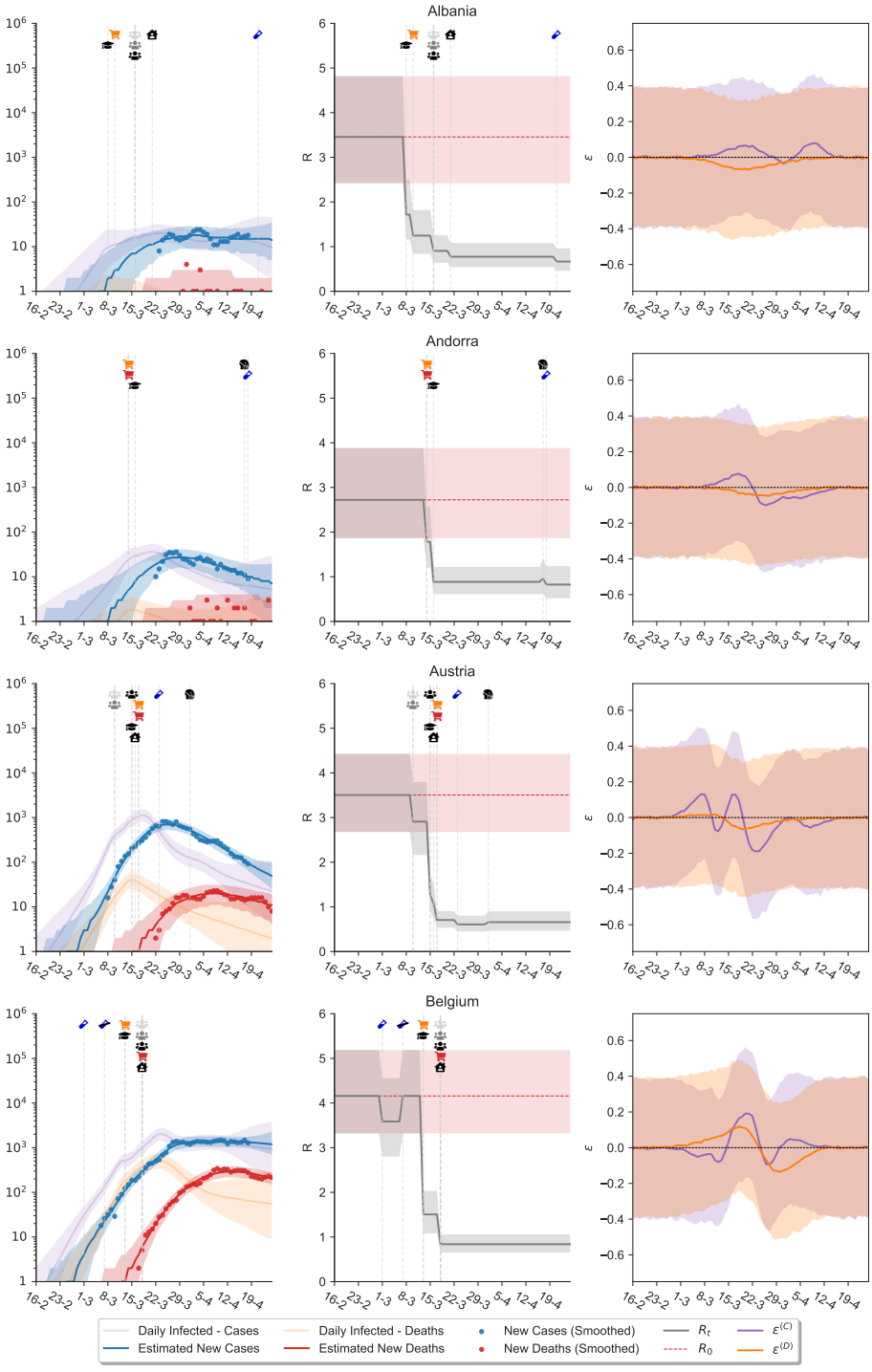

Figure E.30

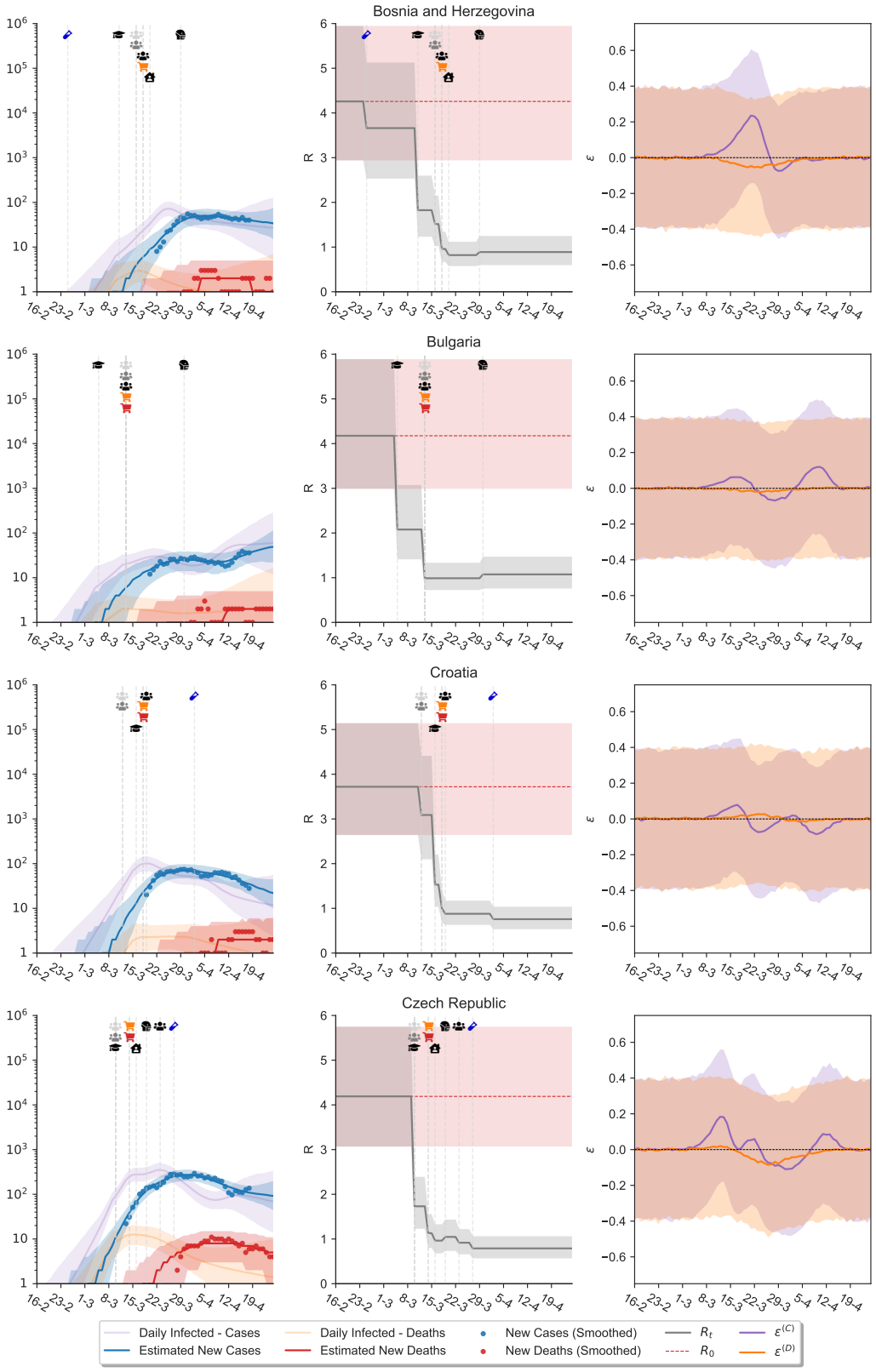
Bosnia and Herzegovina
Bulgaria
Croatia
Czech Republic
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.31

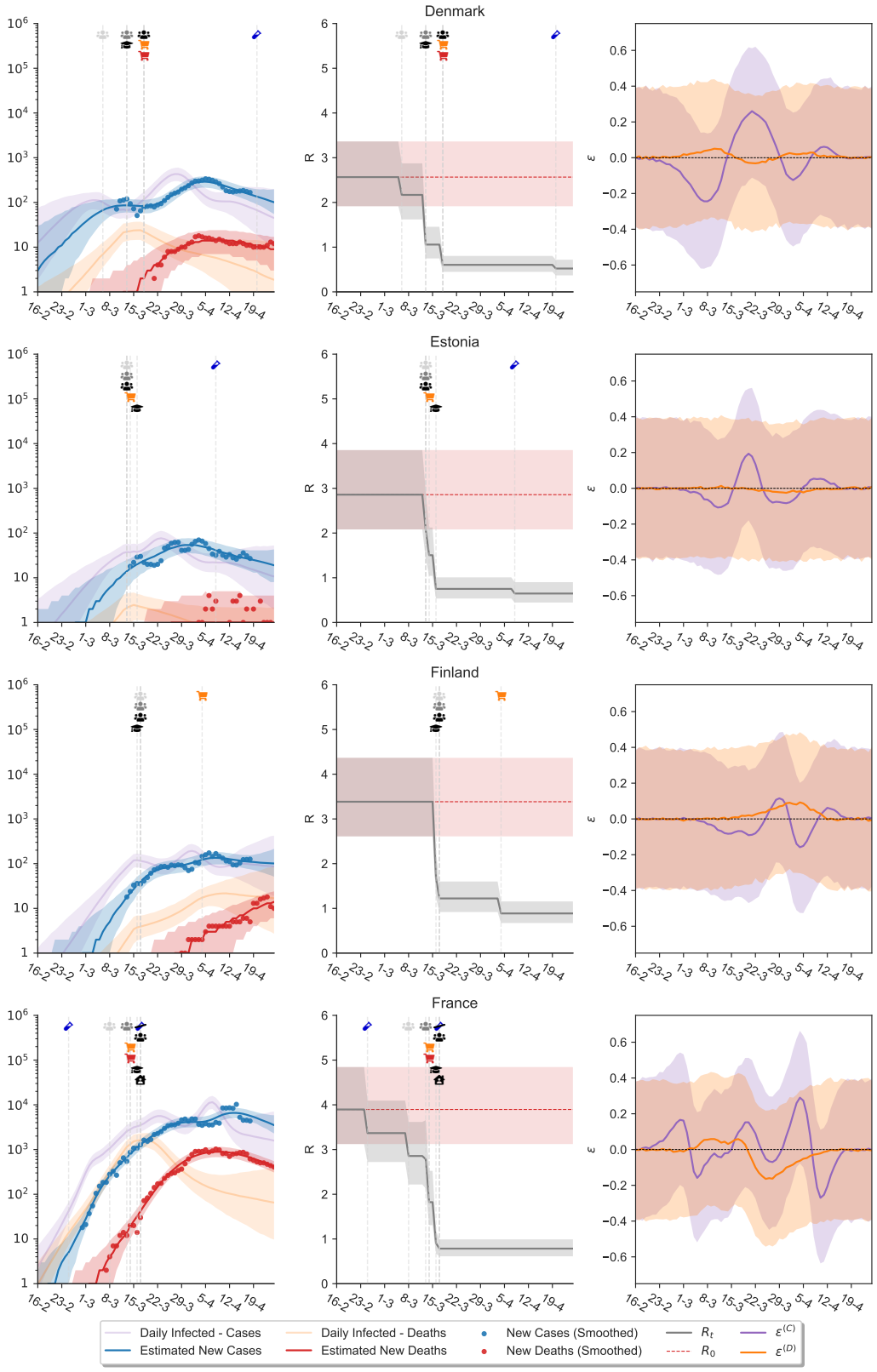
Denmark
Estonia
Finland
France
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.32

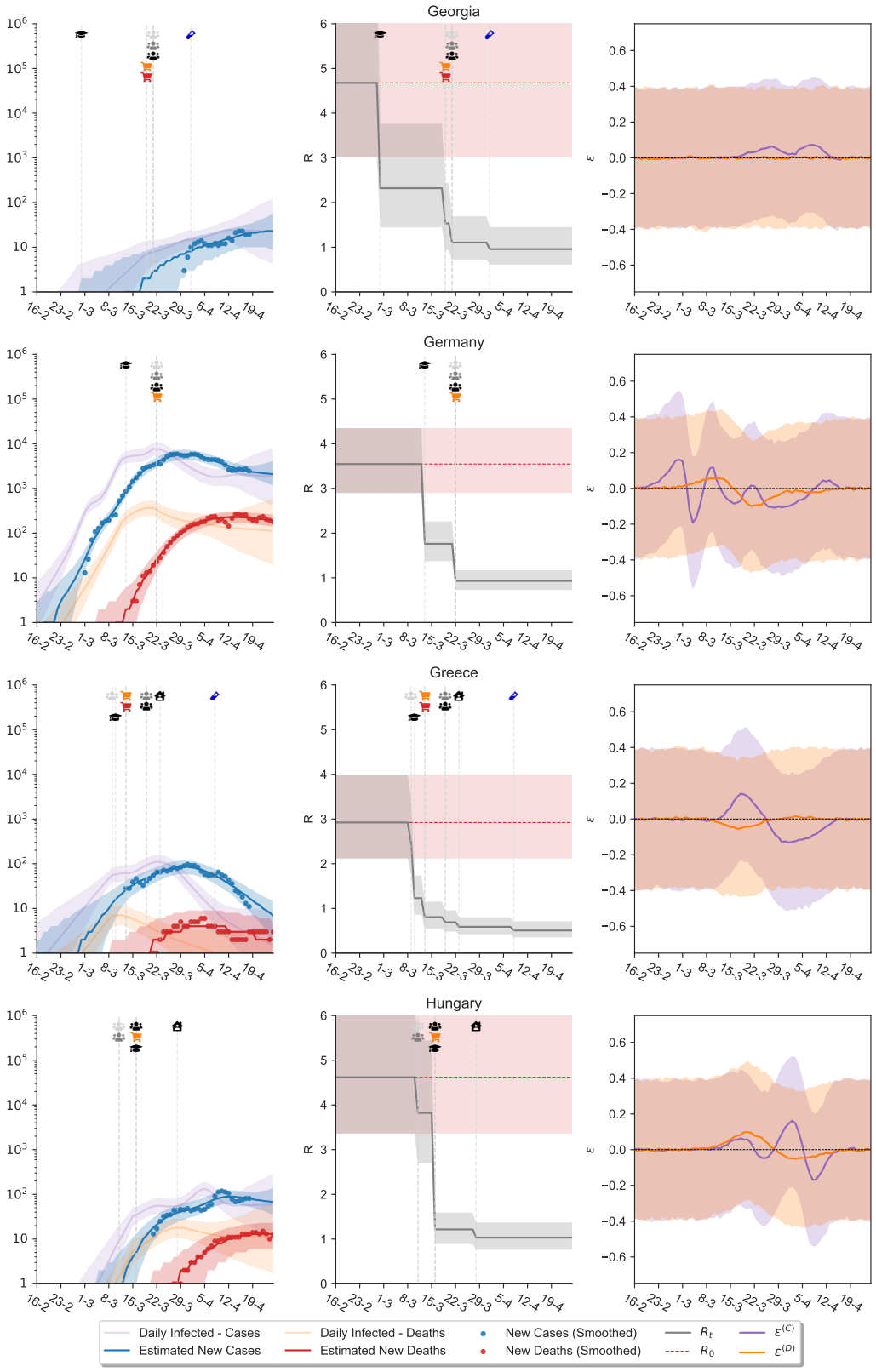
Georgia
Germany
Greece
Hungary
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.33

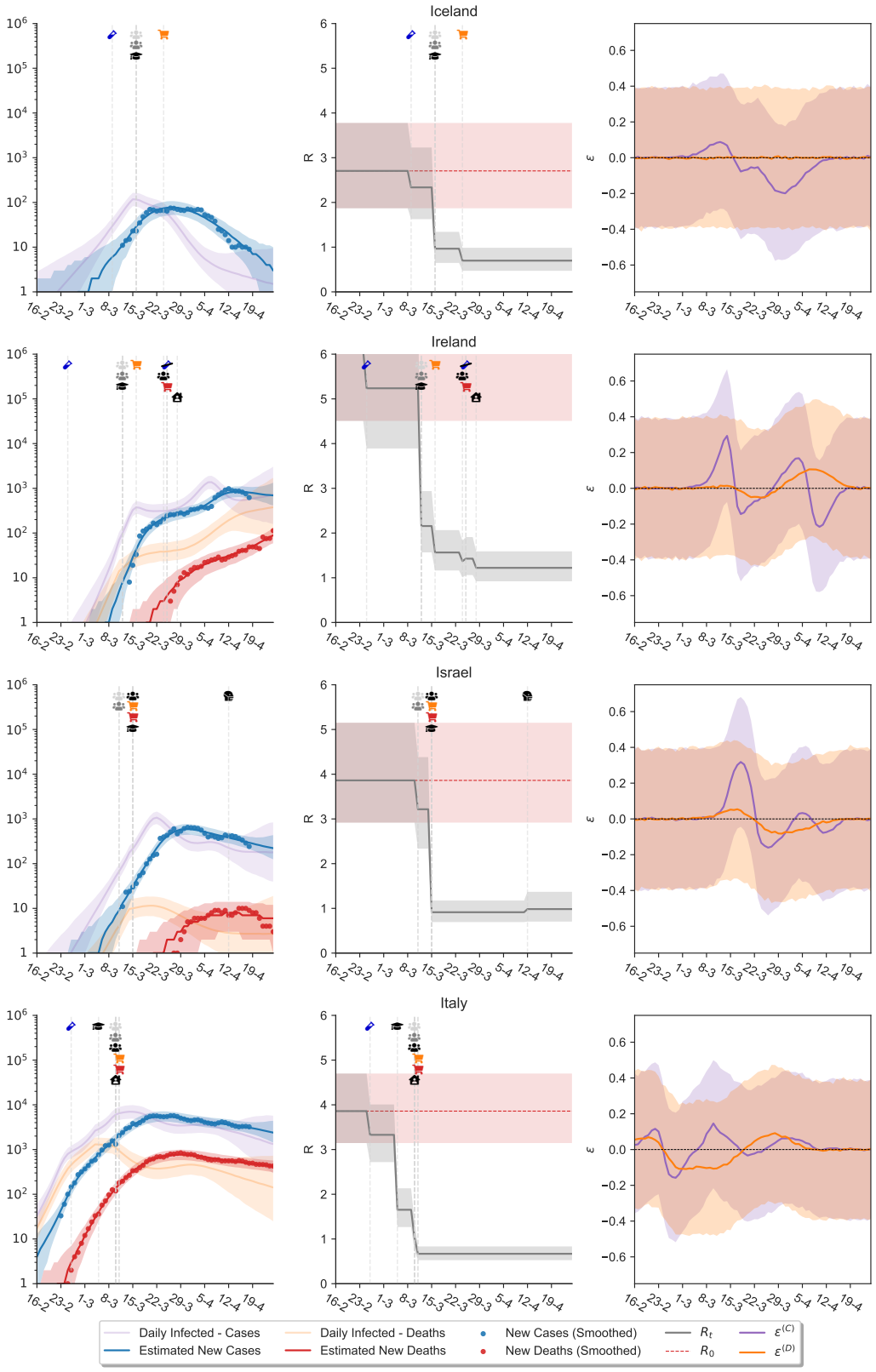
Iceland
Ireland
Israel
Italy
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.34

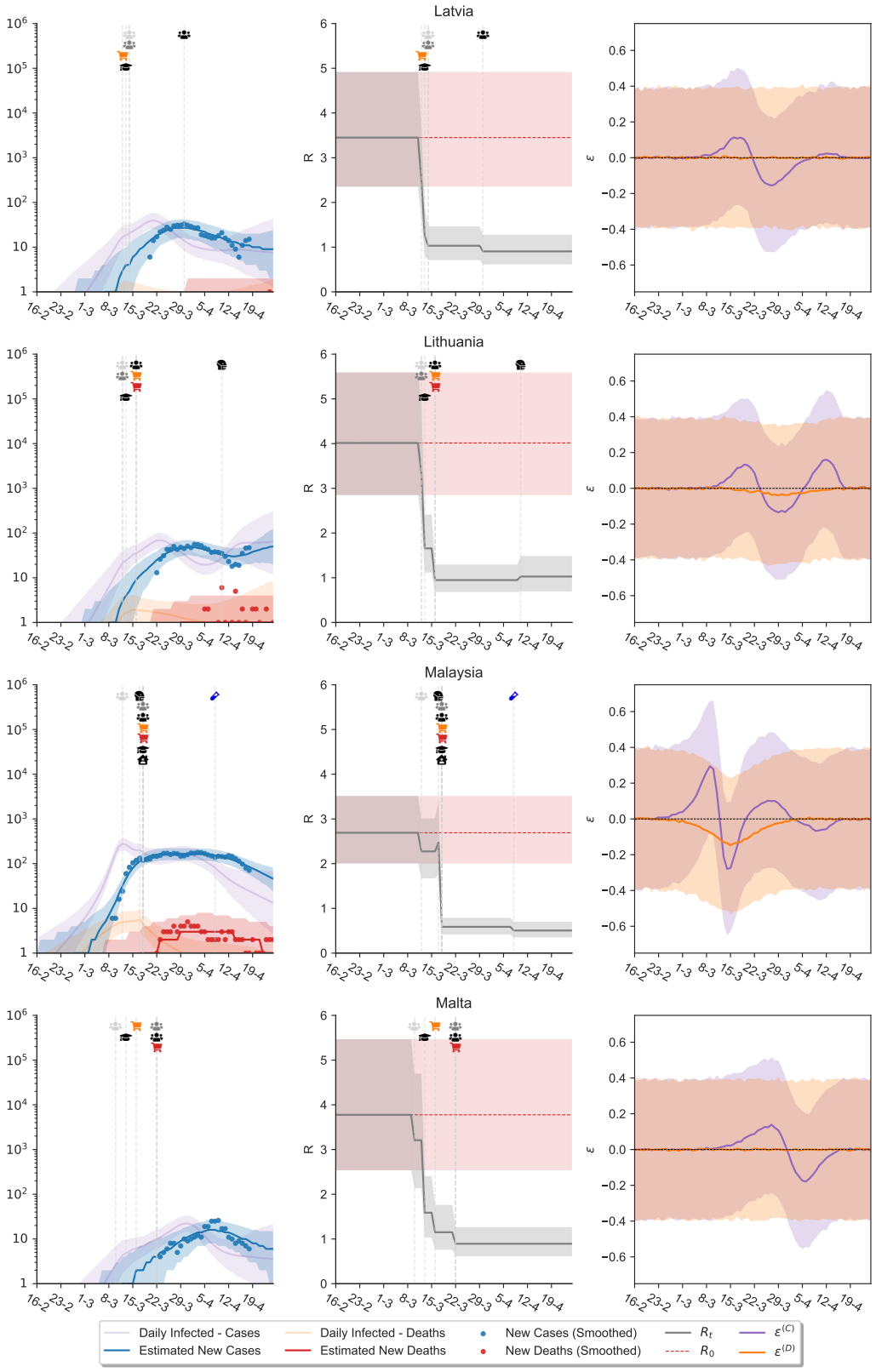
Latvia
Lithuania
Malaysia
Malta
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.35

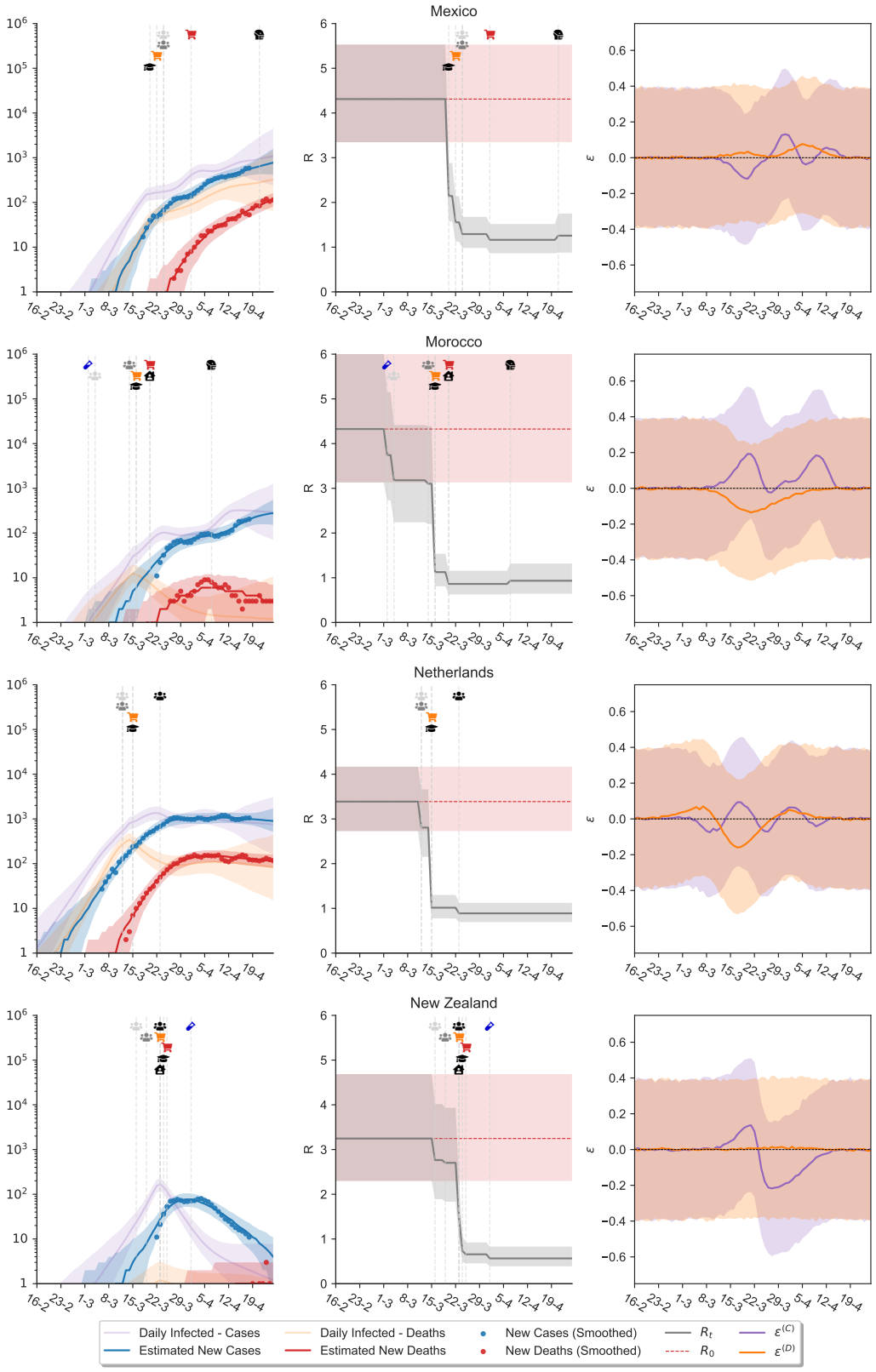
Mexico
Morocco
Netherlands
New Zealand
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.36

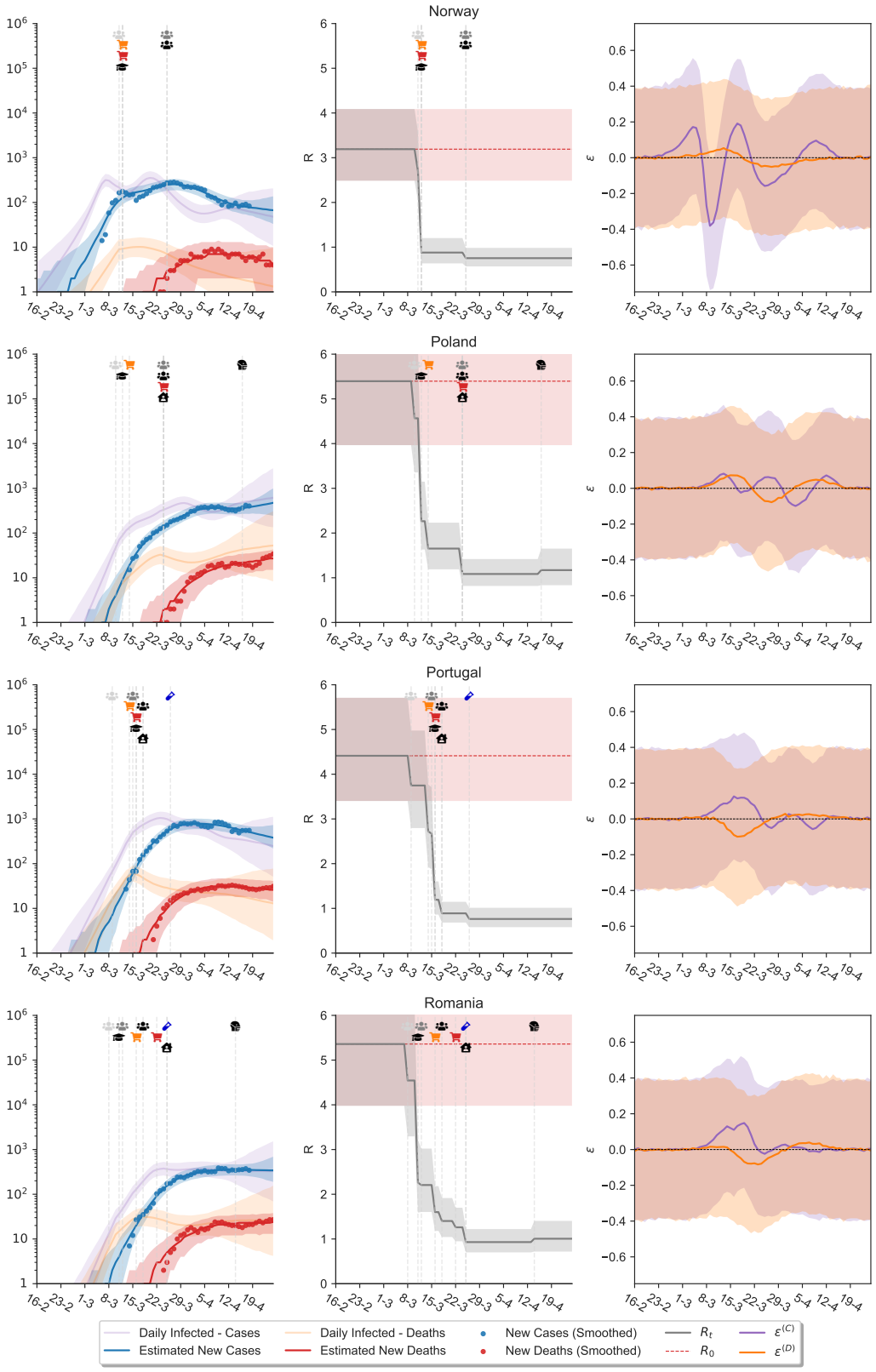
Norway
Poland
Portugal
Romania
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.37

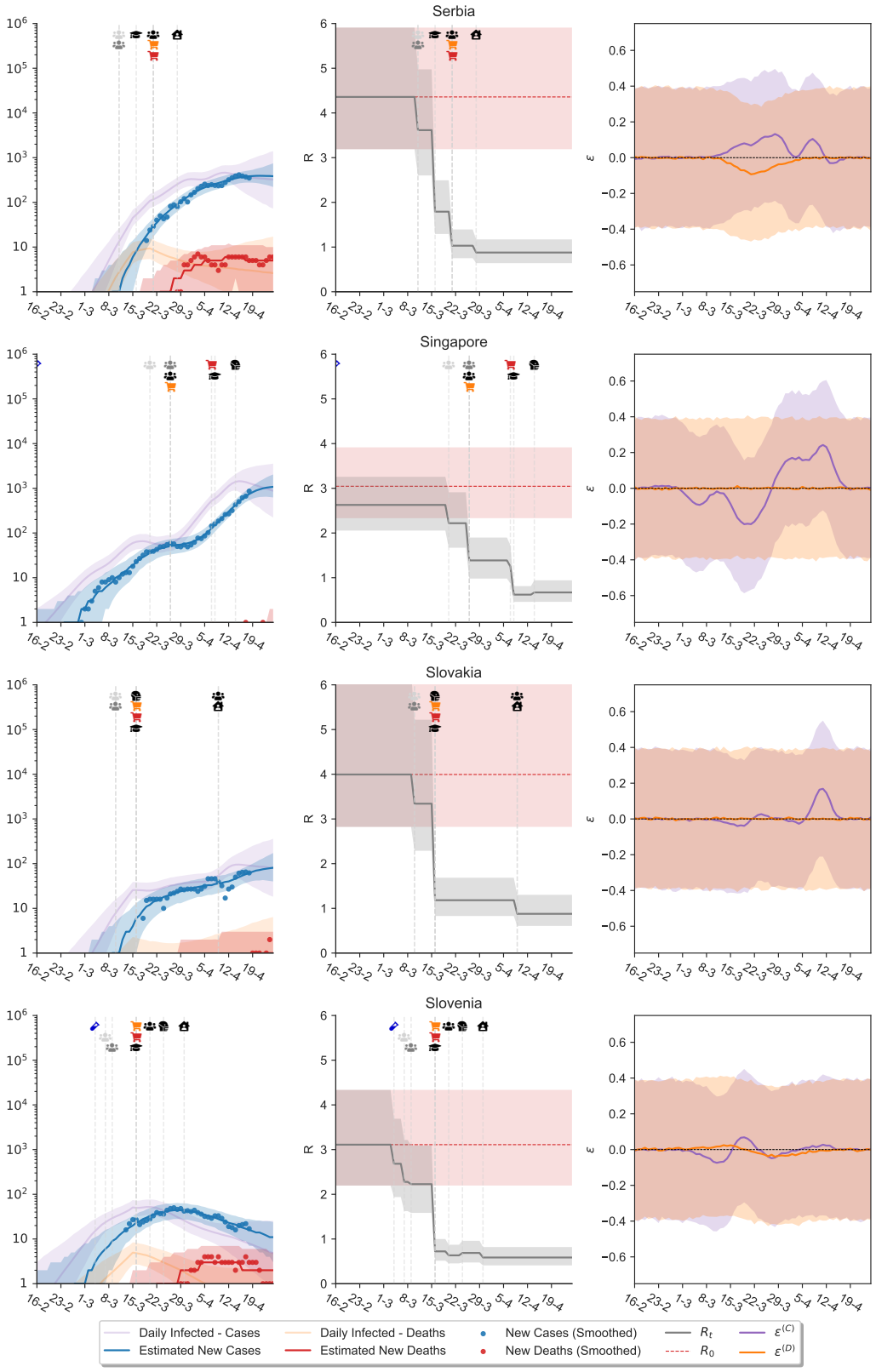
Serbia
Singapore
Slovakia
Slovenia
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.38

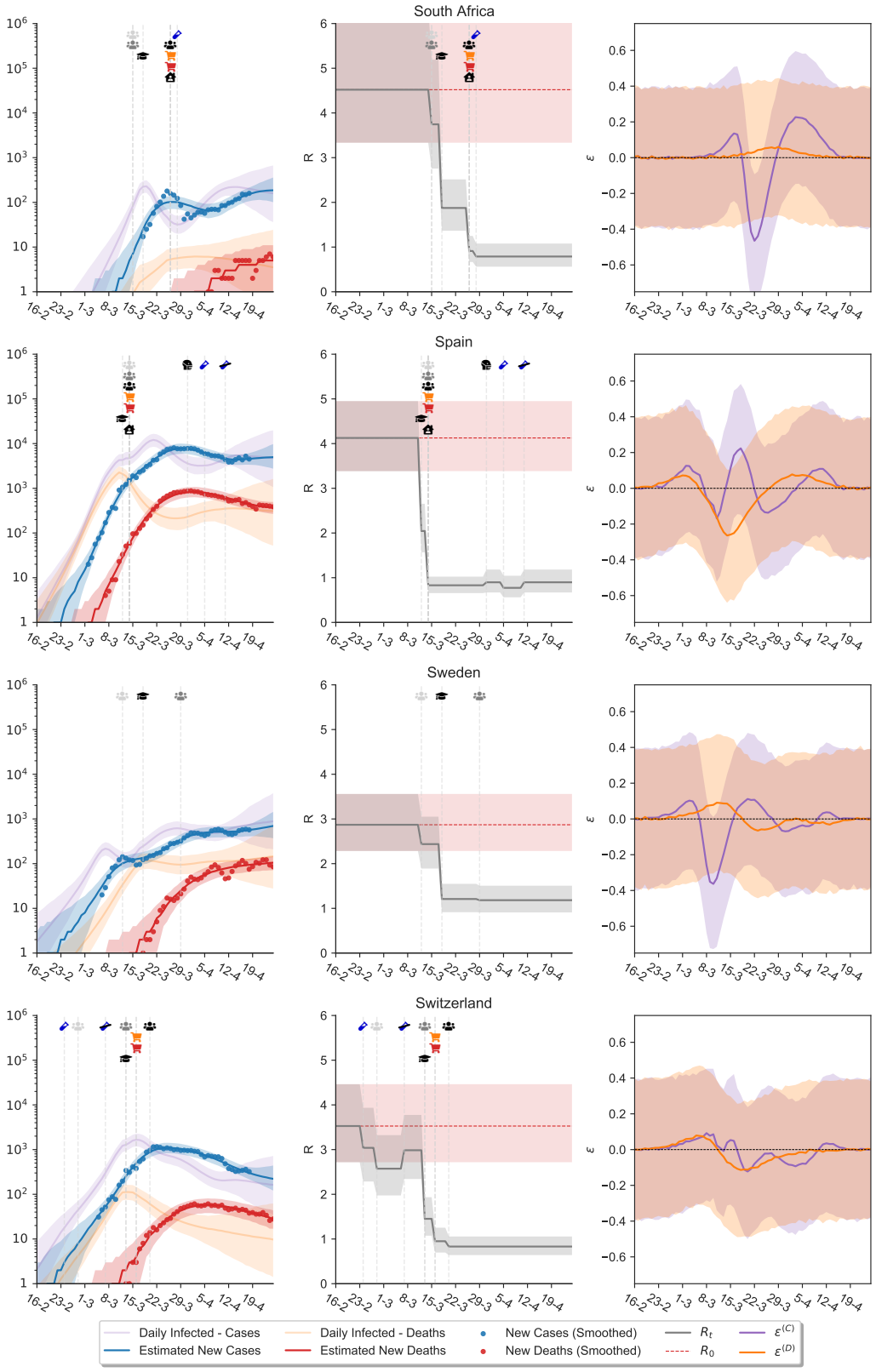
South Africa
Spain
Sweden
Switzerland
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.39

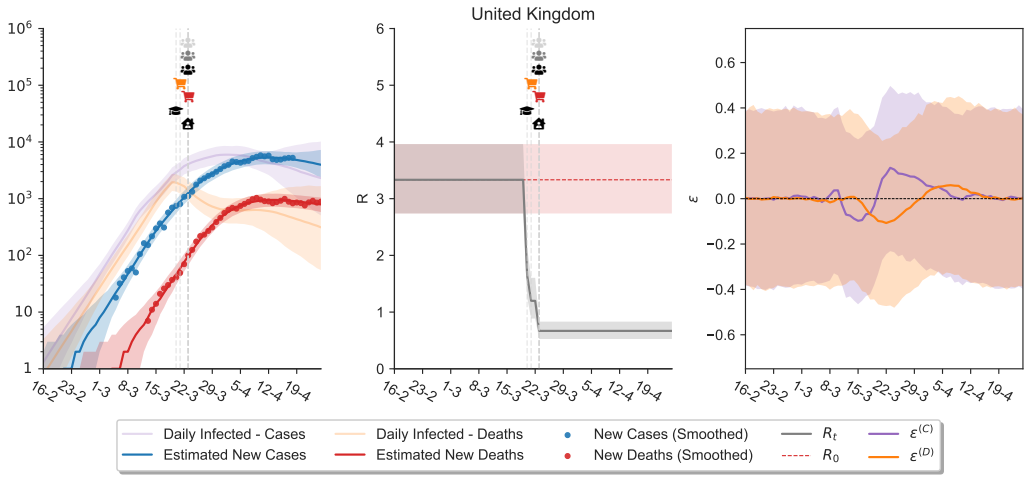
United Kingdom
R
ε
Daily Infected - Cases
Estimated New Cases
Daily Infected - Deaths
Estimated New Deaths
New Cases (Smoothed)
New Deaths (Smoothed)
R_t
R_0
$\varepsilon^{(C)}$
$\varepsilon^{(D)}$

Figure E.40

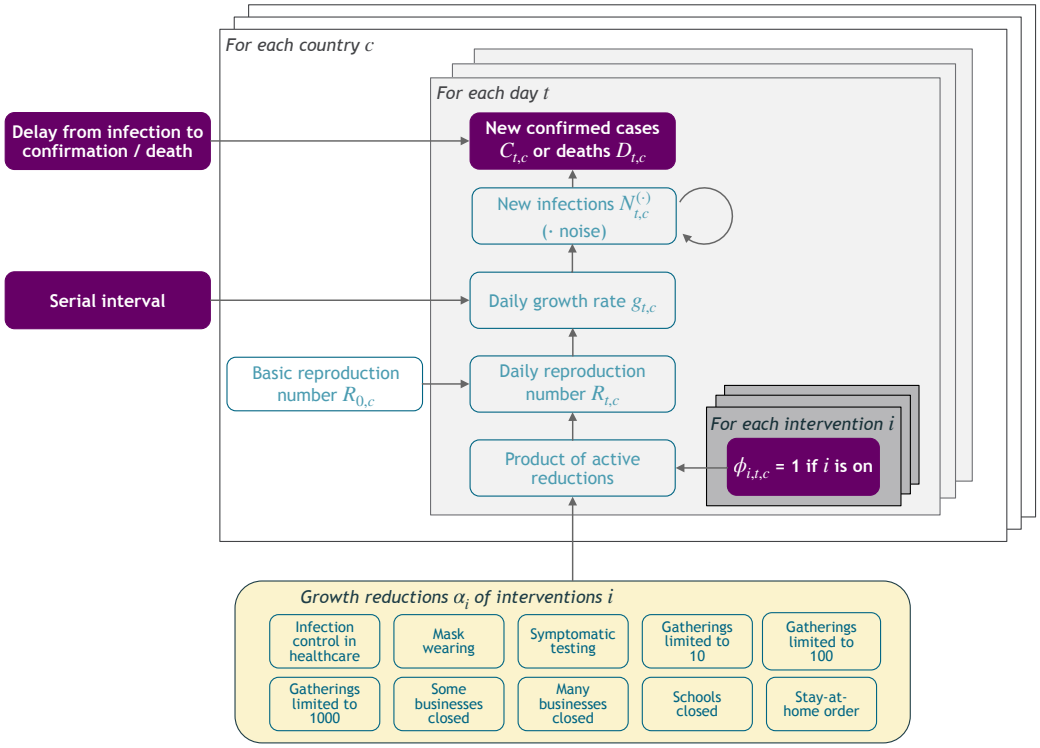

Figure F.41: Model overview. Purple nodes are observed or have a fixed distribution. The same structure is used for both deaths and confirmed cases. Our primary model combines both observations; it splits all nodes above the daily growth rate $g_{t,c}$ into separate branches for deaths and cases.

Appendix F. Additional Modeling Details

Appendix F.1. Data Preprocessing

We perform the following data preprocessing:

- Our data for confirmed cases and deaths is given by the John Hopkins Centre for Systems Science and Engineering[24,25]. We smooth this data by averaging the number of cases and deaths in a five day period around every day, assuming the data is symmetric at the boundaries.
- We mask new cases before a country has reached 100 confirmed cases. This accounts for cases being imported from other countries and rapid changes in testing regime when the case count is small.
- To avoid bias from imported deaths, we mask new deaths before a country has reached 10 deaths.
- Days where there are zero cases or deaths do not provide information about the *relative* change in the size of the epidemic. Therefore, they are masked.

Appendix F.2. Concise 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**
  1. **Country-specific** R_0,

$$R_{0,c} = \exp(\bar{R} + \sigma_R z_c) \tag{F.1}$$
$$\bar{R} \sim \texttt{Student T}(\mu = \log(3.25), \sigma = 0.2, \nu = 10) \tag{F.2}$$
$$\sigma_R \sim \texttt{Half Student T}(\sigma = 0.2, \nu = 10) \tag{F.3}$$
$$z_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 1) \tag{F.4}$$

  2. **NPI Effectiveness**:

$$\alpha_i \sim \texttt{Normal}(\mu = 0, \sigma_R^2 = 0.2) \tag{F.5}$$
$$\tag{F.6}$$

  3. **Infection Initial Counts**.

$$N_{0,c}^{(C)} = \exp(\zeta_c^{(C)}) \tag{F.7}$$
$$N_{0,c}^{(D)} = \exp(\zeta_c^{(D)}) \tag{F.8}$$
$$\zeta_c^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{F.9}$$
$$\zeta_c^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{F.10}$$
$$\tag{F.11}$$

  4. **Observation Noise Dispersion Parameter**

$$\Psi \sim \texttt{Half Normal}(\mu = 0, \sigma^2 = 5^2) \tag{F.12}$$

- **Hyperparameters**
  1. **Infection Noise Scale,** $\sigma_N = 0.1$ (selected by cross-validation).
  2. **Serial Interval Parameters**. The serial interval is assumed to have a `Gamma` distribution with $\alpha = 1.87$ and $\beta = 0.28$.[30]
  3. **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:

$$\mathcal{T}^{(C)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Gamma}(\mu = 18.8, \frac{\sigma}{\mu} = 0.45) \tag{F.13}$$

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, $\mathscr{T}^{(D)}$ is:[31–34]

$$\mathscr{T}^{(D)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Negative Binomial}(\mu = 5.25, \alpha = 1.57), \quad \text{(F.14)}$$

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 serial interval distribution. This is the exact conversion *under exponential growth*, following eq. (2.9) in Wallinga & Lipsitch.[29] (Note that we use daily growth rates.)
  3.

$$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(C)}\right], \quad \text{(F.15)}$$

$$N_{t,c}^{(D)} = N_{0,c}^{(D)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(D)})\right], \text{with noise} \quad \text{(F.16)}$$

$$\varepsilon_{\tau,c}^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2), \quad \text{(F.17)}$$

$$\varepsilon_{\tau,c}^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2) \quad \text{(F.18)}$$

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

$$\bar{C}_{t,c} = \sum_{\tau=1}^{32} N_{t-\tau,c}^{(C)} \pi_C[\tau], \quad \text{(F.19)}$$

$$\bar{D}_{t,c} = \sum_{\tau=1}^{64} N_{t-\tau,c}^{(D)} \pi_D[\tau]. \quad \text{(F.20)}$$

Finally, the output distribution follows a Negative Binomial noise distribution as proposed by Flaxman et al.[2]

$$C_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{C}_{t,c}, \alpha = \Psi) \quad \text{(F.21)}$$

$$D_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{D}_{t,c}, \alpha = \Psi) \quad \text{(F.22)}$$

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[56] with the NUTS MCMC sampling algorithm[37].

Appendix F.3. Interpreting α_i - Proof Sketch

We have previously noted that the effectiveness of each NPI, α_i, may depend on the presence of other NPIs. For example, masks may be less effective when a stay-at-home order has been issued because more of the remaining transmission occurs in private spaces. We claimed that, in such a situation, we can roughly interpret the inferred effect α_i of NPI i as the average additional effect it had in the *contexts* (i.e., the sets of simultaneously active NPIs) in which it was active. The average is over days and countries in which it was active.

Here, we formalize this claim for the maximum likelihood estimator (MLE) of α_i with a simplified model in which we know the true values of $R_{c,t}$ (perhaps from another model). In reality, these values are not known but rather estimated by our model. Although, we are performing Bayesian inference, the posterior density will be high where the likelihood is high, and thus this interpretation is still insightful. The maximum of our posterior (the MAP) will be close to the maximum of the likelihood (the MLE) since the influence of our prior distribution on α_i is, empirically, small.

Simplified Model. We have NPI activations $\phi_{i,c,t}$, where $\phi_{i,c,t} = 1$ represents NPI i being active in country c on day t. Assume that the true values of $R_{c,t}, R_{0,c}$ have been provided to us. Our simplified model is:

$$R_{c,t}^{\text{predicted}} = R_{0,c} \exp[-\sum_i \alpha_i \phi_{i,c,t}], \tag{F.23}$$

$$\log R_{c,t} = \log R_{c,t}^{\text{predicted}} + z_{c,t} \text{ where } z_{c,t} \sim \mathcal{N}(0, \sigma_N^2). \tag{F.24}$$

The log-likelihood can be written as:

$$\begin{aligned} \mathcal{L}(\{\alpha_i\}) &= \log p(\{R_{t,c}\} | \{R_{o,c}\}, \{\alpha_i\}, \{\phi_{i,c,t}\}) && \text{(F.25)} \\ &= \sum_{c,t} \log p(R_{t,c} | R_{o,c}, \{\alpha_i\}, \{\phi_{i,c,t}\}) \\ &= \sum_{c,t} \frac{-1}{2\sigma_N^2} (\log R_{c,t}^{\text{predicted}} - \log R_{c,t})^2 + \text{constant}. && \text{(F.26)} \end{aligned}$$

Taking derivatives with respect to α_i yields:

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \alpha_i} &\propto \sum_{c,t} (\log R_{c,t}^{\text{predicted}} - \log R_{c,t}) \frac{\log R_{c,t}^{\text{predicted}}}{\partial \alpha_i} \\ &\propto \sum_{c,t} (\log R_{c,t}^{\text{predicted}} - \log R_{c,t}) \phi_{i,c,t} \\ &\propto \sum_{c,t \in i \text{ active}} (\log R_{c,t}^{\text{predicted}} - \log R_{c,t}). \end{aligned} \tag{F.27}$$

Finally, setting $\frac{\partial \mathscr{L}}{\partial \alpha_i} = 0$ gives.

$$\begin{aligned}&\sum_{c,t \in i \text{ active}} (\log R_{c,t}^{\text{predicted}} - \log R_{c,t}) = 0 \\ \Rightarrow &\sum_{c,t \in i \text{ active}} (\log R_{0,c} - \sum_{j \neq i} \alpha_j \phi_{j,c,t} - \log R_{c,t}) = N_i \alpha_i^{\text{MLE}}, \end{aligned} \tag{F.28}$$

where N_i is the number of days that NPI i was active. Rearranging gives the desired result:

$$\alpha_i^{\text{MLE}} = \frac{1}{N_i} \sum_{c,t \in i \text{ active}} (\underbrace{[\log R_{0,c} - \sum_{j \neq i} \alpha_j \phi_{j,c,t}]}_{\text{Predicted} \log R \text{ based on other NPIs}} - \log R_{c,t}). \tag{F.29}$$

α_i^{MLE} is the average additional effect that NPI i had over the simultaneously active NPIs, where the average is taken over the days where NPI i was active.

Appendix F.4. Choice of σ_N

The value of σ_N is chosen by evaluating holdout country performance across a range of different values of σ_N.

Figure F.42 shows heldout predictive performance for The Netherlands across different values of σ_N. We choose values $\sigma_N = 0.2$ because it is the gives good holdout calibration. We included other countries in our analysis, leading to similar results.

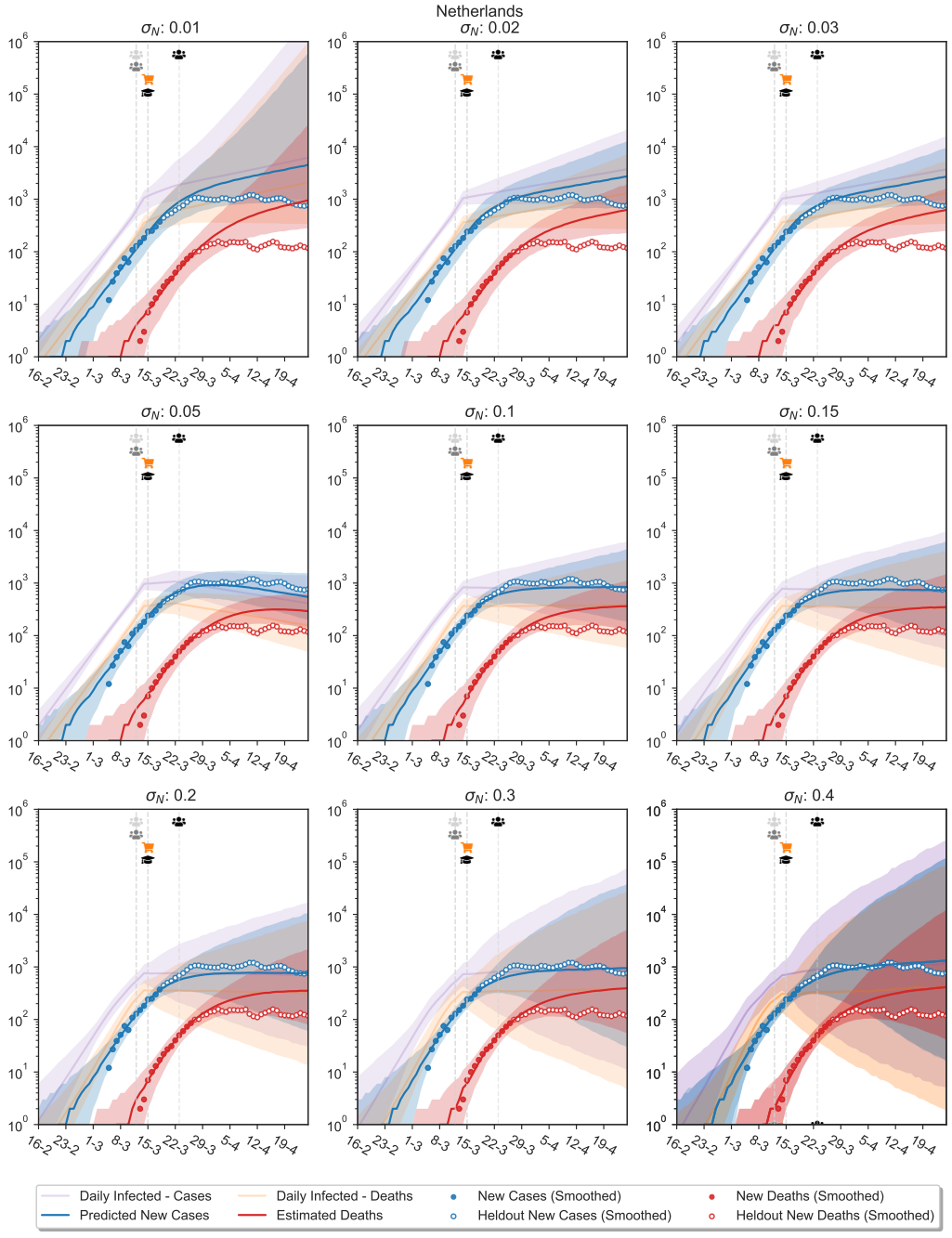

Figure F.42: Holdout performance for The Netherlands for a range of different noise scales, σ_N. Note that this graph was produced with a hyperprior on the effectiveness of each intervention, removed in the final version of the model.

Appendix G. Preference survey details

Appendix G.1. Description of NPIs as shown to participants

All school closed

All levels of schools are closed.

Restrictions on gatherings

All events and gatherings above a certain size are banned.

Most risky businesses closed

Selected businesses with a high risk of infection are closed, such as most restaurants or bars.

All non-essential businesses closed

Essential businesses like grocery stores and pharmacies remain open, but all other customer-facing businesses are closed.

Stay-at-home order

People are required to not leave their house, with exceptions for daily exercise, grocery shopping, and essential trips. Usually, this means that many non-essential businesses are closed as well. You can usually still go to work, but many companies will switch to work-from-home where possible.

Public health authorities tracing contacts

People who are infected have to share their contact history with epidemiologists and at-risk people are quarantined.

Special precautions in clinics and hospitals

People are screened for COVID before entering hospitals. People with COVID symptoms are given a face mask before they enter a clinic, or have to go to a dedicated COVID clinic.

Wearing masks

Wearing a face mask is mandatory when in the public.

Appendix G.2. Example question

This survey focuses on how socially and personally burdensome people perceive various COVID-19 mitigation measures to be.

In order to understand how to best react to the COVID-19 pandemic, we need to find out how different mitigation measures compare to each other. In this survey, we are only interested in how mitigation measures affect people's personal lives, but not in how effective different measures are at reducing the spread of COVID-19 nor what their effects are on the economy as a whole.

As such, we only ask about how different measures affect your life, not about how they affect the course of the pandemic.

Which of these mitigation measures would you find least burdensome, and which most burdensome?

The following shows a selection of mitigation measures that may occur as part of the response against Covid-19. Note that the measures differ in type and duration of deployment. Consider how burdensome would the measures be **if they had the same effect on the reduction of COVID spreading.**

Mitigation measure	**Best (Least burdensome)**	**Worst (Most burdensome)**
Stay-at-home order for 2 weeks	◯	◯
All non-essential businesses closed for 1 week	◯	◯
All schools closed for 3 months	◯	◯
All schools closed for 1 year	◯	◯
Wearing masks for2 weeks	◯	◯
Special precautions in clinics and hospitals for 1 week	◯	◯

Appendix G.3. Estimation of perceived intervention costs (ratio scale) from utility scores (interval scale)

Let $u(i,d)$ be the average population utility score for a pair of intervention i and duration d. We now make two additional assumptions, which are well justified by the empirical data (Figure 7):

1. The utility can be expressed as the sum of two terms, where one term only depends on the intervention and the other term only on the duration: $u(i,d) = a_i + b(d)$
2. The dependence on duration is logarithmic: $b(d) = b \ln d$

We can thus express the utility score as: $u(i,d) = a_i + b \ln d$

We can then define the cost of intervention i as $c_i = e^{\frac{a_i}{b}}$. This cost has the desired ratio property: The cost of intervention i_2 is x times larger than the cost of intervention i_1 iff the average survey participant would be indifferent between enduring i_2 for some duration d' and enduring i_1 for duration $x \cdot d'$.

Proof that the cost c_i has the desired ratio property:

$$
\begin{aligned}
& \frac{c_{i_2}}{c_{i_2}} = x \\
\Leftrightarrow \quad & \frac{e^{\frac{a_2}{b}}}{e^{\frac{a_1}{b}}} = x \\
\Leftrightarrow \quad & a_2 = a_1 + b \ln x \\
\Leftrightarrow \quad & b \ln d' + a_2 = a_1 + b \ln x + b \ln d' \\
\Leftrightarrow \quad & u(i_2, d') = u(i_1, x d')
\end{aligned}
$$

We use a linear model to find parameters a_i and b, using utility scores for all pairs of measures and durations.

Appendix G.4. Demographics

	Number	Percentage
Age		
21-29	25	7%
30-39	116	35%
40-49	66	20%
50-59	41	12%
60 or older	24	7%
No answer	62	19%
Gender		
Male	153	46%
Female	119	36%
Other	1	0%
No answer	61	18%
Martial status		
Never married	135	40%
Married	116	35%
Widowed	2	1%
Separated	2	1%
Divorced	15	4%
Other	2	1%
No answer	62	19%
Highest level of education		
Less than high school degree	1	0%
High school degree or equivalent (e.g. GED)	32	10%
Some college but no degree	57	17%
Associate degree	30	9%
Bachelor degree	127	38%
Graduate degree	28	8%
No answer	59	18%
Employment status		
Employed, working 1-39 hours per week	79	24%
Employed, working 40 or more hours per week	156	47%
Not employed, looking for work	9	3%
Not employed, NOT looking for work	7	2%
Disabled, not able to work	2	1%
Retired	14	4%
Student	1	0%
Other	6	2%
No answer	60	18%
Pre-tax household income		
$0-9,999$	8	2%
$10,000-19,999$	19	6%
$20,000-29,999$	24	7%

Continued on next page

Table G.9 – *Continued from previous page*

	Number	Percentage
30,000–39,999	35	10%
40,000–49,999	34	10%
50,000–59,999	24	7%
60,000–69,999	23	7%
70,000–79,999	35	10%
80,000–89,999	14	4%
90,000–99,999	13	4%
$100,000 or more	40	12%
No answer	65	20%

Appendix H. Assumptions and limitations

Appendix H.1. Limitations of the data

We only record NPIs implemented nationally. For example, several regions in Germany implemented stay-at-home orders even though this was not ordered nationally. Regional orders do not appear in our data. Additionally, while we included more NPIs than previous work (Table 1), 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 H.2. Model Limitations

Independence of country and time. We assume that the effect of NPIs on growth rates is similar across countries and time. However, the exact implementation and adherence of each NPIs is likely to vary. Our uncertainty estimates in Figure 6 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.

Unobserved changes in behavior. Our method assumes that changes in the reproduction number are caused by the observed NPIs rather than unobserved factors such as spontaneous behaviour changes. We test the sensitivity of our results to unobserved interventions by hiding observed NPIs and by including mobility data. Our conclusions were stable (see Figure D.15), but removing our most effective NPI, school closure, increased the inferred effectiveness for gathering bans and business closures.

Testing, reporting, and the IFR. Our model can account for differences in testing (and IFR/reporting) between countries and over time, as discussed in Section 2). 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 Section 3, our model only reports the average additional effect each NPI had in the contexts where it was active in our data (derivation in Appendix F). Figure 6 shows 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.

Growth rates. The functional form of the relationship between the daily growth rate in 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.

Signalling effect of NPIs. As we explained in Section 4 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.

Subgroups. We work under the standard assumption of a well-mixed population (Anderson & May[57]). 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 stable across countries.

Appendix H.3. Limitations of burden estimation

We estimate the burden that different NPIs put on people's lives. Of course, implementation of NPIs has many other costs (and benefits) than just the encumbrance on daily life. Many long-term costs of NPIs will also be codetermined by the economic policy response they engender, their impacts on global supply chains, their structural damage to networks of business contacts, and many other similar effects. Estimating these long-term impacts might be prohibitively difficult and is out of scope for this study. Nevertheless, these factors should be considered for policy decisions to the degree possible.

Our preference data is a sample of US residents only, in particular those working on the Amazon Mechanical Turk platform. This may limit the international applicability of our cost-effectiveness estimates. Even though recruitment on Amazon Mechanical Turk usually results in greater demographic diversity than typical internet samples,[38] there will still be selection bias. It's also important to note that, due to ethical reasons, the sample does not include participants under 18 years of age, which is a main limitation when estimating the perceived costs of closing schools.

Finally, using the mean population preference for policy decisions may be problematic in itself. For example, the closure of schools will likely strongly affect the parents of school children but pose little burden on the majority of people that are not parents of school children. The *mean* burden of closing schools may then just be moderate, but for policy decisions it is necessary to also take considerations around fairness and inequality into account.

References

1 World Health Organization. Non-pharmaceutical public health measures for mitigating the risk and impact of epidemic and pandemic influenza; 2019.

2 Flaxman S, Mishra S, Gandy A, Unwin H, Coupland H, Mellan T, et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries; 2020. Available from: https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-03-30-COVID19-Report-13.pdf.

3 Eichenbaum M, Rebelo S, Trabandt M. The Macroeconomics of Epidemics; 2020.

4 Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. The Lancet Psychiatry. 2020 jun;7(6):547–560.

5 Chen X, Qiu Z. Scenario analysis of non-pharmaceutical interventions on global COVID-19 transmissions; 2020. https://arxiv.org/abs/2004.04529.

6 Banholzer N, van Weenen E, Kratzwald B, Seeliger A, Tschernutter D, Bottrighi P, et al. Impact of non-pharmaceutical interventions on documented cases of COVID-19. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 apr;Available from: https://www.medrxiv.org/content/10.1101/2020.04.16.20062141v3.

7 Hale T, Webster S, Petherick A, Phillips T, Kira B. Oxford COVID-19 Government Response Tracker. Blavatnik School of Government; 2020. https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker.

8 ACAPS. #COVID 19 Government Measures Dataset; 2020. https://www.acaps.org/covid19-government-measures-dataset.

9 Louviere JJ, Woodworth GG. Best-worst scaling: A model for the largest difference judgments. University of Alberta: Working Paper. 1991;.

10 Flynn TN. Valuing citizen and patient preferences in health: recent developments in three types of best–worst scaling. Expert Review of Pharmacoeconomics & Outcomes Research. 2010 jun;10(3):259–267.

11 Adda J. Economic Activity and the Spread of Viral Diseases: Evidence from High Frequency Data. Institute of Labor Economics (IZA); 2015. 9326. Available from: http://ftp.iza.org/dp9326.pdf.

12 Naude J, Mellado B, Choma J, Correa F, Dahbi S, Dwolatzky B, et al. Worldwide Effectiveness of Various Non-Pharmaceutical Intervention Control Strategies on the Global COVID-19 Pandemic: A Linearised Control Model. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 may;Available from: https://www.medrxiv.org/content/early/2020/05/12/2020.04.30.20085316.

13 Siedner MJ, Harling G, Reynolds Z, Gilbert RF, Venkataramani A, Tsai AC. Social distancing to slow the U.S. COVID-19 epidemic: an interrupted time-series analysis. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 apr;Available from: https://www.medrxiv.org/content/10.1101/2020.04.03.20052373v2.

14 Kraemer MUG, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science. 2020 mar;368(6490):493–497.

15 Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases. 2020 may;20(5):553–558.

16 Dandekar R, Barbastathis G. Neural Network aided quarantine control model estimation of global Covid-19 spread;Available from: https://arxiv.org/abs/2004.02752.

17 Maier BF, Brockmann D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. Science. 2020 apr;368(6492):742–746.

18 Villas-Boas SB, Sears J, Villas-Boas M, Villas-Boas V. Are We #StayingHome to Flatten the Curve? UC Berkeley: Department of Agricultural and Resource Economics; 2020.

19 Jarvis CI, , Zandvoort KV, Gimma A, Prem K, Klepac P, et al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. BMC Medicine. 2020 may;18(1).

20 Orea L, Álvarez I. How effective has been the Spanish lockdown to battle COVID-19? A spatial analysis of the coronavirus propagation across provinces. FEDEA; 2020. 2020-03. Available from: http://documentos.fedea.net/pubs/dt/2020/dt2020-03.pdf.

21 Lorch L, Trouleau W, Tsirtsis S, Szanto A, Schölkopf B, Gomez-Rodriguez M. A Spatiotemporal Epidemic Model to Quantify the Effects of Contact Tracing, Testing, and Containment;Available from: https://arxiv.org/abs/2004.07641.

22 Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, et al. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. Proceedings of the National Academy of Sciences. 2020 apr;117(19):10484–10491.

23 Quilty BJ, Diamond C, Liu Y, Gibbs H, Russell TW, Jarvis CI, et al. The effect of inter-city travel restrictions on geographical spread of COVID-19: Evidence from Wuhan, China. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020;Available from: https://www.medrxiv.org/content/early/2020/04/21/2020.04.16.20067504.

24 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. 2020 may;20(5):533–534.

25 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. Github; 2020. https://github.com/CSSEGISandData/COVID-19.

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

27 Our World in Data. Number of tests per confirmed case vs. Total confirmed COVID-19 cases per million people;. (Accessed on 04/06/2020). https://ourworldindata.org/grapher/number-of-tests-per-confirmed-case-vs-total-confirmed-cases-of-covid-19-per-million-people?time=2020-03-31..2020-04-06.

28 Yadav S, Yadav PK. Basic Reproduction Rate and Case Fatality Rate of COVID-19: Application of Meta-analysis. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 may;Available from: https://www.medrxiv.org/content/10.1101/2020.05.13.20100750v1.

29 Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings of the Royal Society B: Biological Sciences. 2006 nov;274(1609):599–604.

30 Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen X, et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. The Lancet Infectious Diseases. 2020 apr;.

31 Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, mok Jung S, et al. Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 jan;Available from: https://www.medrxiv.org/content/10.1101/2020.01.26.20018754v2.

32 Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine. 2020 mar;382(13):1199–1207.

33 Cereda D, Tirani M, Rovida F, Demicheli V, Ajelli M, Poletti P, et al. The early phase of the COVID-19 outbreak in Lombardy, Italy;Available from: https://arxiv.org/abs/2003.09320.

34 Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 mar;Available from: https://www.medrxiv.org/content/10.1101/2020.03.03.20028423v3.

35 Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases. 2020 mar;.

36 Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid Monte Carlo. Physics Letters B. 1987 sep;195(2):216–222.

37 Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research. 2014;15(47):1593–1623. Available from: http://jmlr.org/papers/v15/hoffman14a.html.

38 Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk. Perspectives on Psychological Science. 2011 jan;6(1):3–5.

39 Lipovetsky S, Conklin M. Best-Worst Scaling in analytical closed-form solution. Journal of Choice Modelling. 2014 mar;10:60–68.

40 White M. bwsTools: Tools for Case 1 Best-Worst Scaling (MaxDiff) Designs;. (Accessed on 05/25/2020). https://cran.r-project.org/web/packages/bwsTools/index.html.

41 Sawtooth Software, Inc. Proceedings of the Sawtooth Software Conference; 2003. Available from: https://www.sawtoothsoftware.com/download/techpap/2003Proceedings.pdf.

42 Taylor JR. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books; 1997.

43 Gelman A, Carlin JB, Stern HS, Rubin DB. Model checking and improvement. In: Bayesian Data Analysis, Second Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis; 2003. Available from: https://books.google.com.mx/books?id=TNYhnkXQSjAC.

44 Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992 nov;7(4):457–472.

45 Google: COVID-19 Community Mobility Reports. See how your community is moving around differently due to COVID-19;. (Accessed on 05/26/2020). https://www.google.com/covid19/mobility/.

46 Piguillem F, Shi L. Optimal COVID-19 quarantine and testing policies. 2020;.

47 Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology. 2012 feb;66(1):8–38.

48 Rosenbaum PR, Rubin DB. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. Journal of the Royal Statistical Society: Series B (Methodological). 1983;45(2):212–218. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1983.tb01242.x.

49 Rosenbaum PR. Observational Studies. Springer New York; 2002.

50 Mehta NS, Mytton OT, Mullins EWS, Fowler TA, Falconer CL, Murphy OB, et al. SARS-CoV-2 (COVID-19): What do we know about children? A systematic review. Clinical Infectious Diseases. 2020 may;.

51 Zimmermann P, Curtis N. Coronavirus Infections in Children Including COVID-19. The Pediatric Infectious Disease Journal. 2020 may;39(5):355–368.

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

53 Jones TC, Mühlemann B, Veith T, Zuchowski M, Hofmann J, Stein A, et al. An analysis of SARS-CoV-2 viral load by patient age; 2020.

54 L'Huillier AG, Torriani G, Pigny F, Kaiser L, Eckerle I. Shedding of infectious SARS-CoV-2 in symptomatic neonates, children and adolescents. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 may;.

55 Fontanet A, Tondeur L, Madec Y, Grant R, Besombes C, Jolly N, et al. Cluster of COVID-19 in northern France: A retrospective closed cohort study. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. 2020 apr;.

56 Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science. 2016;2:e55.

57 Roy M Anderson PAL. A framework for discussing the population biology of infectious diseases. In: Infectious Diseases of Humans. OUP Oxford; 1992. .