On the Robustness of Effectiveness Estimation of Nonpharmaceutical Interventions Against COVID-19 Transmission

Mrinank Sharma, Sören Mindermann, Jan Brauner
Department of Computer Science,
University of Oxford
jan.brauner@eng.ox.ac.uk

Gavin Leech,
School of Computer Science,
University of Bristol

Tomáš Gavenčiak,
Independent researcher

Leonid Chindelevitch
Computational Epidemiology Lab,
School of Computing Science,
Simon Fraser University

Yarin Gal,
OATML,
University of Oxford

Abstract

There remains much uncertainty about the relative effectiveness of different non-pharmaceutical interventions (NPIs) against COVID-19 transmission. Several studies attempt to infer NPI effectiveness under specific sets of epidemiological assumptions, but even if they test robustness to epidemiological parameters, they rarely include structural sensitivity analyses. Such analyses would ensure that the inferences made are consistent under plausible alternative assumptions. Without these, effectiveness estimates cannot be used to guide policy. We investigate four model structures similar to a recent state-of-the-art Bayesian hierarchical model that models both confirmed cases and deaths, using a high-quality NPI dataset. We find that all but one of our models perform well when making long-range predictions in held-out countries and crucially, the conclusions of the models that perform well are remarkably consistent. We further investigate the common assumption that the effect of an NPI is independent of coactivated NPIs, and mathematically ground the interpretation of model inferences in settings where this assumption does not hold.

1 Introduction

Nonpharmaceutical interventions (NPIs), such as business closures, gathering bans, and stay-at-home orders, are a central part of the fight against COVID-19. Yet it is largely unknown how effective different NPIs are at reducing transmission [2, 7]. Better understanding is urgently needed to guide policy and help countries efficiently suppress the disease without putting unnecessary burden on the population. We can infer NPI effectiveness by assuming that the implementation of an NPI affects the course of a country's epidemic in a particular way, and then observing that course. Standard tools such as PyMC3 [26] and Stan [3] allow this inference to be performed in an automated way. This data-driven analysis allows us to disentangle the effects of individual NPIs, as different countries implemented different NPIs in different orders and at different points of their epidemics.

However, creating a good data-driven model of NPI effectiveness is challenging and must be accompanied by extensive validation. Firstly, an analysis of holdout predictive performance is required [11, 12]. If a model cannot predict cases and deaths in held-out countries and periods, there is no reason to trust its NPI effectiveness estimates. In our previous work, we analyse holdout performance by holding out the last 20 days in all countries in parallel, and holding out each country one at a time

Preprint. Under review.

[2]. However, holdout performance validation is often limited or absent in other previous work. The majority of studies do not report holdout performance [1, 4, 19, 20, 10, 15].[1] Flaxman et al. [7] hold out the last three days in all countries in parallel.

Secondly, results need to be robust to variations in all components of the analysis for which there is no strong justification for choosing one particular setting [28, 23]. This applies to the following components:

Epidemiological parameters.[2] Key epidemiological parameters remain uncertain, and conflicting results are common (Supplement Table 1).

Model structure. How NPIs interact is unknown, and the process of transmission can only be modeled with approximations [25, 9]. Indeed, models commonly make a large number of approximating assumptions (Section 2).

Data. Given the difficulty of collecting data during an ongoing pandemic, any analysis will be limited to a subset of countries and NPIs, and there is no strong justification for in-or excluding one particular country or NPI.

Hyperparameters (sometimes).[3] If a model has good holdout performance over a range of hyperparameter values, there is no strong justification for picking one particular value.

While sensitivity analyses are more common than holdout validation in previous work, often only a small subset of epidemiological parameters is examined and structural sensitivity analysis is absent. Flaxman et al. [7] check sensitivity to the serial interval and leaving out individual countries, fit the reproduction number (R, the expected number of infections directly generated by one infected individual) with a non-parametric model, and compare to an alternative model of R_0. Banholzer et al. [1] check the sensitivity of their results to the delay from infection to reporting, the threshold initial case count, influential single data points, the form of the influence function, and restricting NPI effectiveness to be positive. Jarvis et al. [15] varied the post-lockdown contact reduction among young people. Many NPI studies do not mention 'sensitivity' or 'model checking' at all [4, 22, 5, 17, 21, 20, 10].

Our contributions:

i) We aggregate in one place the many assumptions commonly used in NPI effectiveness studies.

ii) We reproduce two state-of-the-art models [2, 7] and construct three novel alternative models. We empirically validate them, finding that one of the state-of-the-art models performs poorly.

iii) We analyse the robustness of our previously reported NPI effectiveness results [2] across XYZ structurally different models with different modelling assumptions. We find that results and policy-relevant conclusions are remarkably robust to changes in epidemiological parameters, data, and model structure. This is by far the most extensive sensitivity analysis of NPI efectiveness results to date.

iv) All previous work assumes that the effect of an NPI does not depend on other active NPIs, on the country, or the date. We mathematically show how to interpret effectiveness estimates when these assumptions are violated. Furthermore, we provide empirical evidence that, *in our data*, the assumption that NPI effectiveness is independent of other active NPIs is not as significant as assuming the effect of each NPI is the same across countries.

2 Common assumptions in NPI modelling

We now reproduce the model of our previous work [2], using this as our baseline, and explicitly collect the assumptions used in the model. We discuss key assumptions in detail but defer a full discussion of the implications of all assumptions to the . We further list other works which make use of these assumptions and propose a number of plausible models with different structural assumptions. Note that all of the models we consider make the standard epidemiological assumption that the population is homogeneous and well-mixed. Since this widely-used assumption is well understood (see [14, 24]), we omit it from the following text.

[1] We focus on multi-NPI studies here, but holdout validation is also absent from all single-NPI studies we are aware of.

[2] By 'epidemiological parameters' we mean those that describe properties of the disease or NPIs investigated.

[3] By hyperparameters we mean unlearned parameters which do not correspond to properties of the disease or NPIs.

Notation. The basic reproduction number for country c (i.e. the reproduction number in the absence of any NPIs) is $R_{0,c}$. The time-varying (instantaneous [9]) reproduction number at time t in country c is $R_{t,c}$, which we use as the measure of transmission. We include other work that uses the discrete time growth rate $g_{t,c}$ as the measure of transmission instead (e.g., [1]), rewriting their assumptions in terms of $R_{t,c}$. $\phi_{i,t,c}$ are binary NPI activation features with $\phi_{i,t,c} = 1$ indicating that NPI i is active in country c at time t. $C_{t,c}$ and $D_{t,c}$ represent the number of daily reported cases and deaths respectively. The set of NPIs is denoted as $\mathcal{I}$. $N_{t,c}$ represents (scaled) numbers of new daily infections. $\alpha_i \in \mathbb{R}$ parameterizes the effectiveness of NPI i.

2.1 Baseline (Model 1)

Assumption 1. Epidemiological parameters are constant across countries and time [7, 1, 4, 19, 22, 2].

Remark. Supplement Table 1 outlines the epidemiological parameters required by all our models. In addition, our models place i.i.d. prior distributions over: NPI effectiveness, α_i; initial outbreak sizes, $N_{0,c}$; country-specific base reproduction rates $R_{0,c}$ (through a hyper-prior). Other than the prior over $N_{0,c}$, which is uninformative, we consider these priors to be epidemiological parameters and include them in our sensitivity analysis.

Assumption 2. The effectiveness of NPI i is independent of other NPIs being active [7, 1, 6, 2].

Remark. The manner and extent to which different NPIs interact is unclear, and depends on the specific NPIs. For example, *school closure* and *symptomatic testing* are unlikely to interact, whilst social distancing measures may reduce the effectiveness of *mask wearing*. See Section 5 for a more detailed discussion.

Assumption 3. The effectiveness of NPI i is independent of the country [7, 1, 4, 22, 2].

Assumption 4. The effectiveness of NPI i is independent of time [7, 1, 4, 22, 2].

Assumption 5. The effectiveness of NPI i is independent of its implementation date [7, 1, 4, 22, 19, 2].

Assumption 6. Each NPI has a multiplicative effect on $R_{t,c}$ [7, 1, 4, 2].

Assumption 7. $R_{t,c}$ depends only on $R_{0,c}$ and active NPIs $\{\phi_{i,c,t}\}_{i\in\mathcal{I}}$ [7, 1, 4, 2]. Corollary: each NPI has its full effect on $R_{t,c}$ immediately. In Appendix A.3, we discuss to what extent this assumption allows *causal* inference.

Assumptions 2 to 7 lead to:

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

with $\alpha_i > 0$ to be interpreted as NPI i being effective.

Let the *discrete time growth rate* be $g_{t,c}$, such that $N_{t,c} = g_{t,c} N_{t-1,c}$.

Assumption 8. *(Approximation).* It is valid to convert $R_{t,c}$ to $g_{t,c}$ under exponential growth [27]:

$$g_{t,c} = \exp\left(M_{\text{SI}}^{-1}(R_{t,c}^{-1})\right), \tag{2}$$

where M_{SI}^{-1} is the inverse of the moment-generating function of the serial interval distribution [2], [7] (in sensitivity analysis).

Assumption 9. The Infection Fatality Rate (IFR_c), the proportion of infected cases that subsequently die, and the Ascertainment Rate (AR_c), the proportion of infected cases that are subsequently reported positive (both in country c) change slowly over time [2].

Assumption 10. The effective growth rate, in expectation, is the same for both cases and deaths. [2].

We can now write:

$$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{t'=1}^{t} \left[g_{t',c} \cdot \exp\left(\varepsilon_{t',c}^{(C)}\right) \right], \quad N_{t,c}^{(D)} = N_{0,c}^{(D)} \prod_{t'=1}^{t} \left[g_{t',c} \cdot \exp\left(\varepsilon_{t',c}^{(D)}\right) \right] \tag{3}$$

with noise terms $\varepsilon_{t',c}^{(C)}, \varepsilon_{t',c}^{(D)} \sim \mathcal{N}(0, \sigma_g^2)$. $N_{t,c}^{(C)}$ and $N_{t,c}^{(D)}$ represent the daily infections on day t, in country c, which will become confirmed COVID-19 cases and fatalities, respectively.

Remark. The noise terms allow for small, gradual changes in the AR and IFR [2]. Crucially, noise $\varepsilon_{t',c}^{(C)}$ affects $N_{t,c}^{(C)}$ for all $t \geq t'$. Differences in IFR_c and AR_c are accounted for by latents $N_{0,c}^{(C)}$

and $N_{0,c}^{(D)}$, which represent initial outbreak sizes. Concretely, if the true number of infections in country c is the same as country c' $\forall t$ but c tests a greater proportion of the population, we can infer $N_{0,c} > N_{0,c'}$.

Remark. These noise terms also partially relax Assumption 7, as they can account for the effects of unobserved NPIs, providing that they are uncorrelated with the observed NPIs [6]. However, if unobserved NPI i is correlated with observed NPI j, the effect of NPI i may be attributed to NPI j.

Discrete convolutions produce the expected number of new reported cases $\bar{C}_{t,c}$ and deaths $\bar{D}_{t,c}$ on a given day:

$$\bar{C}_{t,c} = \sum_{\tau=1}^{31} N_{t-\tau,c}^{(C)} \pi_C[\tau], \quad \bar{D}_{t,c} = \sum_{\tau=1}^{63} N_{t-\tau,c}^{(D)} \pi_D[\tau], \tag{4}$$

where $\pi_C[\tau]$ represents the probability of the delay between infection and confirmation being τ days (likewise for $\pi_D[\tau]$), produced by summing the incubation period with the onset-to-death (or confirmation) distributions and discretising.

Assumption 11. The output distribution of confirmed cases $C_{t,c}$ and deaths $D_{t,c}$ follows a Negative Binomial noise distribution [7, 2, 1].

$$C_{t,c} \sim \text{NB}(\mu = \bar{C}_{t,c}, A = \Psi), \qquad D_{t,c} \sim \text{NB}(\mu = \bar{D}_{t,c}, A = \Psi) \tag{5}$$

A is the dispersion parameter of the distribution, with larger values of A corresponding to less noise; Ψ is its inferred estimate. This distribution is suitable as it has support over $\mathbb{N}_0$, and has independent mean and variance parameters.

The models we describe next branch from this baseline model. Please refer to the Supplement for full model descriptions.

2.2 Additive Effect Model (Model 2)

To investigate Assumption 6, we propose a model where interventions have additive effects on $R_{t,c}$.

Assumption 12. NPI i has a linear effect on $R_{t,c}$ by affecting a non-overlapping, constant proportion of transmission. The introduction of NPI i eliminates all transmission related to i.

This leads to:

$$R_{t,c} = R_{0,c} \left(\hat{\alpha} + \sum_{i \in \mathcal{I}} \alpha_i \left(1 - \phi_{i,t,c}\right) \right), \quad \text{with } \hat{\alpha} + \sum_{i \in \mathcal{I}} \alpha_i = 1, \tag{6}$$

$\alpha_i > 0 \; \forall i$ and $\hat{\alpha} > 0$. α_i is the proportion of transmission eliminated by introducing NPI i.

2.3 Noisy-R Model (Model 3)

Instead of adding noise to $N_{t,c}$ as in Eq. 3, we could add noise directly to $R_{t,c}$ as:

$$R_{t,c}^{(C)} = \bar{R}_{t,c} \exp \varepsilon_{t,c}^{(C)}, \quad R_{t,c}^{(D)} = \bar{R}_{t,c} \exp \varepsilon_{t,c}^{(D)}, \tag{7}$$

where $\varepsilon_{t,c}^{(C)}, \varepsilon_{t,c}^{(D)} \sim \mathcal{N}(0, \sigma_R^2)$. This noise has the same implications as in Eq. 3, so it would be concerning if results were not robust across these models.

2.4 Discrete Renewal Model (Model 4)

Model 4 is based on Flaxman et al. [7].

Assumption 13. IFR_c and AR_c are constant over time. [7, 19, 1].

Remark. While the *true* IFR may be approximately constant, reporting is imperfect, and the proportion of COVID-19 related deaths captured by official statistics changes over time. Indeed, Fig. 4 (Supplement) shows the ratio between reported deaths and excess deaths, i.e., the deaths in excess of the historical average: their ratio changes notably in every country studied. As the modelled IFR_c is the ratio of *confirmed* COVID deaths to infections, we thus do not expect it to be constant over time. Furthermore, the assumption that AR_c is constant is especially problematic, since we expect testing capacity to vary over time. Nevertheless, these assumptions are common.

We investigate Assumption 8 by using a discrete renewal process ([8, 22]) in place of Eq.(3). Let $\pi_{SI}[\tau]$ represent the discretised serial interval distribution. We then write:

$$N_{t,c}^{(C)} = R_{t,c} \sum_{\tau=1}^{t} N_{t-\tau,c}^{(C)} \cdot \pi_{SI}[\tau], \quad N_{t,c}^{(D)} = R_{t,c} \sum_{\tau=1}^{t} N_{t-\tau,c}^{(D)} \cdot \pi_{SI}[\tau], \tag{8}$$

Modelling only deaths yields the model of Flaxman et al. [7] with $N_{c,t}^{(D)}$ taking the place of $\text{IFR}_c \cdot N_{c,t}^{\text{true}}$ in [7]. While our baseline partially relaxes Assumption 7, this model relies on $R_{0,c}$, active NPIs, and output noise to explain unobserved effects.

2.5 Different Effects Model (Model 5)

Our previous models share Assumptions 2 and 3 which we now relax. Denote the effectiveness of NPI i in country c as $\alpha_{i,c}$.

Assumption 14. For each NPIs i, $\{\alpha_{i,c}\}_c$ are drawn i.i.d. according to $\mathcal{N}(\alpha_i, \sigma_\alpha^2)$. σ_α is a noise scale hyper-parameter.

Remark. The effect of NPI i in country c may depend on the other NPIs which are active (since Assumption 2 may not hold) as well as the exact implementation of i in that country. Thus Assumption 14 relaxes both Assumptions 2 and 3.

3 Experiments

Data. We use data from [2], which comprises fact-checked data on the implementation of 9 NPIs in 41 countries between January and April 2020, including data on reported COVID-19 cases and deaths from the Johns Hopkins CSSE tracker [16]. See the Supplement for pre-processing details.

Implementation. We implement our models in PyMC3 [26] with NUTS [13] for inference. We typically use 4 chains, 2000 samples per chain (occasionally 8 chains, 1000 samples each). We ensure that Gelman-Rubin $\hat{R}$ convergence is less than 1.05 (Supplement Fig. **??**) and that there are no divergent transitions. Our sensitivity analyses and model implementations are available here.

Hyperparameters. Noise scales (σ_g or σ_R, σ_α) are chosen to give good holdout performance.

Model evaluation. We evaluate holdout performance by 4-fold cross-validation, holding out all but the first 14 days of cases and deaths (to allow estimation of $R_{0,c}$ and $N_{0,c}$). We only evaluate on countries with >100 deaths, as we want to evaluate models by their accuracy and calibration over long time periods, and countries with fewer deaths usually have fewer relevant days. We measure holdout performance as the held out log-likelihood, averaged over days, outputs, and countries.

As in our previous work [2], we perform extensive sensitivity experiments across 7 categories. *Epidemiological parameter sensitivity*, including: varying the infection-to-death and infection-to-confirmation delay distribution means, the serial interval distribution mean, the prior distributions on α_i (NPI effectiveness), and the hyperprior distribution on R_0. *Data sensitivity*, including: leaving out one country one at a time; leaving out one NPI at a time; varying the cumulative-case-threshold below which days are masked; switching the *school closure* NPI in Sweden on/off [4]. To evaluate sensitivity, we record posterior effectiveness estimates α_i under each experiment condition.

We summarise the sensitivity of a model using two *worst-case categorised sensitivity losses*: $\mathcal{L}_{\text{med}}$, which describes the sensitivity in median effectiveness, and $\mathcal{L}_\sigma$, which describes the sensitivity of the model's posterior standard deviation, or confidence.

$$\mathcal{L}_{\text{med}} = \sum_{c \in \text{categories}} \max_{i \in \mathcal{I}, \text{test} \in c} \left\{ \left| \text{median}[\tilde{\alpha}_i^{(\text{test})}] - \text{median}[\tilde{\alpha}_i^{(\text{default})}] \right| \right\} \tag{9}$$

test $\in c$ represents a specific test of category c. For example, c might represent the category of varying the distribution of a specific epidemiological parameter, and test would correspond to one particular value for the parameter(s) of that distribution. $\tilde{\alpha}_i$ is the effectiveness of NPI i converted into a percentage reduction of R, $\tilde{\alpha}_i^{(\text{default})}$ is the effectiveness under default data and epidemiological

[4] Sweden closed high schools and universities, but not elementary schools. Brauner et al. [2] counted this as "schools closed", but Banholzer et al. [1] counted this as "schools open".

parameters. The definition of $\mathcal{L}_\sigma$ is analogous, but the median is replaced by the standard deviation of the effectiveness. Larger values of $\mathcal{L}_{\text{med}}$ and $\mathcal{L}_\sigma$ correspond to more sensitive models. We take the maximum over NPIs and tests because we find that tests often significantly affect a small subset of the NPIs.

4 Results & Discussion

Baseline model (Model 1). We first reproduce the model, results, and sensitivity analyses of our previous work [2]. The key conclusions of this work are summarised in Supplementary Table 2. For brevity, we do not further discuss the results and their implications here, but refer the reader to [2]. The results are remarkably robust to changes in the data and hyperparameters, as well as across plausible ranges of epidemiological parameters (Supplement Fig. 6). In particular, the results are robust to leaving out one NPI at a time, indicating that the model can successfully ignore unobserved confounders (i.e., is robust to violations of Assumption 7). We tend to see systematic trends when varying epidemiological parameters (delays and serial interval), suggesting that, while they may affect effectiveness estimates, they do not affect conclusions about relative effectiveness.

Model comparison. We implement 4 alternative models with different model structure (Models 2 - 5), tuning their hyperparameters and comparing them based on holdout predictive performance and robustness. Models 1,2,3, and 5 have very similar holdout performance (average log-likelihood of -6.8, -6.6, -5.7, -6.6, respectively). The renewal model of Flaxman et al. [7], despite fitting the training data well, has worse holdout performance (average log-likelihood of -7.5). Fits and holdout predictions are shown in the Supplement. Models 1,2,3, and 5 also have broadly comparable robustness, with the renewal model faring significantly worse (Fig. 1). In general, the median effectiveness of each model varies more than the model's confidence, for most categories.

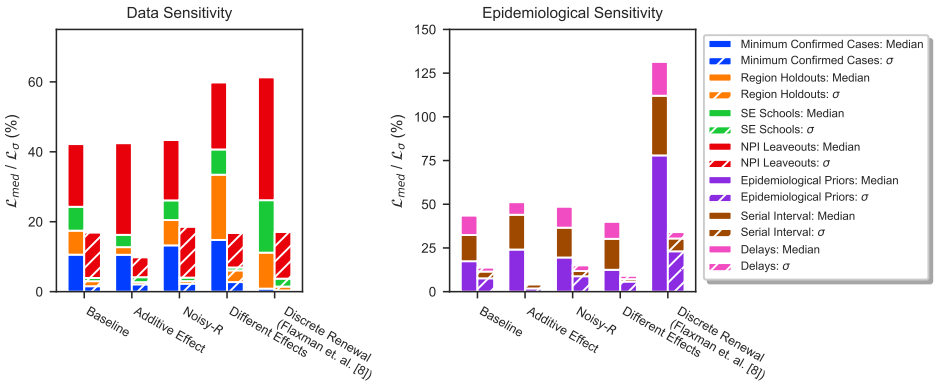

Figure 1: Contributions to sensitivity score from different categories: *Left*: categorised sensitivity to data perturbations, by model. *Right*: categorised sensitivity to epidemiological parameters.

Structural sensitivity analysis. We only require results to be robust across different model structures if there is no strong justification for choosing one structure over the other. The poorer holdout performance and robustness of the renewal model ([7]), justifies not choosing it, so we do not include it in our structural sensitivity analysis. Fig. 2 displays the inferred NPI effectiveness for models 1,2,3, and 5, using the same "best guess" (default) setting for epidemiological parameters for all models. The inferred NPI effectiveness estimates are remarkably robust across all models based on multiplicative effects. The results of the additive model cannot be directly compared as α_i have different meaning between in the additive and the multiplicative models. However, trends in the result are consistent across additive and multiplicative models.

It is particularly interesting to evaluate the robustness of high-level conclusions that might guide policymakers. Ranked by median effectiveness, we find remarkable robustness in NPI rankings (Fig. 2R): for example, the most and least effective NPI are the same across more than 150 experiment settings. Furthermore, we list policy-relevant conclusions from Brauner et al. [2], operationalise them, and count the fraction of tested parameter and model structure settings each conclusion holds in (Supplement Table 3). Again, all high-level conclusions are highly robust.

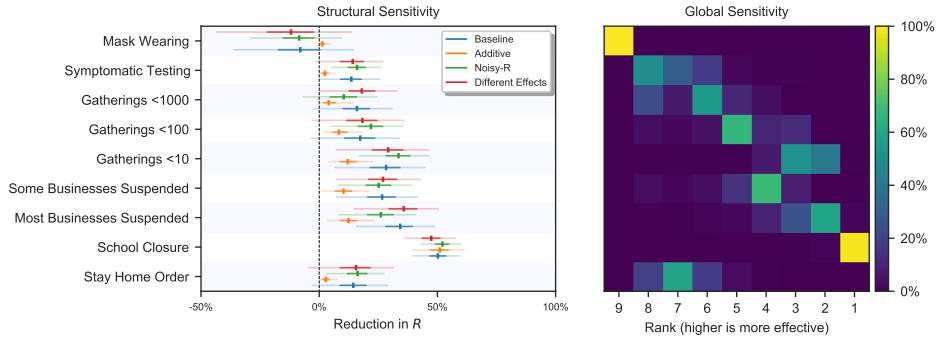

Figure 2: *Left*: NPI effectiveness under different structural assumptions, using the same "best guess" (default) setting for epidemiological parameters for all models. The figure shows the median, inter-quartile range and 95% confidence intervals of the posterior distributions of α_i. *Right*: Distribution of NPI effectiveness ranks. We performed more than 150 different experiments (4 structurally different models, 7 categories of data and parameter variation, several settings per category). The colour of the square indicates in what percentage of experiments the NPI was the xth most effective (judged by median posterior effectiveness).

5 Context-dependent Effectiveness

Except Model 5, our models rely on Assumptions 2 and 3, i.e., they assume an NPI's effect is independent of the other NPIs that are active and of the implementing country. In practice, these assumptions are likely violated. For instance, *mask wearing* probably has a greater effect on R when no social distancing measures are in place. Furthermore, the implementation of an NPI differs across countries: Brauner et al. [2] record *school closure* as active in both the UK (where schools but not universities were ordered to close) and Sweden (where universities and secondary schools, but not primary schools were ordered to close), not to mention that models also assume that NPI effects are also constant over time (Assumption 4) which appears dubious.

How should effectiveness estimates be interpreted when these assumptions are violated? To gain insight, we assume that ground truth values of $g_{t,c}$ and $R_{0,c}$ have been provided to us. Consider Model 1, where the noise is applied to $g_{t,c}$ and Model 3, where the noise is applied to R.

Model 1. $g_{t,c} = g(R_{t,c}) \exp(\varepsilon_{t,c})$, with $R_{t,c} = R_{0,c} \prod_{i \in \mathcal{I}} \exp(-\alpha_i \phi_{i,t,c})$.
Model 3. $g_{t,c} = g(R_{t,c})$, with $R_{t,c} = R_{0,c} \exp(\varepsilon_{t,c}) \prod_{i \in \mathcal{I}} \exp(-\alpha_i \phi_{i,t,c})$.
As usual, $\varepsilon_{t,c} \sim \mathcal{N}(0, \sigma^2)$. Recall that Assumption 8 lets us write $\log g(R) = \beta \left(R^{1/\nu} - 1\right)$, where ν is the shape and β is the inverse scale of the serial interval distribution, assumed to be Gamma(ν, β) [2, 7]. We have used the well known analytical expression for $M_{\text{SI}}(\cdot)$. Let $\Phi_i = \{(t,c) | \phi_{i,t,c} = 1\}$ be the days and countries with NPI i active. Let $\tilde{R}_{(-i),t,c} = R_{0,c} \prod_{j \in \mathcal{I} \setminus \{i\}} \exp(\alpha_j \phi_{j,t,c})$ i.e., $\tilde{R}_{(-i),t,c}$ is the predicted R ignoring the effect of NPI i. We present the following results in terms of $\exp(-\alpha_i)$, the factor by which NPI i reduces R.

Theorem 1. The Maximum Likelihood (ML) solution of α_i, given $\{\alpha_j\}_{j \neq i}$, under Model 3 satisfies:

$$\exp(-\alpha_i) = \frac{\left(\prod_{(t,c) \in \Phi_i} R_{t,c}\right)^{1/|\Phi_i|}}{\left(\prod_{(t,c) \in \Phi_i} \tilde{R}_{(-i),t,c}\right)^{1/|\Phi_i|}} = \frac{M_0(\{R_{t,c}\}_{\Phi_i})}{M_0(\{\tilde{R}_{(-i),t,c}\}_{\Phi_i})}, \tag{10}$$

where $M_0(\mathcal{S})$ denotes the geometric mean of set $\mathcal{S}$. The ML solution for $\exp(-\alpha_i)$ is thus the ratio of two geometric means over all country-days when NPI i is active: the numerator is the mean of $R_{t,c}$ and the denominator is the mean of the (hypothetical) predicted value of $R_{t,c}$ if NPI i was deactivated.

Theorem 2. The ML solution of α_i, given $\{\alpha_j\}_{j \neq i}$, under Model 1 satisfies:

$$\exp(-\alpha_i) = \left(\sum_{(t,c) \in \Phi_i} \tilde{R}_{(-i),t,c}^{1/\nu} \bar{R}_{t,c}^{1/\nu}\right)^{\nu} \Bigg/ \left(\sum_{(t,c) \in \Phi_i} \tilde{R}_{(-i),t,c}^{1/\nu} \tilde{R}_{(-i),t,c}^{1/\nu}\right)^{\nu} = \frac{M_{1/\nu}^{W_i}(\{\bar{R}_{t,c}\}_{\Phi_i})}{M_{1/\nu}^{W_i}(\{\tilde{R}_{(-i),t,c}\}_{\Phi_i})} \tag{11}$$

where $M_{1/\nu}^{W_i}(\mathcal{S})$ is the generalized *weighted* mean of set a $\mathcal{S}$, with exponent $1/\nu$ and weights $W_i = \{w_{c,t} = (\tilde{R}_{(-i),t,c})^{\frac{1}{\nu}}\}$. $\bar{R}_{t,c}$ is the "observed" R that exactly corresponds to the observed g.

Proofs: See Supplement.

For both models, α_i only depends on the days and countries it is active, and only depends on other $\alpha_j, j \neq i$ for days and countries when both NPIs i and j are active. Notably, the minor variation in model structure gives a significant difference in ML solutions; the ML solution of Model 3 is a ratio of geometric means whilst that of Model 1 is a ratio of generalized weighted means with exponent $p = 1/\nu$ that weighs observations (t, c) higher when the effect of the other active NPIs $(-i)$ is smaller. In both models, however, when assumptions 2, 3 and 4 do not hold, α_i can be interpreted as an average additional effectiveness added to the simultaneously active NPIs, in our data.

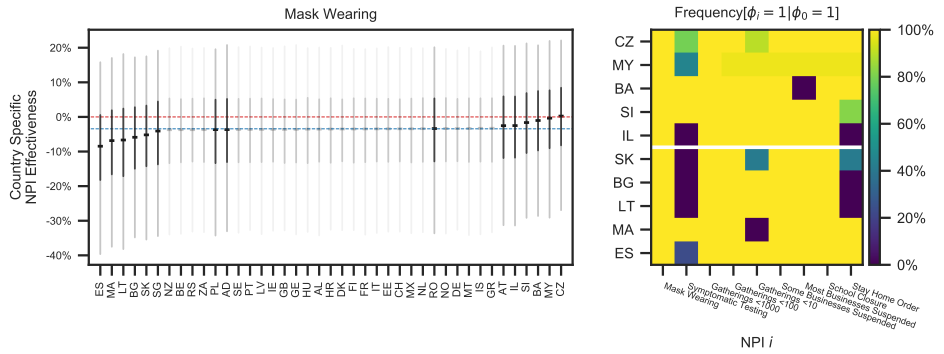

Figure 3: *Left*: Country-specific effectiveness of *mask wearing*, sorted by posterior median effectiveness (Different Effects Model). *Right*: Similar NPIs were co-activated with masks for countries where *mask wearing* was estimated to have the most effect (above white line) and the least effective (below line).

Empirical investigation. Recall that the Different Effects model relaxes Assumptions 2 and 3. Noise on α_c is able to model both country-level differences as well as interactions with other NPIs: if NPI j is less effective when NPI i is active, the noise will reduce the estimated effectiveness of NPI j in countries where i and j are both active. Fig. 3 shows country-specific effectiveness estimates for *mask wearing*, as well as conditional activation plots for the countries in which this NPI was most and least effective. If NPI effectiveness was heavily influenced by the presence of other NPIs, we would expect countries in which *mask wearing* was effective to have similar conditional-activation matrices (likewise for countries in which it was not effective). Surprisingly, we do not find this difference in NPI activation pattern for *mask wearing* or any other NPIs (Supplement, Figs. 8, 9). We know that Assumption 2 does not hold in practice, but we conclude that Assumption 3 is more significant *in this data*, since NPI effectiveness estimates seem to be more influenced by other country-specific factors than the presence or absence of other NPIs.

6 Conclusions

We perform by far the most extensive robustness analysis of NPI effectiveness estimates to date. We show that our previously reported NPI effectiveness results [2] are remarkably robust across a wide range of plausible epidemiological parameters, variations in the data, and several well-performing model structures. For a discussion of these results and their implications we refer the reader to [2]. While the robustness of these results is promising, the numerous assumptions and limitations inherent to data-driven NPI modelling imply that we should not treat these results as the last word on NPI effectiveness. Instead, decision-makers should draw on diverse sources of evidence, including other retrospective studies, experimental methods, and clinical experience. We have publicly released our sensitivity analysis suite and model implementations, and we urge those working on estimating NPI effectiveness to not only systematically validate their models, but also to report this.

Broader Impact

We validate the conclusions of a sophisticated COVID-19 countermeasure model, finding that the qualitative conclusions are robust across our benchmark sensitivity analyses. This finding is directly relevant to global policy.

The rapid pace of the COVID-19 research cycle has led to an increased number of erroneous and misreported findings reaching popular attention [18]. It is critical that such errors are caught before publication; the sensitivity analyses developed in this work can uncover faulty assumptions, and so prevent overconfidence or misinformation. We intend for our methodology to aid other modelling teams in producing highly reliable, policy-guiding estimates of NPI effects; to this end we release our sensitivity analysis suite and model implementations.

This work is written as many governments are selecting the time and order in which to lift NPIs, and anticipating second wave epidemics. It offers vital validation of the evidence, to help minimise harm to the world population.

One potential risk of this work stems from miscommunication: if readers came away mistaking high robustness (independence of conclusions from assumptions) for excessively high certainty. Even after extensive study of countries across the world, our estimates remain uncertain (Fig. 2L), and we expect both results and conclusions to change as more data and better modelling rolls in. In addition, the subtle issues of interpretation raised in Section 5 are difficult to convey to nontechnical audiences, and could easily be misread as unconditional effects or extrapolated outside the NPI context we have shown to be essential. Finally, the result on mask wearing also demands pre-emptive clarification: our confidence in the estimate is very low owing to the combination of a lack of data, late implementations, and high coactivation with other NPIs. It would be extremely regrettable if a lack of evidence was taken as a lack of effectiveness.

Acknowledgments and Disclosure of Funding

We thank Laurence Aitchison for helpful comments leading to the Additive Effect Model.

The following is a complete list of author funding sources: 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].

References

[1] Nicolas Banholzer, Eva van Weenen, Bernhard Kratzwald, Arne Seeliger, Daniel Tschernutter, Pierluigi Bottrighi, Alberto Cenedese, Joan Puig Salles, Werner Vach, and Stefan Feuerriegel. Impact of non-pharmaceutical interventions on documented cases of COVID-19. *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv*, apr 2020. doi: 10.1101/2020.04.16.20062141. URL `https://www.medrxiv.org/content/10.1101/2020.04.16.20062141v3`.

[2] Jan Markus Brauner, Mrinank Sharma, Sören Mindermann, Anna B Stephenson, Tomáš Gavenčiak, David Johnston, John Salvatier, Gavin Leech, Tamay Besiroglu, George Altman, Hong Ge, Vladimir Mikulik, Meghan Hartwick, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal, and Jan Kulveit. The effectiveness and perceived burden of nonpharmaceutical interventions against COVID-19 transmission: a modelling study with 41 countries. *medRxiv*, 2020. doi: 10.1101/2020.05.28.20116129. URL `https://www.medrxiv.org/content/early/2020/05/29/2020.05.28.20116129`.

[3] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[4] Xiaohui Chen and Ziyi Qiu. Scenario analysis of non-pharmaceutical interventions on global COVID-19 transmissions. `https://arxiv.org/abs/2004.04529`, 2020.

[5] Raj Dandekar and George Barbastathis. Neural network aided quarantine control model estimation of global covid-19 spread. *arXiv*, 2020. URL https://arxiv.org/abs/2004.02752.

[6] Jonas Dehning, Johannes Zierenberg, F Paul Spitzner, Michael Wibral, Joao Pinheiro Neto, Michael Wilczek, and Viola Priesemann. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, 2020.

[7] S Flaxman, S Mishra, A Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman, N Schmit, L Cilloni, K Ainslie, M Baguelin, I Blake, A Boonyasiri, O Boyd, L Cattarino, C Ciavarella, L Cooper, Z Cucunuba Perez, G Cuomo-Dannenburg, A Dighe, A Djaafara, I Dorigatti, S Van Elsland, R Fitzjohn, H Fu, K Gaythorpe, L Geidelberg, N Grassly, W Green, T Hallett, A Hamlet, W Hinsley, B Jeffrey, D Jorgensen, E Knock, D Laydon, G Nedjati Gilani, P Nouvellet, K Parag, I Siveroni, H Thompson, R Verity, E Volz, C Walters, H Wang, Y Wang, O Watson, P Winskill, X Xi, C Whittaker, P Walker, A Ghani, C Donnelly, S Riley, L Okell, M Vollmer, N Ferguson, and S Bhatt. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Technical report, Imperial College London, 2020. URL https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-03-30-COVID19-Report-13.pdf.

[8] S Flaxman, S Mishra, A Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman, N Schmit, L Cilloni, K Ainslie, M Baguelin, I Blake, A Boonyasiri, O Boyd, L Cattarino, C Ciavarella, L Cooper, Z Cucunuba Perez, G Cuomo-Dannenburg, A Dighe, A Djaafara, I Dorigatti, S Van Elsland, R Fitzjohn, H Fu, K Gaythorpe, L Geidelberg, N Grassly, W Green, T Hallett, A Hamlet, W Hinsley, B Jeffrey, D Jorgensen, E Knock, D Laydon, G Nedjati Gilani, P Nouvellet, K Parag, I Siveroni, H Thompson, R Verity, E Volz, C Walters, H Wang, Y Wang, O Watson, P Winskill, X Xi, C Whittaker, P Walker, A Ghani, C Donnelly, S Riley, L Okell, M Vollmer, N Ferguson, and S Bhatt. Code for modelling estimated deaths and cases for COVID-19 from report 13 published by MRC Centre for Global Infectious Disease Analysis, Imperial College London: Estimating the number of infections and the impact of nonpharmaceutical interventions on COVID-19 in 11 European countries. https://mrc-ide.github.io/covid19estimates/#/interventions, 2020.

[9] Christophe Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PloS one*, 2(8), 2007.

[10] Marino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19):10484–10491, apr 2020. doi: 10.1073/pnas.2004978117.

[11] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis, Second Edition*, chapter Model checking and improvement. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2003. ISBN 9781420057294. URL https://books.google.com.mx/books?id=TNYhnkXQSjAC.

[12] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

[13] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. URL http://jmlr.org/papers/v15/hoffman14a.html.

[14] A. Huppert and G. Katriel. Mathematical modelling and prediction in infectious disease epidemiology. *Clinical Microbiology and Infection*, 19(11):999 – 1005, 2013. ISSN 1198-743X. doi: https://doi.org/10.1111/1469-0691.12308. URL http://www.sciencedirect.com/science/article/pii/S1198743X14630019.

[15] Christopher I. Jarvis, , Kevin Van Zandvoort, Amy Gimma, Kiesha Prem, Petra Klepac, G. James Rubin, and W. John Edmunds. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Medicine*, 18(1), may 2020. doi: 10.1186/s12916-020-01597-8.

[16] Johns Hopkins University Center for Systems Science and Engineering. COVID-19 data repository by the center for systems science and engineering (CSSE) at johns hopkins university. https://github.com/CSSEGISandData/COVID-19, 2020.

[17] Moritz U. G. Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M. Pigott, Louis du Plessis, Nuno R. Faria, Ruoran Li, William P. Hanage, John S. Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G. Pybus, and Samuel V. Scarpino and. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490):493–497, mar 2020. doi: 10.1126/science.abb4218.

[18] Diana Kwon. How swamped preprint servers are blocking bad coronavirus research. https://www.nature.com/articles/d41586-020-01394-6, 2020. *Nature News*.

[19] Joseph Chadi Lemaitre, Javier Perez-Saez, Andrew Azman, Andrea Rinaldo, and Jacques Fellay. Assessing the impact of non-pharmaceutical interventions on SARS-CoV-2 transmission in switzerland. *medRxiv*, 2020. doi: 10.1101/2020.05.04.20090639. URL https://www.medrxiv.org/content/early/2020/05/08/2020.05.04.20090639.

[20] Lars Lorch, William Trouleau, Stratis Tsirtsis, Aron Szanto, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. A spatiotemporal epidemic model to quantify the effects of contact tracing, testing, and containment. *arXiv*, 2020. URL https://arxiv.org/abs/2004.07641.

[21] Benjamin F. Maier and Dirk Brockmann. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, 368(6492):742–746, apr 2020. doi: 10.1126/science.abb4557.

[22] Jacques Naude, Bruce Mellado, Joshua Choma, Fabio Correa, Salah Dahbi, Barry Dwolatzky, Leslie Dwolatzky, Kentaro Hayasi, Benjamin Lieberman, Caroline Maslo, Kgomotso Monnakgotla, Xifeng Ruan, and Finn Stevenson. 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*, may 2020. doi: 10.1101/2020.04.30.20085316. URL https://www.medrxiv.org/content/early/2020/05/12/2020.04.30.20085316.

[23] Elaine O Nsoesie, Richard J Beckman, and Madhav V Marathe. Sensitivity analysis of an individual-based model for simulation of influenza epidemics. *PloS one*, 7(10), 2012.

[24] Özgür Özmen, James J Nutaro, Laura L Pullum, and Arvind Ramanathan. Analyzing the impact of modeling choices and assumptions in compartmental epidemiological models. *Simulation*, 92(5):459–472, 2016.

[25] Pamela Anderson Lee Roy M. Anderson. *Infectious Diseases of Humans*, chapter A framework for discussing the population biology of infectious diseases. OUP Oxford, 1992. ISBN 019854040X.

[26] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[27] J Wallinga and M Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274 (1609):599–604, nov 2006. doi: 10.1098/rspb.2006.3754.

[28] Jianyong Wu, Radhika Dhingra, Manoj Gambhir, and Justin V Remais. Sensitivity analysis of infectious disease models: methods, advances and their application. *Journal of The Royal Society Interface*, 10(86):20121018, 2013.

A Supplementary material for 'On the Robustness of Effectiveness Estimation of Nonpharmaceutical Interventions Against COVID-19 Transmission'

Contents

A.1 Additional Figures and Tables

Table 1: Key epidemiological parameters, with reported estimates for COVID-19. Epidemiological parameters remain uncertain, and conflicting results are common.

Parameter	Definition	Reported values	
Serial-Interval (SI) distribution	The distribution of the period (in days) taken for one infected individual to generate R new infections. Gamma distributed.	μ=5.1, σ=2.7 μ=7.5, σ=3.4 μ=6.7, σ=4.9	[3] [4] [5]
Incubation period distribution	The distribution of the period (in days) between infection and onset of symptoms. Gamma distributed.	μ=5.2, σ=2.5 μ=6.0, σ=3.1 μ=5.1, σ=4.4	[3] [6] [7]
Onset-to-confirmation distribution	The distribution of the period (in days) between symptom onset and case confirmation. Gamma in [3]; Negative Binomial in [5]; lognormal in [3].	μ=2.6, σ=3.22 μ=3.68, σ=16.6 μ=5.25, σ=4.8	[3] [3] [5]
Onset-to-death distribution	The distribution of the period (in days) between symptom onset and death. Gamma distributed.	μ=18.8, σ=8.5 μ=15, σ=6.9	[8] [6]

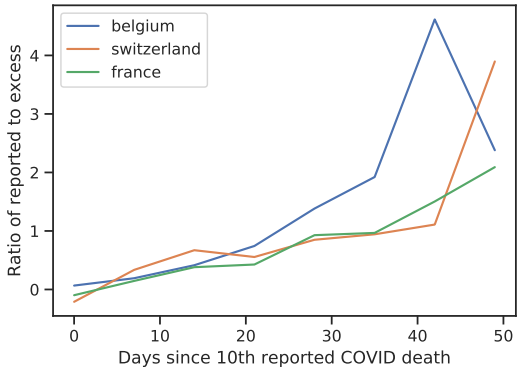

Figure 4: Ratio of reported COVID-19 deaths to overall excess deaths (deaths in excess of the historical average), over time & by country. At the start of the epidemic in each country, we might expect the number of excess deaths to closely track the true number of COVID-19 deaths. The ratio between reported COVID-19 deaths and excess changes notably over time, suggesting that Assumption 13 is violated. Data from [9].

Table 2: Conclusions of Brauner et al. [1]

Conclusion
Closing schools seems to play a surprisingly large role in reducing COVID-19 transmission.
Closing high-risk businesses, such as bars, restaurants, and gyms, appears only slightly less effective than closing most nonessential businesses (while imposing a substantially smaller burden on the population).
Testing of symptomatic patients has a demonstrable effect on reducing COVID-19 transmission.
A stay-at-home order (with exemptions) has a comparatively small effect on reducing COVID-19 transmission. (Explanation from Brauner et al. [1]: *"We estimate a comparatively small effect for stay-at-home orders. The 'stay-at-home order (with exemptions)' NPI 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. 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."*)

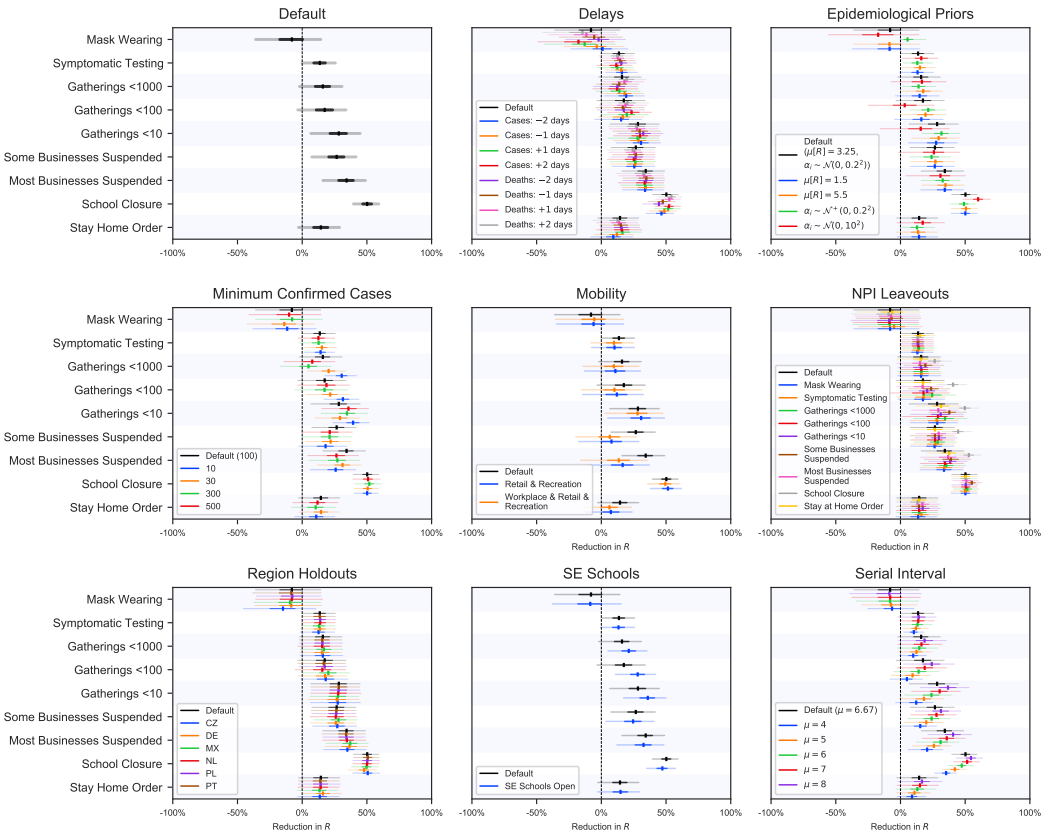

Figure 6: Categorised sensitivity analysis results for the baseline model. Median, inter-quartile range and 95% confidence intervals shown. Black values correspond to default parameter settings. Mobility shows the results if Google mobility data is added as an "NPI" as a proxy for unobserved behaviour changes. We expect some NPIs to be mediated through changes in mobility (e.g. business closures, stay-at-home orders) and thus expect to see changes in these NPIs, but not others. We include mobility in retail and recreation areas, and additionally in workplace areas.

Table 3: Robustness of the main conclusions of Brauner et al. [1]. *First column.* Brief restatement of the conclusion. See Table 2. *Second column.* Operationalisation to make the qualitative conclusions testable in the sensitivity analyses. *Third column.* Percentage of experiments in which the conclusion holds. We performed more than 150 different experiments (4 structurally different models, 7 categories of data and parameter variation, several settings per category). *The Additive Effects model constrains α_i to positive values, so we change the operationalisation of the conclusion about testing for the Additive Effects model, to "The probability of the 'Symptomatic testing' NPI reducing R by more than 1% is $\geq$ 90%." This conclusion is generally not supported by the additive effects model, which finds this to hold for only approximately 10% of sensitivity tests.

Conclusion	**Operationalisation**	**Holds in [%]**
Closing schools has a surprisingly large effect	The 'School closure' NPI has the highest median effectiveness of all NPIs.	100%
Closing most nonessential businesses has limited benefit over closing only high-risk businesses	The percentage reduction in R from closing high-risk businesses is at least 150% of the additional percentage reduction in R from closing most nonessential businesses (over closing just high-risk businesses).	100%
Symptomatic testing has a demonstrable effect	The posterior probability of the 'Symptomatic testing' NPI being effective is $\geq$ 90%.*	74.8%
A stay-at-home order (with exemptions) has comparatively small effect	The 'Stay-at-home order' NPI was in the bottom 4 NPIs in terms of median effectiveness.	100%

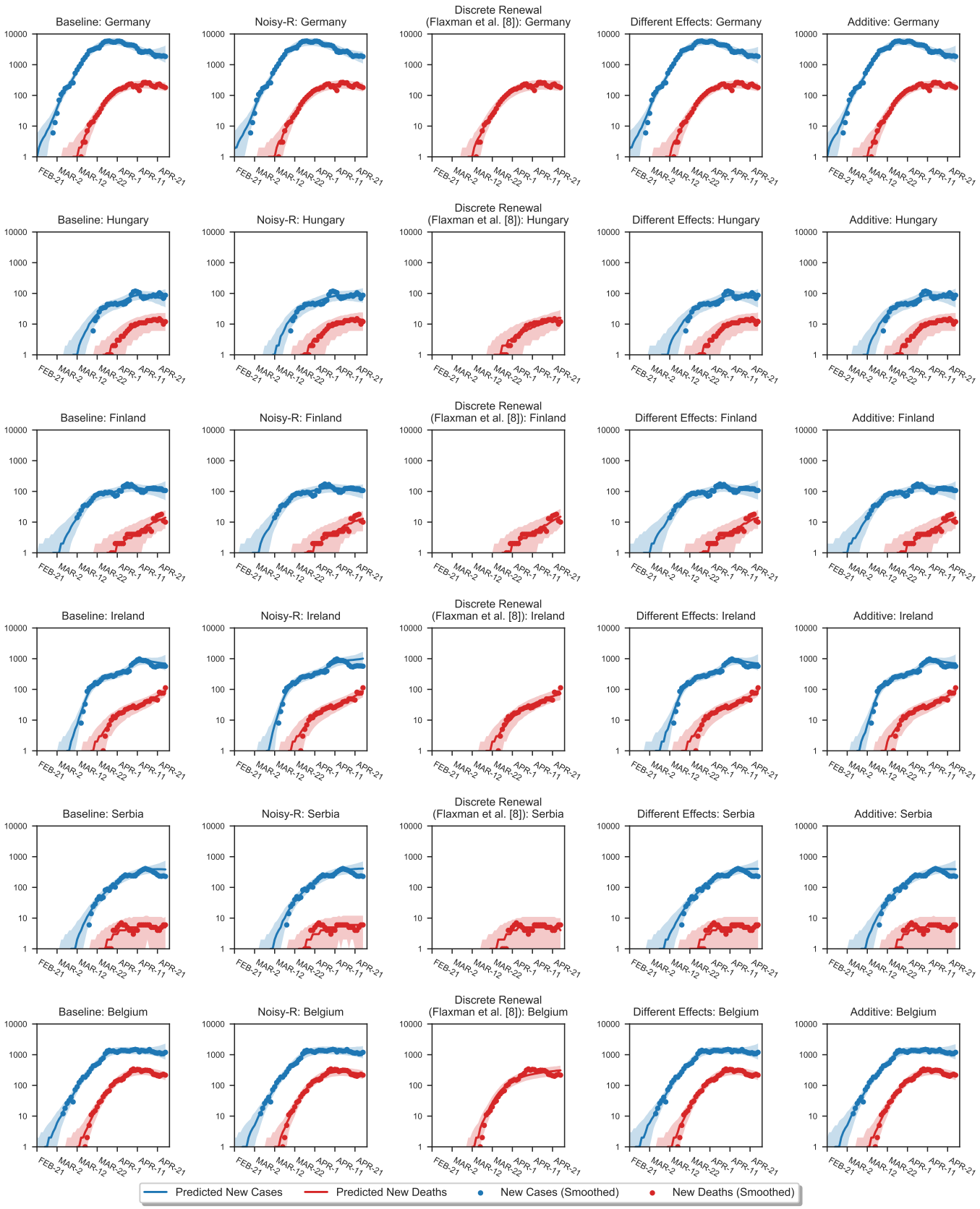

Figure 7a: Country fits, plotted for countries used in the first fold of cross-validation.

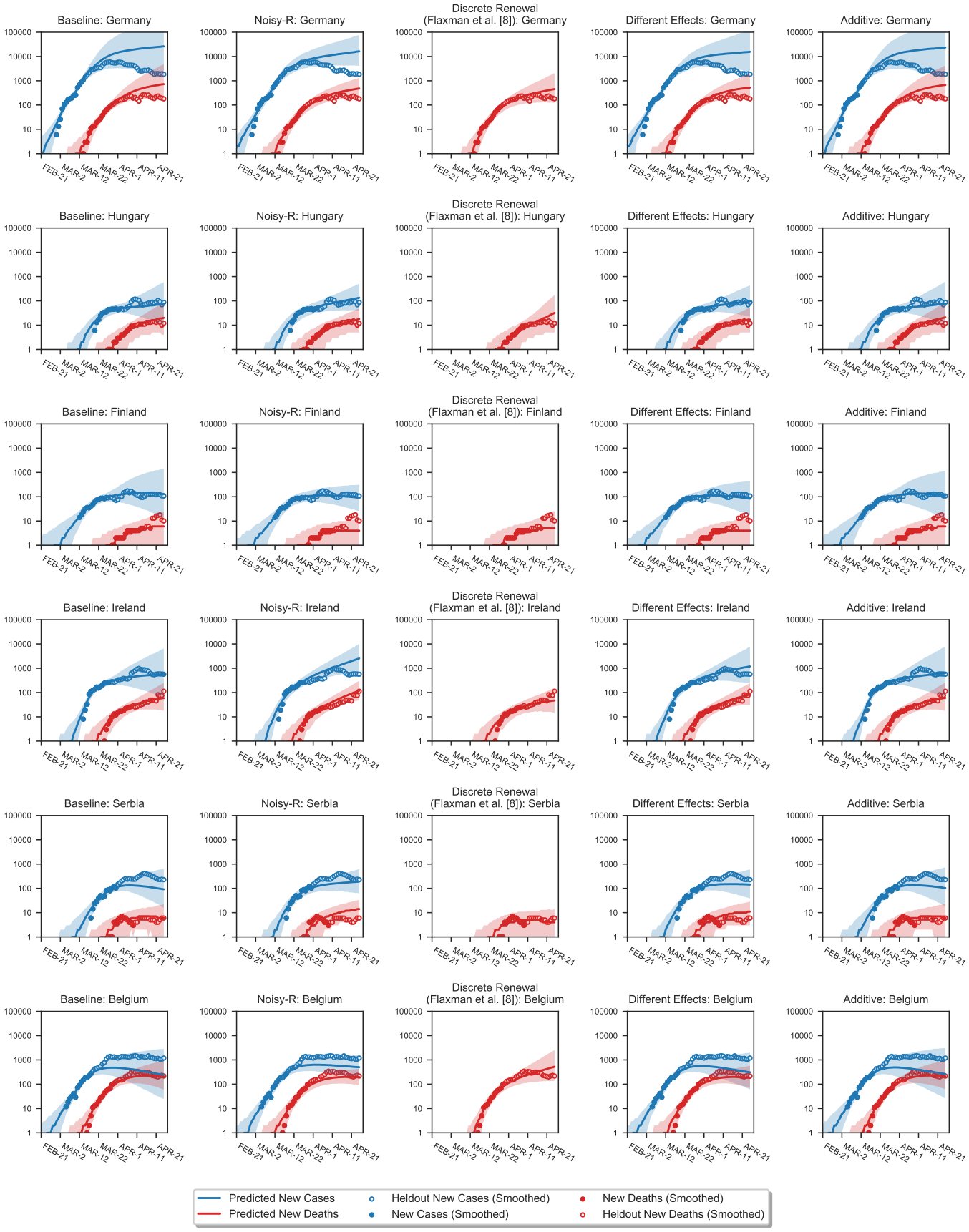

Figure 7b: Holdout country plots for the first fold of cross-validation.

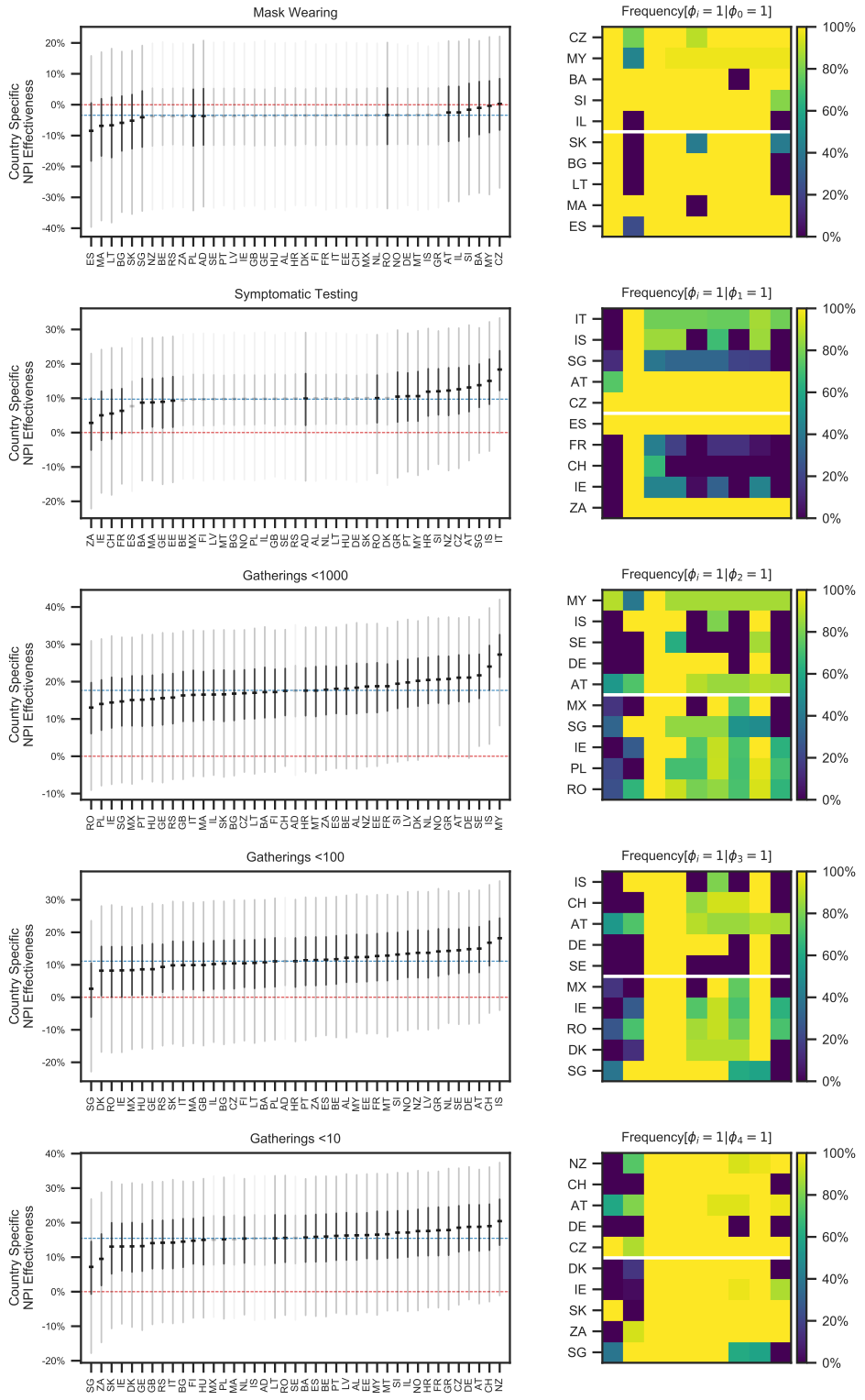

Figure 8: *Left*: Country-specific effectiveness for different NPIs, sorted by posterior median effectiveness (Different Effects Model). *Right*: Denote the NPI in question as NPI j. Similar NPIs are co-activated in the countries where NPI j is estimated to be the most effective (above white line) and the least effective (below white line). This suggests that country specific factors are more significant than the context independence assumption.

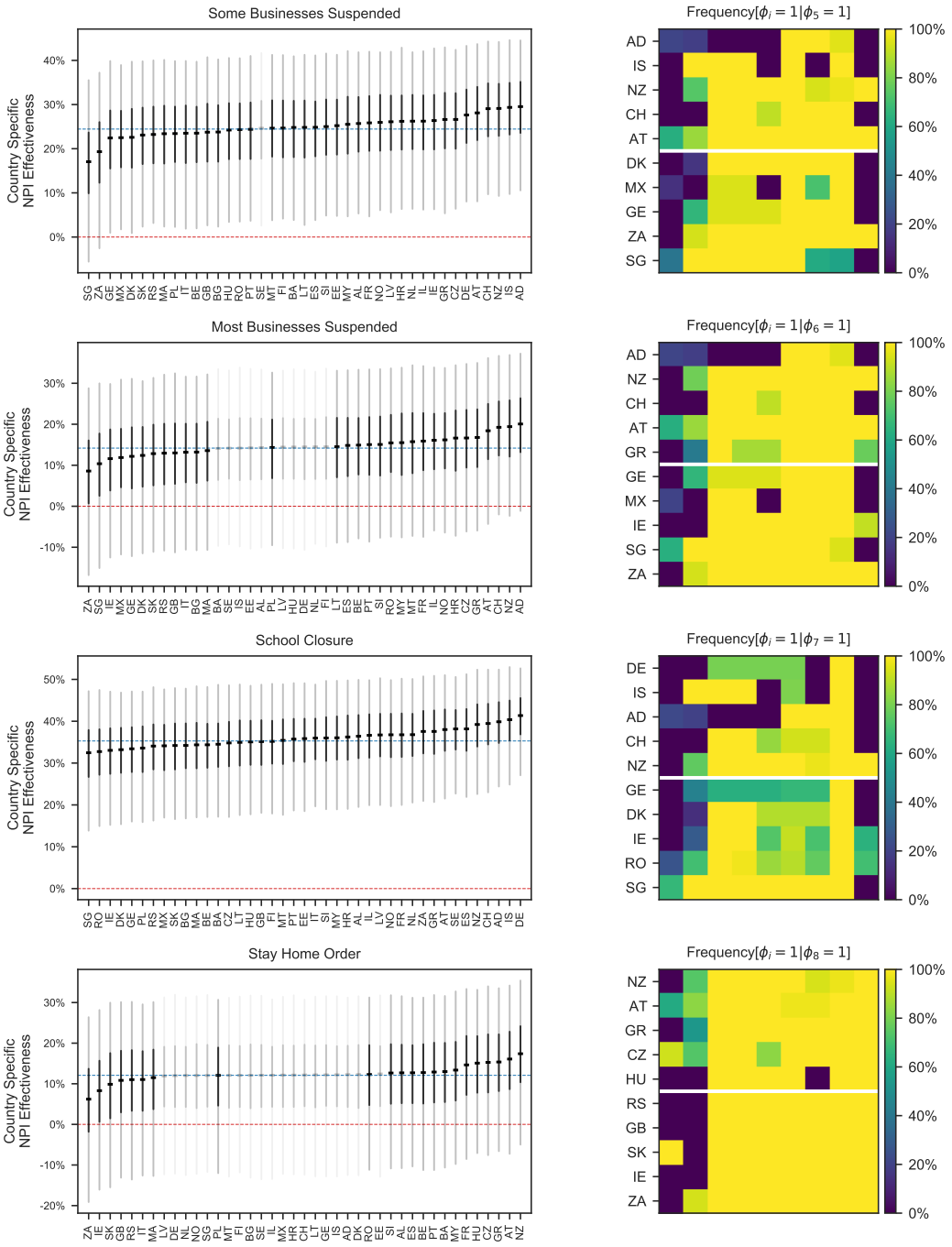

Figure 9: *Left*: Country-specific effectiveness for remaining NPIs, sorted by posterior median effectiveness (Different Effects Model). *Right*: Denote the NPI in question as NPI j. Similar NPIs are co-activated in the countries where NPI j is estimated to be the most effective (above white line) and the least effective (below white line). This suggests that country specific factors are more significant than the context independence assumption.

A.2 Full sensitivity results for all models

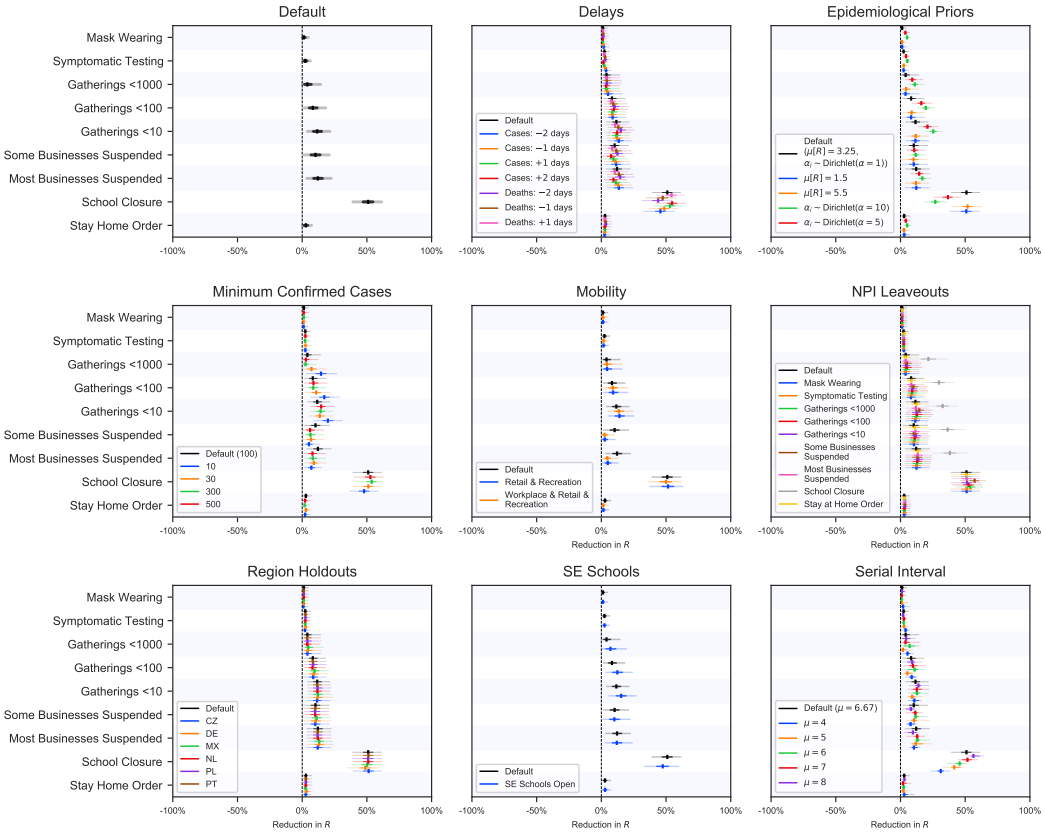

Figure 10: Categorised sensitivity analysis results for the additive model. Median, inter-quartile range and 95% confidence intervals shown. Black values correspond to default parameter settings. Mobility shows the results if Google mobility data is added as an "NPI" as a proxy for unobserved behaviour changes. We expect some NPIs to be mediated through changes in mobility (e.g. business closures, stay-at-home orders) and thus expect to see changes in these NPIs, but not others. We include mobility in retail and recreation areas, and additionally in workplace areas.

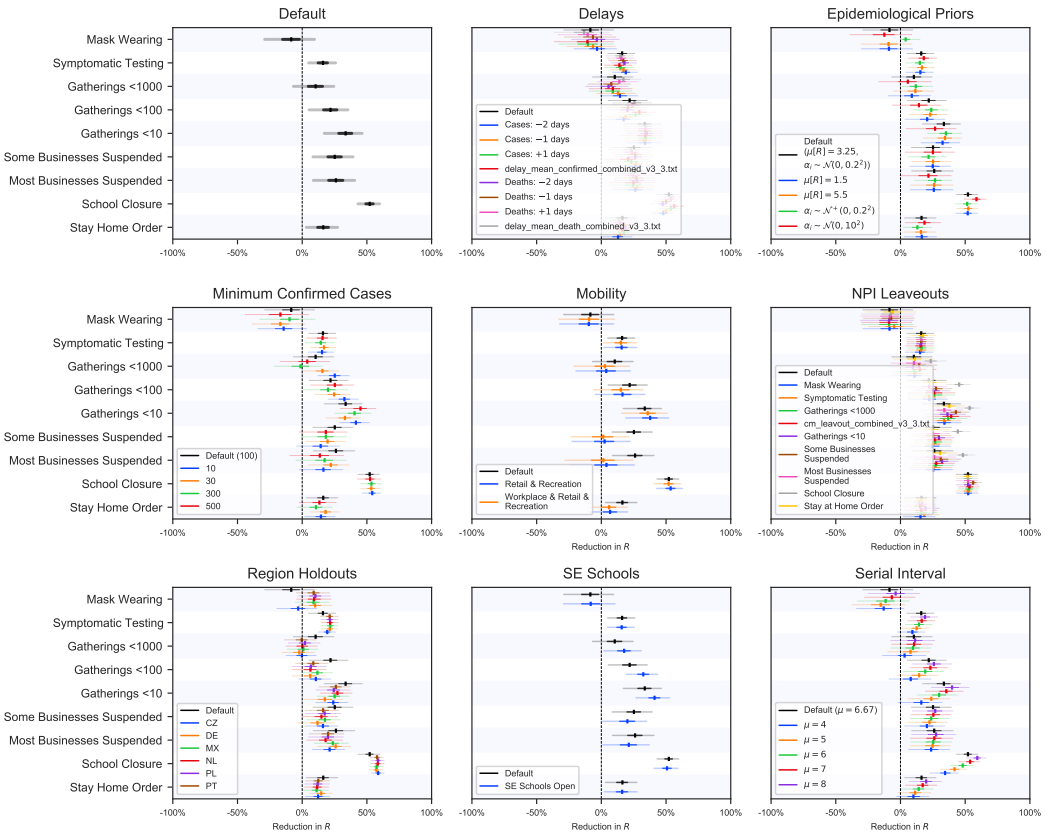

Figure 11: Categorised sensitivity analysis results for the Noisy-R model. Median, inter-quartile range and 95% confidence intervals shown. Black values correspond to default parameter settings. Mobility shows the results if Google mobility data is added as an "NPI" as a proxy for unobserved behaviour changes. We expect some NPIs to be mediated through changes in mobility (e.g. business closures, stay-at-home orders) and thus expect to see changes in these NPIs, but not others. We include mobility in retail and recreation areas, and additionally in workplace areas.

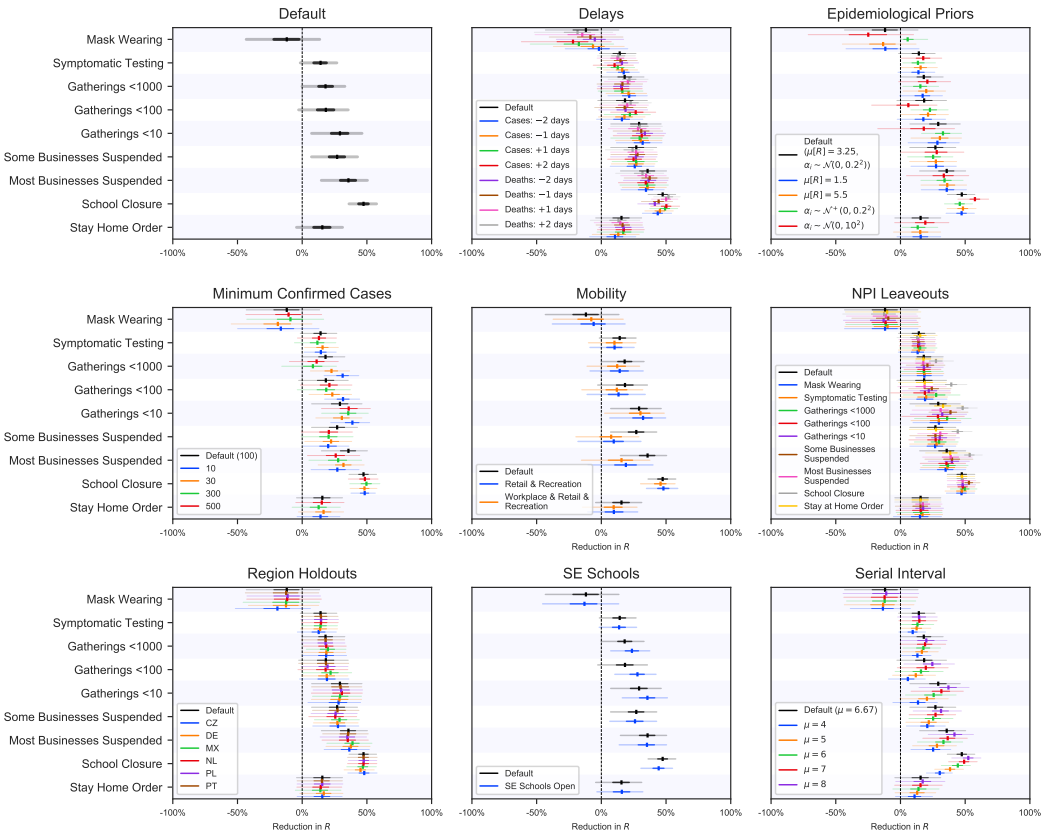

Figure 12: Categorised sensitivity analysis results for the different effects model. Median, interquartile range and 95% confidence intervals shown. Black values correspond to default parameter settings.Mobility shows the results if Google mobility data is added as an "NPI" as a proxy for unobserved behaviour changes. We expect some NPIs to be mediated through changes in mobility (e.g. business closures, stay-at-home orders) and thus expect to see changes in these NPIs, but not others. We include mobility in retail and recreation areas, and additionally in workplace areas.

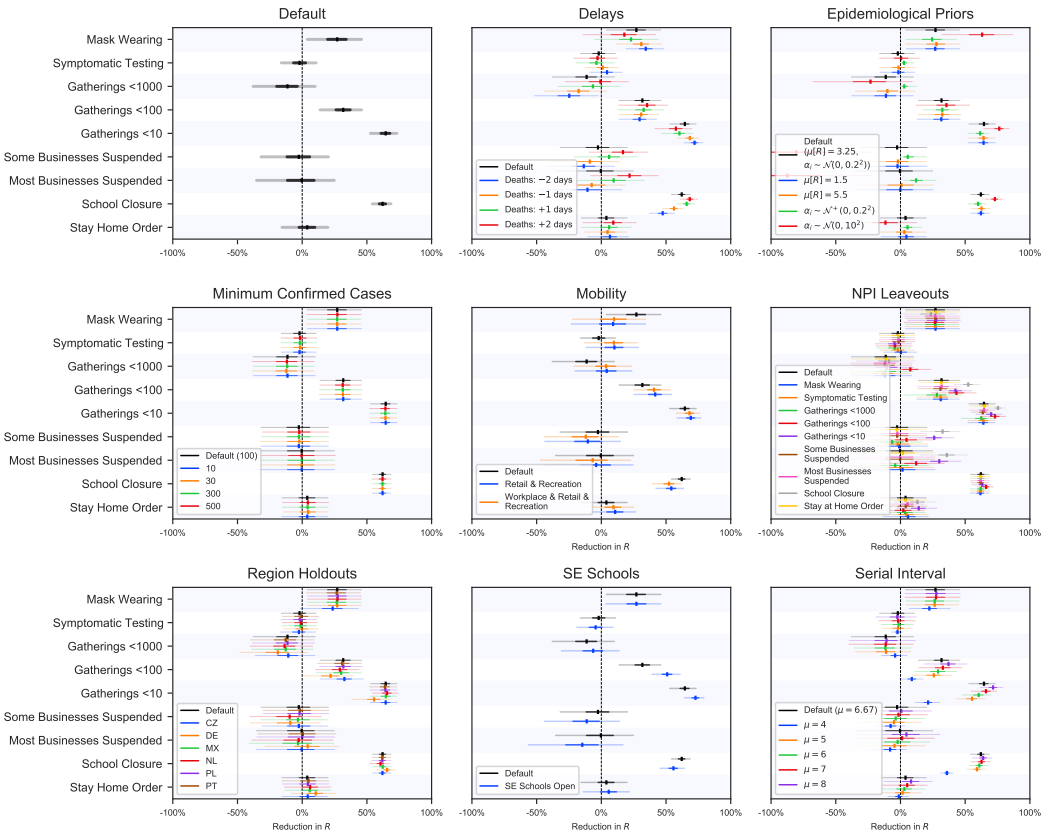

Figure 13: Categorised sensitivity analysis results for the discrete renewal model of [2]. Median, inter-quartile range and 95% confidence intervals are shown. Black values correspond to default parameter settings.Mobility shows the results if Google mobility data is added as an "NPI" as a proxy for unobserved behaviour changes. We expect some NPIs to be mediated through changes in mobility (e.g. business closures, stay-at-home orders) and thus expect to see changes in these NPIs, but not others. We include mobility in retail and recreation areas, and additionally in workplace areas.

A.3 Discussion of assumptions

Here we discuss assumptions (and their implications) listed in section 2 but not discussed in the main text, for which further discussion is necessary.

Assumption 6 states that each NPI's effect on $R_{t,c}$ is multiplicative. This implies that each NPI has a smaller effect when $R_{t,c}$ is already lowered by other NPIs. Such an assumption may be appropriate because e.g. an active stay-home order decreases the effect of wearing masks in public spaces. However, it may be inappropriate for other NPIs. For example, suppose a given proportion of transmission happens in schools and a given proportion in businesses and gatherings. In such a situation, closing schools is expected to decrease $R_{t,c}$ by a constant, whether or not businesses or gatherings are closed. This leads to an alternative model based on **Assumption 12**, where the effect of each NPI is additive (reprinted from equation (6)):

$$R_{t,c} = R_{0,c}\left(\hat{\alpha} + \sum_{i\in\mathcal{I}} \alpha_i \left(1 - \phi_{i,t,c}\right)\right), \quad \text{with } \hat{\alpha} + \sum_{i\in\mathcal{I}} \alpha_i = 1, \tag{12}$$

where the parameter $\hat{\alpha}$ represents the proportion of transmission that still happens when all NPIs are active.

Assumption 7 states that $R_{t,c}$ depends only on each country's initial reproduction number $R_{0,c}$ and the active NPIs. In other words, no unobserved factors are changing $R_{t,c}$, such as spontaneous social distancing. This is a crucial assumption since the effect of unobserved factors may otherwise be attributed to the active NPIs. This can happen under specific conditions. Firstly, the unobserved effect cannot be present throughout the entire study period since otherwise $R_{0,c}$ accounts for it. Secondly, its timing must be correlated with that of an NPI since otherwise it will be modeled as noise. Under these conditions, an unobserved effect constitutes an *unobserved confounder* [10, 11].

Without unobserved confounders, our models can infer the *causal* effects of the studied NPIs. This is a property of regression models, such as ours, when their specification is correct [12]. To understand this point intuitively, it is worth examining the simplified models used in section 5.

The effect of unobserved confounders is usually examined by introducing artificial confounders and observing how much this affects results [10, 11]. We tested each model's sensitivity to unobserved confounders by making each NPI unobserved, in turn (Figures 6 and 10– 13). Results were stable except when hiding the most effective NPI, school closures. Results may therefore be affected by confounding if unobserved confounders caused a large reduction in $R_{0,c}$ (ca. 50%, which is the inferred effect for school closures).

Brauner et al. [1] performed a further sensitivity check where they added mobility data as an 'NPI' that serves as a proxy for unobserved confounders. We did not include this for the calculation of our sensitivity metrics, the worst-case categorised sensitivity losses $\mathcal{L}_{\text{med}}$ and $\mathcal{L}_{\sigma}$, because mobility is not just a confounder but a *mediator* for the effect of some NPIs, so the inferred effects of these NPIs *should* change. (we do, however, show the test with mobility data in Figures 6 and 10– 13). Brauner et al. [1] found that indeed the inferred effects are significantly decreased only for those NPIs whose effect is mediated through retail and recreation mobility (business closures and stay-home orders, as well as a slight change for gathering bans). However, for the other NPIs, adding mobility data appears to be a sensible check for sensitivity to unobserved confounding.

Assumption 9 states that the country-specific infection fatality rate (IFR_c) and ascertainment rate (AR_c) only change in small steps. **Assumption 13**, implicitly made in [2], states that they do not change at all. Small changes over time are modeled by the noise terms $\varepsilon^{(C)}_{t',c}$, $\varepsilon^{(D)}_{t',c} \sim \mathcal{N}(0, \sigma_g^2)$. Noise at time t changes all future infections, mimicking the effect of a change in the IFR or AR. In the sensitivity analyses, models that lack these noise terms perform significantly worse than others (Figure 1).

Note that, in principle, it is possible to distinguish changes in the IFR and AR from the NPIs' effects: decreasing the ascertainment rate decreases future cases $C_{t,c}$ by a constant factor whereas the introduction of an NPI decreases them by a factor that grows exponentially over time.

A.4 Proofs of Theorems 1 and 2

Proof of Theorem 1. For this model, assume that ground truth values of $R_{t,c}$ have been given to us. By definition, we can write:

$$\log R_{t,c} = \log R_{0,c} - \sum_{i \in \mathcal{I}} \alpha_i \phi_{i,t,c} + \varepsilon_{t,c} \tag{13}$$

where $\varepsilon_{t,c} \sim \mathcal{N}(\mu = 0, \sigma^2 = \sigma_R^2)$; σ_R and $R_{0,c}$ are fixed parameters, $\phi_{i,t,c} \in \{0, 1\}$ and $R_{t,c}$ are given. We want to find the maximum likelihood solution for $\{\alpha_i\}_{i \in \mathcal{I}}$.

The log-likelihood $\mathcal{L}$ is given as

$$\mathcal{L} = \sum_{t,c} \log \mathcal{N}(\varepsilon_{t,c} | 0, \sigma_R^2) = -\frac{1}{2\sigma_R^2} \sum_{t,c} \varepsilon_{t,c}^2 + \text{constant}, \tag{14}$$

where the constant does not depend on the values of $\{\alpha_i\}_{i \in \mathcal{I}}$. Assume that values $\{\alpha_j\}_{j \in \mathcal{I}, j \neq i}$ are fixed and we are finding the ML solution for α_i. Then,

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} \propto \sum_{t,c} \frac{\partial \varepsilon_{t,c}^2}{\partial \alpha_i} \propto \sum_{t,c} \varepsilon_{t,c} \phi_{i,t,c} = \sum_{(t,c) \in \Phi_i} \varepsilon_{t,c} = \sum_{(t,c) \in \Phi_i} (\log \frac{R_{t,c}}{\tilde{R}_{(-i),t,c}} + \alpha_i), \tag{15}$$

where, as in the main text, $\Phi_i = \{(t, c) | \phi_{i,t,c} = 1\}$ is the set of days and countries with NPI i active, and $\tilde{R}_{(-i),t,c}$ is the predicted R ignoring the effect of NPI i:

$$\tilde{R}_{(-i),t,c} = R_{0,c} \prod_{j \in \mathcal{I} \setminus \{i\}} \exp(-\alpha_j \phi_{j,t,c}) \tag{16}$$

Setting $\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0$, we obtain:

$$-\alpha_i |\Phi_i| = \sum_{(t,c) \in \Phi_i} \log \frac{R_{t,c}}{\tilde{R}_{(-i),t,c}}. \tag{17}$$

By exponentiation and separation into two products, we obtain the theorem statement.

All that remains to show is that $\frac{\partial^2 \mathcal{L}}{\partial \alpha_i^2} < 0$. Preserving signs, but not constants of proportionality, we have:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} \propto -\sum_{(t,c) \in \Phi_i} \varepsilon_{t,c} \Rightarrow \frac{\partial^2 \mathcal{L}}{\partial \alpha_i^2} \propto -\sum_{(t,c) \in \Phi_i} (1) < 0, \tag{18}$$

as required. □

Note. It can be shown that $\mathcal{L}$ is a convex function of $\boldsymbol{\alpha}$. Using Eq. 13

$$\mathcal{L} = \sum_{t,c} \log \mathcal{N}(\varepsilon_{t,c} | 0, \sigma_R^2) = -\frac{1}{2\sigma_R^2} \sum_{t,c} \varepsilon_{t,c}^2 + \text{constant}. \tag{19}$$

Proof of Theorem 2. For this model, assume that ground truth values of $g_{t,c}$ have been given to us. Expanding the definitions, we obtain

$$\log g_{t,c} = \beta (R_{0,c}^{1/\nu} \prod_{i \in \mathcal{I}} (\exp(-\alpha_i \phi_{i,t,c})^{1/\nu}) - 1) + \varepsilon_{t,c} \tag{20}$$

where $\varepsilon_{t,c} \sim \mathcal{N}(\mu = 0, \sigma^2 = \sigma_R^2)$; σ_R, ν, β and $R_{0,c}$ are fixed parameters, $\phi_{i,t,c} \in \{0, 1\}$ and $g_{t,c}$ are given.

For each $i \in \mathcal{I}$ independently, we find the maximum likelihood solution α_i given the other $\{\alpha_j\}_{j\in\mathcal{I},j\neq i}$ in the point where $\partial\mathcal{L}/\partial\alpha_i = 0$. The log-likelihood takes the same form as in Eq. (14). By differentiating, we obtain:

$$\frac{\partial\mathcal{L}}{\partial\alpha_i} \propto -\sum_{t,c} \varepsilon_{t,c}\frac{\partial\varepsilon_{t,c}}{\partial\alpha_i} \tag{21}$$

where we have dropped constants of proportionality but kept the correct signs. Recalling Eq. 20, we can write:

$$\frac{\partial\varepsilon_{t,c}}{\partial\alpha_i} = \frac{\beta}{\nu}\tilde{R}_{t,c}^{1/\nu}\phi_{i,t,c} \propto \tilde{R}_{t,c}^{1/\nu}\phi_{i,t,c}. \tag{22}$$

$\tilde{R}_{t,c}$ is the predicted value of $R_{t,c}$ given NPI effectiveness estimates $\{\alpha_i\}_{i\in\mathcal{I}}$ (following Eq. 1 in the main text).

Setting $\frac{\partial\mathcal{L}}{\partial\alpha_i} = 0$ now yields:

$$-\sum_{t,c}\varepsilon_{t,c}\phi_{i,t,c}\tilde{R}_{t,c}^{1/\nu} = 0 \Rightarrow \exp(-\alpha_i/\nu)\sum_{(t,c)\in\Phi_i}\varepsilon_{t,c}\tilde{R}_{(-i),t,c}^{1/\nu} = 0 \tag{23}$$

Then by expanding $\varepsilon_{t,c}$ using Eq. 20 and expressing $\log g_{t,c}$ in terms of $R_{t,c}$ i.e., converting using Assumption 8, we obtain:

$$\sum_{(t,c)\in\Phi_i}\tilde{R}_{(-i),t,c}^{1/\nu}\left(\beta(\bar{R}_{t,c}^{1/\nu} - 1) - \beta(\tilde{R}_{(-i),t,c}^{1/\nu}\exp(-\alpha_i)^{1/\nu} - 1)\right) = 0. \tag{24}$$

$\bar{R}_{t,c}$ is the value of $R_{t,c}$ produced by converting ground truth values of $g_{t,c}$ using Assumption 8.

From this we obtain the theorem by simplification and rearranging.

All that remains is to show that $\frac{\partial^2\mathcal{L}}{\partial\alpha_i^2} < 0$. Keeping the signs but dropping constants of proportionality, we have:

$$\frac{\partial\mathcal{L}}{\partial\alpha_i} \propto -\sum_{(t,c)\in\Phi_i}\varepsilon_{t,c}\tilde{R}_{t,c}^{1/\nu}. \tag{25}$$

Therefore:

$$\begin{aligned}\frac{\partial^2\mathcal{L}}{\partial\alpha_i^2} &\propto -\sum_{(t,c)\in\Phi_i}\left[\frac{\partial\varepsilon_{t,c}}{\partial\alpha_i}\tilde{R}_{t,c}^{1/\nu} + \varepsilon_{t,c}\frac{\partial\tilde{R}_{t,c}^{1/\nu}}{\partial\alpha_i}\right] \\ &\propto -\frac{\beta}{\nu}\sum_{(t,c)\in\Phi_i}\left(\tilde{R}_{t,c}^{1/\nu}\right)^2 + \frac{1}{\nu}\underbrace{\sum_{(t,c)\in\Phi_i}\varepsilon_{t,c}\tilde{R}_{t,c}^{1/\nu}}_{\text{0 at ML solution}}\end{aligned} \tag{26}$$

Combining Eqs. 21 and 22, we see that the second term is proportional to $\frac{\partial\mathcal{L}}{\partial\alpha_i}$ and therefore 0 at the maximum likelihood solution. Given this, we have $\frac{\partial^2\mathcal{L}}{\partial\alpha_i^2} < 0$ at α_i satisfying Eq. (24). Therefore, the solution of Eq. (24) is the maximum likelihood solution.

□

A.5 Experiment details

A.5.1 Data Preprocessing

We perform the same data preprocessing as in [1]:

- Our data for confirmed cases and deaths is given by the John Hopkins Centre for Systems Science and Engineering[13, 14]. This data is noisy, so 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 ignore new cases before a country has reached 100 cumulative confirmed cases, accounting 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 ignore new deaths before a country has reached 10 cumulative deaths.

A.5.2 Cross Validation

We perform four fold cross validation, holding all but the first fourteen days of cases and deaths out for 6 regions at a time. We only considering predictions for regions which have greater than 100 deaths at the end of the time period we consider, since we are primarily interested in validating holdout performance over long time periods. We do not hold out the first 14 days of cases and deaths to allow the model to infer $R_{0,c}, N_{0,c}^{(C)}$ and $N_{0,c}^{(D)}$. The only way the model is able to explain held-out data is through these parameters as well as the shared NPI effectiveness parameters, $\{\alpha_i\}$. We chose a fixed set of four folds of cross validation randomly, which are:

$$\texttt{Fold 1} = \texttt{[DE, HU, FI, IE, RS, BE]},$$
$$\texttt{Fold 2} = \texttt{[DK, GR, NO, FR, RO, NA]},$$
$$\texttt{Fold 3} = \texttt{[ES, CZ, NL, CH, PT, AT]},$$
$$\texttt{Fold 4} = \texttt{[IL, SE, IT, MX, GB, PL]}.$$

For these experiments, we run 4 chains with 2000 samples per chain.

A.5.3 Convergence Statistics

For all experiments, we ensure that $\hat{R} < 1.05$ (i.e., there are no PyMC3 warnings) and that there are no divergent transitions. For the baseline model, we have $\hat{R} \in [1.000, 1.004]$ for the vast majority of parameters for the experiment with no held out data and default parameter settings.

A.5.4 Sensitivity Analyis

We summarise the sensitivity analysis tests we perform here. These are mostly as performed in [1]. Default values are highlighted in bold.

Epidemiological Parameter Uncertainty.

1. We shift the mean of the infection-to-confirmation distribution by $[-2, -1, \mathbf{0}, 1, +2]$ days.
2. We shift the mean of the infection-to-death distribution by $[-2, -1, \mathbf{0}, 1, +2]$ days.
3. We consider different serial intervals with mean values $[4, 5, 6, \mathbf{6.67}, 7, 8]$ days. These distributions have the same standard deviation as the default distribution (4.88 days).
4. We change the mean value of the hyperprior of $R_{0,c}$, $\bar{R}$. We consider values $[\log 1.25,\ \mathbf{\log 3.25},\ \log 5.5]$.
5. We change the prior over α_i. For all models except the additive model, we try $\boldsymbol{\mathcal{N}(0, 0.2^2)}$, $\mathcal{N}(0, 10^2)$, $\mathcal{N}^+(0, 0.2^2)$. For the additive model, we use a Dirichlet(α) prior, where the concentration parameter α is the same for all components. We consider values $[\mathbf{1}, 5, 10]$.

Data Sensitivity.

1. We hold out regions [CZ, DE, MX, NL, PL, PT] (chosen randomly) one at a time.

2. We change the cutoff below which confirmed COVID-19 cases are included in $[10, 30, \mathbf{100}, 300, 500]$.
3. We include Google mobility data as an additional NPI.
4. We modify the *school closure* feature in Sweden to be off rather than on. While [15] consider this feature to be not active for Sweden, [1] consider it to be active.
5. We remove each NPI in turn from the data.

For sensitivity experiments, we run 8 chains with 1000 samples per chain.

A.6 Complete Model Descriptions

Note: we use the same notation as in the main text. Unless otherwise stated, all prior distributions are independent.

A.6.1 Baseline Model [1]

- **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,
  $$R_{0,c} = \exp(\bar{R} + \sigma_R z_c) \tag{27}$$
  $$\bar{R} \sim \texttt{Student T}(\mu = \log(3.25), \sigma = 0.2, \nu = 10) \tag{28}$$
  $$\sigma_R \sim \texttt{Half Student T}(\sigma = 0.2, \nu = 10) \tag{29}$$
  $$z_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 1) \tag{30}$$
  2. **NPI Effectiveness**:
  $$\alpha_i \sim \texttt{Normal}(\mu = 0, \sigma_R^2 = 0.2) \tag{31}$$
  $$\tag{32}$$
  3. **Infection Initial Counts**.
  $$N_{0,c}^{(C)} = \exp(\zeta_c^{(C)}) \tag{33}$$
  $$N_{0,c}^{(D)} = \exp(\zeta_c^{(D)}) \tag{34}$$
  $$\zeta_c^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{35}$$
  $$\zeta_c^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{36}$$
  $$\tag{37}$$
  4. **Observation Noise Dispersion Parameter**
  $$\Psi \sim \texttt{Half Normal}(\mu = 0, \sigma^2 = 5^2) \tag{38}$$
- **Hyperparameters**
  1. **Infection Noise Scale,** $\sigma_N = 0.2$ (selected by cross-validation).
  2. **Serial Interval Parameters**. The serial interval is assumed to have a `Gamma` distribution with $\alpha = 1.87$ and $\beta = 0.28$. [5]
  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:[6, 4, 5, 16]
  $$\mathcal{T}^{(C)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Negative Binomial}(\mu = 5.25, \alpha = 1.57) \tag{39}$$
  The time from infection to death is assumed to be the sum of the incubation period and the time taken from symptom onset to death. Therefore, the time taken from infection

to death, $\mathcal{T}^{(D)}$ is:[6, 8]

$$\mathcal{T}^{(D)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Gamma}(\mu = 18.8, \frac{\sigma}{\mu} = 0.45), \tag{40}$$

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. $g_{t,c} = \exp\left(\beta(R_{c,t}^{\frac{1}{\alpha}} - 1)\right)$ 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.[17] (Note that we use daily growth rates.)
  3. $$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{\tau=1}^{t} \left[g_{\tau,c} \cdot \exp \varepsilon_{\tau,c}^{(C)}\right], \tag{41}$$

     $$N_{t,c}^{(D)} = N_{0,c}^{(D)} \prod_{\tau=1}^{t} \left[g_{\tau,c} \cdot \exp \varepsilon_{\tau,c}^{(D)})\right], \text{with noise} \tag{42}$$

     $$\varepsilon_{\tau,c}^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2), \tag{43}$$

     $$\varepsilon_{\tau,c}^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2) \tag{44}$$

     $N_{t,c}^{(C)}$ 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.

$$\bar{C}_{t,c} = \sum_{\tau=1}^{32} N_{t-\tau,c}^{(C)} \pi_C[\tau], \tag{45}$$

$$\bar{D}_{t,c} = \sum_{\tau=1}^{64} N_{t-\tau,c}^{(D)} \pi_D[\tau]. \tag{46}$$

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

$$C_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{C}_{t,c}, \alpha = \Psi) \tag{47}$$

$$D_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{D}_{t,c}, \alpha = \Psi) \tag{48}$$

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.

A.6.2 Additive Effect Model

- **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,

$$R_{0,c} = \exp(\bar{R} + \sigma_R z_c) \tag{49}$$

$$\bar{R} \sim \texttt{Student T}(\mu = \log(3.25), \sigma = 0.2, \nu = 10) \tag{50}$$

$$\sigma_R \sim \texttt{Half Student T}(\sigma = 0.2, \nu = 10) \tag{51}$$

$$z_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 1) \tag{52}$$

2. **NPI Effectiveness**:

$$\hat{\alpha}, \{\alpha_i\} \sim \texttt{Dirichlet}(\alpha = 1) \tag{53}$$

α is the *concentration parameter* of the Dirichlet distribution, and assumed to be the same for all dimensions.

3. **Infection Initial Counts**.

$$N_{0,c}^{(C)} = \exp(\zeta_c^{(C)}) \tag{54}$$

$$N_{0,c}^{(D)} = \exp(\zeta_c^{(D)}) \tag{55}$$

$$\zeta_c^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{56}$$

$$\zeta_c^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{57}$$

$$\tag{58}$$

4. **Observation Noise Dispersion Parameter**

$$\Psi \sim \texttt{Half Normal}(\mu = 0, \sigma^2 = 5^2) \tag{59}$$

- **Hyperparameters**

1. **Infection Noise Scale,** $\sigma_N = 0.2$ (selected by cross-validation).
2. **Serial Interval Parameters**. The serial interval is assumed to have a `Gamma` distribution with $\alpha = 1.87$ and $\beta = 0.28$. [5]
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:[6, 4, 5, 16]

$$\mathcal{T}^{(C)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Negative Binomial}(\mu = 5.25, \alpha = 1.57) \tag{60}$$

The time from infection to death is assumed to be the sum of the incubation period and the time taken from symptom onset to death. Therefore, the time taken from infection to death, $\mathcal{T}^{(D)}$ is:[6, 8]

$$\mathcal{T}^{(D)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Gamma}(\mu = 18.8, \frac{\sigma}{\mu} = 0.45), \tag{61}$$

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. $R_{t,c} = R_{0,c} \cdot \left[\hat{\alpha} + \sum_{i \in \mathcal{I}} \alpha_i (1 - \phi_{i,t,c})\right]$.
2. $g_{t,c} = \exp\left(\beta(R_{c,t}^{\frac{1}{\alpha}} - 1)\right)$ 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.[17] (Note that we use daily growth rates.)

3.

$$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{\tau=1}^{t} \left[g_{\tau,c} \cdot \exp \varepsilon_{\tau,c}^{(C)} \right], \tag{62}$$

$$N_{t,c}^{(D)} = N_{0,c}^{(D)} \prod_{\tau=1}^{t} \left[g_{\tau,c} \cdot \exp \varepsilon_{\tau,c}^{(D)}) \right], \text{with noise} \tag{63}$$

$$\varepsilon_{\tau,c}^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2), \tag{64}$$

$$\varepsilon_{\tau,c}^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2) \tag{65}$$

$N_{t,c}^{(C)}$ 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.

$$\bar{C}_{t,c} = \sum_{\tau=1}^{32} N_{t-\tau,c}^{(C)} \pi_C[\tau], \tag{66}$$

$$\bar{D}_{t,c} = \sum_{\tau=1}^{64} N_{t-\tau,c}^{(D)} \pi_D[\tau]. \tag{67}$$

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

$$C_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{C}_{t,c}, \alpha = \Psi) \tag{68}$$

$$D_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{D}_{t,c}, \alpha = \Psi) \tag{69}$$

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.

A.6.3 Noisy-R Model

- **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,

$$R_{0,c} = \exp(\bar{R} + \sigma_R z_c) \tag{70}$$

$$\bar{R} \sim \texttt{Student T}(\mu = \log(3.25), \sigma = 0.2, \nu = 10) \tag{71}$$

$$\sigma_R \sim \texttt{Half Student T}(\sigma = 0.2, \nu = 10) \tag{72}$$

$$z_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 1) \tag{73}$$

  2. **NPI Effectiveness**:

$$\alpha_i \sim \texttt{Normal}(\mu = 0, \sigma_R^2 = 0.2) \tag{74}$$

$$\tag{75}$$

  3. **Infection Initial Counts**.

$$N_{0,c}^{(C)} = \exp(\zeta_c^{(C)}) \tag{76}$$

$$N_{0,c}^{(D)} = \exp(\zeta_c^{(D)}) \tag{77}$$

$$\zeta_c^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{78}$$

$$\zeta_c^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{79}$$

$$\tag{80}$$

4. **Observation Noise Dispersion Parameter**

$$\Psi \sim \texttt{Half Normal}(\mu = 0, \sigma^2 = 5^2) \tag{81}$$

- **Hyperparameters**

1. **Infection Noise Scale,** $\sigma_R = 0.7$ (selected by cross-validation).
2. **Serial Interval Parameters**. The serial interval is assumed to have a `Gamma` distribution with $\alpha = 1.87$ and $\beta = 0.28$. [5]
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:[6, 4, 5, 16]

$$\mathcal{T}^{(C)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Negative Binomial}(\mu = 5.25, \alpha = 1.57) \tag{82}$$

The time from infection to death is assumed to be the sum of the incubation period and the time taken from symptom onset to death. Therefore, the time taken from infection to death, $\mathcal{T}^{(D)}$ is:[6, 8]

$$\mathcal{T}^{(D)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Gamma}(\mu = 18.8, \frac{\sigma}{\mu} = 0.45), \tag{83}$$

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.

$$R^{(C)}_{t,c} = R_{0,c} \cdot \exp\left(-\sum_{i=1}^{9} \alpha_i\ \phi_{i,t,c}\right) \cdot \exp \varepsilon^{(C)}_{\tau,c} \tag{84}$$

$$R^{(D)}_{t,c} = R_{0,c} \cdot \exp\left(-\sum_{i=1}^{9} \alpha_i\ \phi_{i,t,c}\right) \cdot \exp \varepsilon^{(D)}_{\tau,c}, \text{with noise} \tag{85}$$

$$\varepsilon^{(C)}_{\tau,c} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2), \tag{86}$$

$$\varepsilon^{(D)}_{\tau,c} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2) \tag{87}$$

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.[17] (Note that we use daily growth rates.)
3.

$$N^{(C)}_{t,c} = N^{(C)}_{0,c} \prod_{\tau=1}^{t} g_{\tau,c}, \tag{88}$$

$$N^{(D)}_{t,c} = N^{(D)}_{0,c} \prod_{\tau=1}^{t} g_{\tau,c} \tag{89}$$

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

$$\bar{C}_{t,c} = \sum_{\tau=1}^{32} N^{(C)}_{t-\tau,c} \pi_C[\tau], \tag{90}$$

$$\bar{D}_{t,c} = \sum_{\tau=1}^{64} N^{(D)}_{t-\tau,c} \pi_D[\tau]. \tag{91}$$

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

$$C_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{C}_{t,c}, \alpha = \Psi) \tag{92}$$

$$D_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{D}_{t,c}, \alpha = \Psi) \tag{93}$$

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.

A.6.4 Different Effects Model

- **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,

$$R_{0,c} = \exp(\bar{R} + \sigma_R z_c) \tag{94}$$

$$\bar{R} \sim \texttt{Student T}(\mu = \log(3.25), \sigma = 0.2, \nu = 10) \tag{95}$$

$$\sigma_R \sim \texttt{Half Student T}(\sigma = 0.2, \nu = 10) \tag{96}$$

$$z_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 1) \tag{97}$$

  2. **NPI Effectiveness**:

$$\hat{\alpha}_i \sim \texttt{Normal}(\mu = 0, \sigma_R^2 = 0.2^2) \tag{98}$$

$$\alpha_{i,c} \sim \texttt{Normal}(\mu = \hat{\alpha}_i, \sigma_\alpha^2 = 0.1^2) \tag{99}$$

$$\tag{100}$$

  3. **Infection Initial Counts**.

$$N^{(C)}_{0,c} = \exp(\zeta^{(C)}_c) \tag{101}$$

$$N^{(D)}_{0,c} = \exp(\zeta^{(D)}_c) \tag{102}$$

$$\zeta^{(C)}_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{103}$$

$$\zeta^{(D)}_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 50^2) \tag{104}$$

$$\tag{105}$$

  4. **Observation Noise Dispersion Parameter**

$$\Psi \sim \texttt{Half Normal}(\mu = 0, \sigma^2 = 5^2) \tag{106}$$

- **Hyperparameters**
  1. **Infection Noise Scale,** $\sigma_N = 0.2$ (selected by cross-validation).
  2. **Country-Variation Scale,** $\sigma_\alpha = 0.1$ (selected by cross-validation).
  3. **Serial Interval Parameters**. The serial interval is assumed to have a `Gamma` distribution with $\alpha = 1.87$ and $\beta = 0.28$. [5]
  4. **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:[6, 4,

5, 16]

$$\mathcal{T}^{(C)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Negative Binomial}(\mu = 5.25, \alpha = 1.57) \tag{107}$$

The time from infection to death is assumed to be the sum of the incubation period and the time taken from symptom onset to death. Therefore, the time taken from infection to death, $\mathcal{T}^{(D)}$ is:[6, 8]

$$\mathcal{T}^{(D)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Gamma}(\mu = 18.8, \frac{\sigma}{\mu} = 0.45), \tag{108}$$

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. $R_{t,c} = R_{0,c} \cdot \exp\left(-\sum_{i=1}^{9} \alpha_{i,c}\, \phi_{i,t,c}\right)$.
  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.[17] (Note that we use daily growth rates.)
  3. $$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(C)}\right], \tag{109}$$

     $$N_{t,c}^{(D)} = N_{0,c}^{(D)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(D)})\right], \text{with noise} \tag{110}$$

     $$\varepsilon_{\tau,c}^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2), \tag{111}$$

     $$\varepsilon_{\tau,c}^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2) \tag{112}$$

     $N_{t,c}^{(C)}$ 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.

$$\bar{C}_{t,c} = \sum_{\tau=1}^{32} N_{t-\tau,c}^{(C)} \pi_C[\tau], \tag{113}$$

$$\bar{D}_{t,c} = \sum_{\tau=1}^{64} N_{t-\tau,c}^{(D)} \pi_D[\tau]. \tag{114}$$

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

$$C_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{C}_{t,c}, \alpha = \Psi) \tag{115}$$

$$D_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{D}_{t,c}, \alpha = \Psi) \tag{116}$$

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.

A.6.5 Discrete Renewal Model [2]

Note: this model uses only deaths as observations. In [2], they do not smooth the death observations. Our implemention of this model uses different priors as compared to Flaxman et al. [2].

- **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,
  $$R_{0,c} = \exp(\bar{R} + \sigma_R z_c) \tag{117}$$
  $$\bar{R} \sim \texttt{Student T}(\mu = \log(3.25), \sigma = 0.2, \nu = 10) \tag{118}$$
  $$\sigma_R \sim \texttt{Half Student T}(\sigma = 0.2, \nu = 10) \tag{119}$$
  $$z_c \sim \texttt{Normal}(\mu = 0, \sigma^2 = 1) \tag{120}$$
  2. **NPI Effectiveness**:
  $$\alpha_i \sim \texttt{Normal}(\mu = 0, \sigma_R^2 = 0.2) \tag{121}$$
  $$\tag{122}$$
  3. **Infection Initial Counts**.
  $$N_{0,c}^{(D)} = \exp(\zeta_c^{(D)}) \tag{123}$$
  $$\zeta_c^{(D)} \sim \texttt{ Normal}(\mu = 0, \sigma^2 = 50^2) \tag{124}$$
  $$\tag{125}$$
  4. **Observation Noise Dispersion Parameter**
  $$\Psi \sim \texttt{Half Normal}(\mu = 0, \sigma^2 = 5^2) \tag{126}$$
- **Hyperparameters**
  1. **Serial Interval Parameters**. The serial interval is assumed to have a `Gamma` distribution with $\alpha = 1.87$ and $\beta = 0.28$. [5] We discretise using Monte Carlo sampling to form discrete array $\pi_{\text{SI}}[i]$ with a maximum delay of 27 days.
  2. **Delay Distributions**. The time from infection to confirmation is assumed to be the sum of the incubation period and the time taken from symptom onset to laboratory confirmation. Therefore, the time taken from infection to confirmation, $\mathcal{T}^{(C)}$ is:[6, 4, 5, 16]
  $$\mathcal{T}^{(C)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Negative Binomial}(\mu = 5.25, \alpha = 1.57) \tag{127}$$
  The time from infection to death is assumed to be the sum of the incubation period and the time taken from symptom onset to death. Therefore, the time taken from infection to death, $\mathcal{T}^{(D)}$ is:[6, 8]
  $$\mathcal{T}^{(D)} \sim \texttt{Gamma}(\mu = 5.1, \frac{\sigma}{\mu} = 0.86) + \texttt{Gamma}(\mu = 18.8, \frac{\sigma}{\mu} = 0.45), \tag{128}$$
  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.[17] (Note that we use daily growth rates.)
3.

$$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(C)}\right], \tag{129}$$

$$N_{t,c}^{(D)} = N_{0,c}^{(D)} \prod_{\tau=1}^{t} \left[(g_{\tau,c} + 1) \cdot \exp \varepsilon_{\tau,c}^{(D)})\right], \text{with noise} \tag{130}$$

$$\varepsilon_{\tau,c}^{(C)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2), \tag{131}$$

$$\varepsilon_{\tau,c}^{(D)} \sim \texttt{Normal}(\mu = 0, \sigma^2 = \sigma_N^2) \tag{132}$$

$N_{t,c}^{(C)}$ 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.

$$\bar{C}_{t,c} = \sum_{\tau=1}^{32} N_{t-\tau,c}^{(C)} \pi_C[\tau], \tag{133}$$

$$\bar{D}_{t,c} = \sum_{\tau=1}^{64} N_{t-\tau,c}^{(D)} \pi_D[\tau]. \tag{134}$$

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

$$C_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{C}_{t,c}, \alpha = \Psi) \tag{135}$$

$$D_{t,c} \sim \texttt{Negative Binomial}(\mu = \bar{D}_{t,c}, \alpha = \Psi) \tag{136}$$

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.

A.7 Bibliography

References

[1] Jan Markus Brauner, Mrinank Sharma, Sören Mindermann, Anna B Stephenson, Tomáš Gavenčiak, David Johnston, John Salvatier, Gavin Leech, Tamay Besiroglu, George Altman, Hong Ge, Vladimir Mikulik, Meghan Hartwick, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal, and Jan Kulveit. The effectiveness and perceived burden of nonpharmaceutical interventions against COVID-19 transmission: a modelling study with 41 countries. *medRxiv*, 2020.

[2] S Flaxman, S Mishra, A Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman, N Schmit, L Cilloni, K Ainslie, M Baguelin, I Blake, A Boonyasiri, O Boyd, L Cattarino, C Ciavarella, L Cooper, Z Cucunuba Perez, G Cuomo-Dannenburg, A Dighe, A Djaafara, I Dorigatti, S Van Elsland, R Fitzjohn, H Fu, K Gaythorpe, L Geidelberg, N Grassly, W Green, T Hallett, A Hamlet, W Hinsley, B Jeffrey, D Jorgensen, E Knock, D Laydon, G Nedjati Gilani, P Nouvellet, K Parag, I Siveroni, H Thompson, R Verity, E Volz, C Walters, H Wang, Y Wang, O Watson, P Winskill, X Xi, C Whittaker, P Walker, A Ghani, C Donnelly, S Riley, L Okell, M Vollmer, N Ferguson, and S Bhatt. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Technical report, Imperial College London, 2020.

[3] Juanjuan Zhang, Maria Litvinova, Wei Wang, Yan Wang, Xiaowei Deng, Xinghui Chen, Mei Li, Wen Zheng, Lan Yi, Xinhua Chen, Qianhui Wu, Yuxia Liang, Xiling Wang, Juan Yang, Kaiyuan Sun, Ira M Longini, M Elizabeth Halloran, Peng Wu, Benjamin J Cowling, Stefano Merler,

Cecile Viboud, Alessandro Vespignani, Marco Ajelli, and Hongjie Yu. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *The Lancet Infectious Diseases*, apr 2020.

[4] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy S.M. Leung, Eric H.Y. Lau, Jessica Y. Wong, Xuesen Xing, Nijuan Xiang, Yang Wu, Chao Li, Qi Chen, Dan Li, Tian Liu, Jing Zhao, Man Liu, Wenxiao Tu, Chuding Chen, Lianmei Jin, Rui Yang, Qi Wang, Suhua Zhou, Rui Wang, Hui Liu, Yinbo Luo, Yuan Liu, Ge Shao, Huan Li, Zhongfa Tao, Yang Yang, Zhiqiang Deng, Boxi Liu, Zhitao Ma, Yanping Zhang, Guoqing Shi, Tommy T.Y. Lam, Joseph T. Wu, George F. Gao, Benjamin J. Cowling, Bo Yang, Gabriel M. Leung, and Zijian Feng. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13):1199–1207, mar 2020.

[5] D Cereda, M Tirani, F Rovida, V Demicheli, M Ajelli, P Poletti, F Trentini, G Guzzetta, V Marziano, A Barone, M Magoni, S Deandrea, G Diurno, M Lombardo, M Faccini, A Pan, R Bruno, E Pariani, G Grasselli, A Piatti, M Gramegna, F Baldanti, A Melegaro, and S Merler. The early phase of the COVID-19 outbreak in lombardy, italy. *arXiv*, 2020.

[6] Natalie M. Linton, Tetsuro Kobayashi, Yichi Yang, Katsuma Hayashi, Andrei R. Akhmetzhanov, Sung mok Jung, Baoyin Yuan, Ryo Kinoshita, and Hiroshi Nishiura. 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*, jan 2020.

[7] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Helen Coupland, Thomas A Mellan, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, Pablo N P Guzman, Nora Schmit, Lucia Callizo, Imperial College COVID-19 Response Team, Charles Whittaker, Peter Winskill, Xiaoyue Xi, Azra Ghani, Christl A. Donnelly, Steven Riley, Lucy C Okell, Michaela A C Vollmer, Neil M. Ferguson, and Samir Bhatt. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European countries: technical description update. `https://arxiv.org/abs/2004.11342`, 2020.

[8] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick G T Walker, Han Fu, Amy Dighe, Jamie T Griffin, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Anne Cori, Zulma Cucunubá, Rich FitzJohn, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Daniel Laydon, Gemma Nedjati-Gilani, Steven Riley, Sabine van Elsland, Erik Volz, Haowei Wang, Yuanrong Wang, Xiaoyue Xi, Christl A Donnelly, Azra C Ghani, and Neil M Ferguson. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, mar 2020.

[9] James Tozer and Martín González. The Economist COVID-19 Excess Deaths Tracker. `https://github.com/TheEconomist/covid-19-excess-deaths-tracker`, 2020.

[10] Mohammad Ali Mansournia, Mahyar Etminan, Goodarz Danaei, Jay S Kaufman, and Gary Collins. Handling time varying confounding in observational research. *BMJ*, 359:j4587, 2017.

[11] P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

[12] A Gelman and J Hill. Causal inference using regression on the treatment variable. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2007.

[13] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, may 2020.

[14] Johns Hopkins University Center for Systems Science and Engineering. COVID-19 data repository by the center for systems science and engineering (CSSE) at johns hopkins university. `https://github.com/CSSEGISandData/COVID-19`, 2020.

[15] Nicolas Banholzer, Eva van Weenen, Bernhard Kratzwald, Arne Seeliger, Daniel Tschernutter, Pierluigi Bottrighi, Alberto Cenedese, Joan Puig Salles, Werner Vach, and Stefan Feuerriegel. Impact of non-pharmaceutical interventions on documented cases of COVID-19. *COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv*, apr 2020.

[16] Qifang Bi, Yongsheng Wu, Shujiang Mei, Chenfei Ye, Xuan Zou, Zhen Zhang, Xiaojian Liu, Lan Wei, Shaun A Truelove, Tong Zhang, Wei Gao, Cong Cheng, Xiujuan Tang, Xiaoliang Wu, Yu Wu, Binbin Sun, Suli Huang, Yu Sun, Juncen Zhang, Ting Ma, Justin Lessler, and Teijian Feng. 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*, mar 2020.

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