Supplementary Information - Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe

Mrinank Sharma*,‡, Sören Mindermann*,‡, Charlie Rogers-Smith, Gavin Leech, Benedict Snodin, Janvi Ahuja, Jonas B. Sandbrink, Joshua Teperowski Monrad, George Altman, Gurpreet Dhaliwal, Lukas Finnveden, Alexander John Norman, Sebastian B. Oehm, Julia Fabienne Sandkühler, Laurence Aitchison, Tomáš Gavenčiak, Thomas Mellan, Jan Kulveit, Leonid Chindelevitch, Seth Flaxman, Yarin Gal, Swapnil Mishra[+], Samir Bhatt[+], Jan Markus Brauner*,[+]

*Equal contribution to first authorship ‡listing order chosen at random
[+]Equal contribution to senior authorship
Correspondence: mrinank.sharma@eng.ox.ac.uk, soren.mindermann@cs.ox.ac.uk, s.bhatt@imperial.ac.uk, jan.brauner@eng.ox.ac.uk

Supplementary Notes

Table of Contents

Supplementary Note 1. Generalisation of NPI effectiveness estimates to later waves

1.1. Stability of voluntary protective behaviours and organisational safety measures over time

1.1.1. Personal protective behaviours

The YouGov survey (*1*) polls data on protective behaviours since March 2020. For the UK, but not for other countries, data is available from 1 March 2020 where NPIs were largely absent. The YouGov survey is chosen because it contains more relevant countries than other surveys. However, other surveys show similar results, such as the COVID-19 Consumer Sentiment Tracker for the UK (*2*) and the HOPE Dashboard (*3*) (for which data is only available graphically).

In Supplementary Figure S1, we display all included countries that are also part of our study: the UK, Italy, and Germany.

We focus on the behaviour "avoiding crowded places" over time. This behaviour appears especially important in a pandemic that is thought to be dominated by superspreading and not by contact transmission (*4,5,6*).

On 1 March, only 14% of people in the UK reported "avoiding crowded places" (early data not available from other countries), but this steeply increases from March to April 2020, reaching a peak of 80% in the UK, 90% in Italy, and 70% in Germany. The difference between countries could be explained by different behaviour but also by differences in interpreting and answering the polling question. By the start of our period of analysis, August 2020, "avoiding crowded places" has stabilized to a month-average of 64% in the UK, 73% in Italy, and 61% in Germany. These figures remained relatively stable up until the most recent 4 weeks, April-May 2021: 58% in the UK, 77% in Italy, and 63% in Germany.

The average month-on-month change is small after August 2020: on average an absolute change of 2% in the UK, 1.1% in Italy, and 1.4% in Germany. As noted, the behaviour decreased to 58% in the UK in the most recent 4 weeks. This may reflect the reopening of the country. However, this percentage is still close to the one seen in August 2020 (64%), and further from the earliest recorded percentage (14%). This suggests that behaviour returns to pre-pandemic levels at a slow pace in the country with the highest rates of vaccination, lowest incidence, and least strict NPIs. We encourage governments to continue to monitor

4
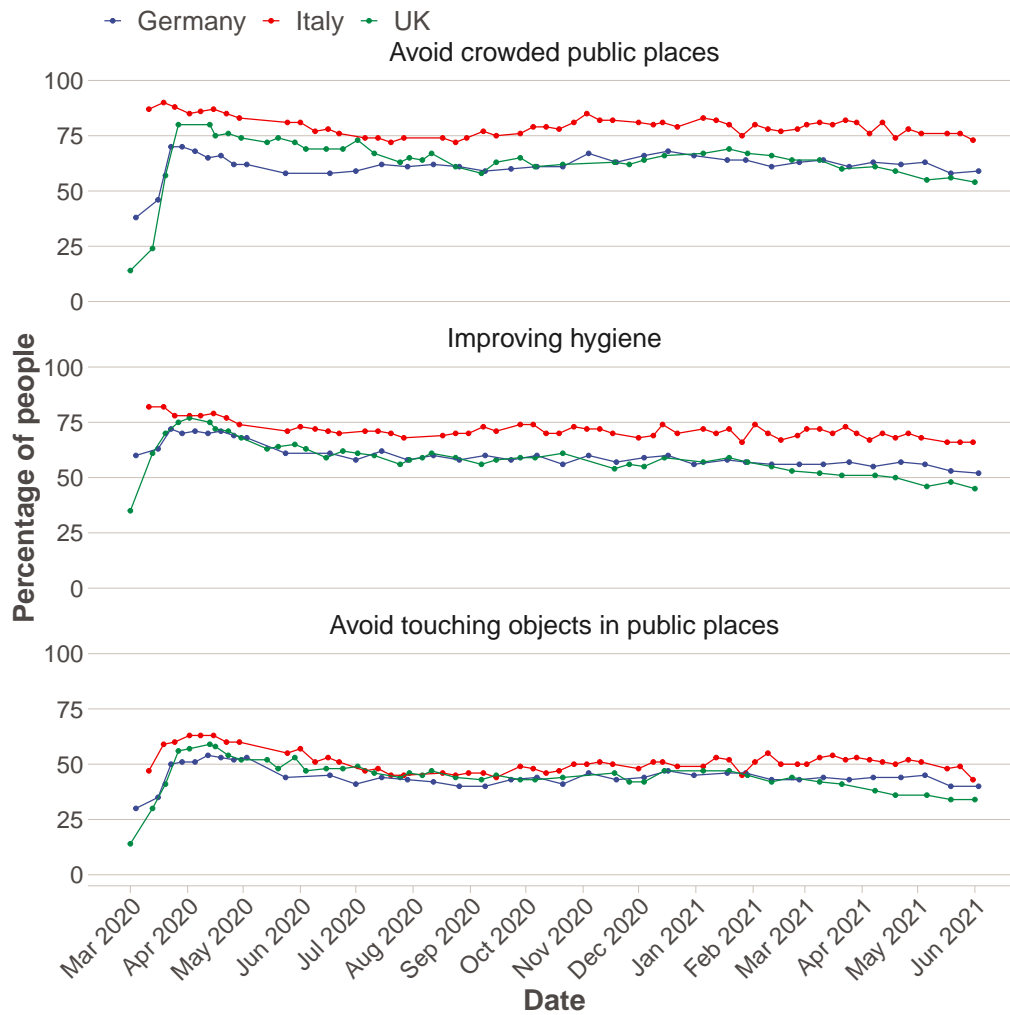
Supplementary Figure S1: Behaviour of people during the second wave as measured by the YouGov survey (*1*) for Germany, Italy and UK. The other countries in our study are not included in this survey. Note that some data is lacking in the earliest weeks.

these trends. Particularly, behaviour change among non-immune individuals could affect transmission.

We find qualitatively similar trends for other voluntary protective behaviours: "improving hygiene" and "avoiding touching objects in public" (Supplementary Figure S1). However, the public health importance of these behaviours is still debated (*6*). We include these to reflect broader behaviour change. "Mask-wearing in public" shows a similar curve but is not shown here since it is largely prescribed by the mask-wearing NPI.

Other surveys display similar trends for infection-reducing voluntary behaviours (*2,3*) in the relevant countries; this also extends to some measures of compliance (*3*). However, no behaviour is expected to stay fully stable, and all may eventually return to pre-pandemic levels.

1.1.2. Organisational safety measures

First wave effectiveness estimates measure the reduction in transmission compared to the earliest phase of the epidemic, where transmission was high due to a lack of protective behaviours and safety measures. However, safety measures largely stayed active since they were first implemented, and in some cases were further strengthened over time. We describe government safety prescriptions from Switzerland as an example, since this was the only country in our study with a readily available safety measure chronology. It is published by the Federal Office of Public Health FOPH, Switzerland (*7,8*).

Official safety prescriptions and guidelines for businesses, schools, and events in Switzerland were introduced in June 2020. Examples of measures are "All persons must be able to wash their hands regularly. Hand sanitiser and, in the case of publicly accessible washbasins, soap must be made available", "All contact surfaces must be regularly cleaned", "Persons must keep a minimum of 1.5 metres distance from each other", "measures that guarantee compliance with the requirement to wear a mask", and "restrict access to the establishment, business or event in such a way that allows the required distance [1.5 metres] to be maintained". These measures have not been rolled back to the pre-pandemic state. All measures have stayed in place until the present time. Furthermore, in October 2020, additional measures, namely "provision must be made for recording the contact details of the persons who are exempted from the requirement to wear a face mask", were added.

Although the above is geographically very limited, it suggests that, despite ongoing re-openings, safety measures are presently not fully relaxed.

1.2. Empirical assessment of generalisation of NPI effects to the third wave

We assess the generalisation of estimates from the first and second wave to the period January-May 2021. In this period, countries both implemented and lifted NPIs several times as they controlled a variant-driven epidemic. We can compare the "observed" changes in R_t (implied by the dynamics of the reported number of COVID-19 cases per day) with the "predicted" changes in R_t (predicted from NPI implementation dates and previously estimated NPI effects). We can then compare predictions using first-wave NPI effects (estimated in previous work (9)) to predictions using second-wave NPI effects from this work.

1.2.1. Data

We collected national NPI data for 6 countries between 9th January and 1st May 2021.

In the third wave, NPIs were implemented on a national, rather than subnational, level in several European countries (similar to the first wave, but unlike the second wave). We thus only include countries where NPIs were implemented nationally, for ease of modelling and data collection. From our study, these are: England, the Netherlands, and the Czech Republic. We exclude Germany, Italy, Switzerland, and Austria due to subnational NPI implementation in 2021. We replaced these with the three major European countries whose NPI implementation was national: Hungary, Belgium, and Ireland. National modelling reduces the number of data points and may introduce some error. However, for a prediction experiment, a smaller dataset suffices.

We recorded NPIs using the NPI definitions both from this study and our previous work analysing the first wave (9). We manually researched all dates by using internet searches and screening government press releases, ordinances, and legislation. All collected data was cross-validated against the OXCGRT dataset (10). Edge cases and judgement calls in the data collection were handled as in the main data collections (see Supplementary Note 5 and Appendix G in Brauner et al. (9)).

We took national case data from the Johns Hopkins Center for Systems Science and Engineering (CSSE) COVID-19 Dataset (11).

1.2.2. Method

Summary of method. We consider using NPI data to predict changes in transmission immediately following NPI implementation changes—short-term predictions. Intuitively, although many factors influence transmission, the implementation of NPIs can be used to predict immediate changes in transmission. Factors such as vaccination coverage and variants of concern will affect R_t, but only gradually. Thus, by focusing on short-term rather than long-term predictions, we can assess the generalisation of NPI effects while minimising assessment bias from time-varying factors.

The model we use is similar to our main model, but without individual NPI effects. Instead of estimating individual NPI effects, we estimate the instanteanous changes, or "shocks", in R_t, whenever any NPIs are implemented or lifted. Our previously estimated NPIs effects, and their effect parameters, are no longer part of the model we use here. Instead, we estimate the change in R_t for each shock individually. This estimates the observed change in R_t at the dates of NPI implementation/lifting. For evaluation, the observed change is then compared to the change predicted by the first or second-wave effects. Since other factors may affect the degree of transmission gradually, we additionally allow R_t to vary smoothly over time using the random walk of our default model.

In order to compute the evaluation metrics (bias and relative prediction errors), we take the posterior median instantaneous change in transmission at each shock as the observed instantaneous change in transmission, and relate this to the predicted change in the transmission using the NPIs. Changes in transmission are measured using percentage changes in R_t.

Details. We allow for shocks in the R_t curve for each country whenever the implemented NPIs change, either according to the NPI definitions used in this study or the ones used in Brauner et al. (*9*) respectively. We place a weakly informative prior distribution over the change in R_t at each shock date ($\beta_{\text{shock}} \sim \text{Normal}(0, 0.25^2)$).

There are some further differences between the model used here and our primary model. We use only reported cases and not deaths, since increased vaccination coverage will change the infection-fatality ratio considerably. Furthermore, we remove the infection noise, or equivalently set $\sigma_N = 0$, as we model on the national level, which has far higher case counts.

Neither this study, nor Brauner et al. (*9*) could disentangle the individual effects of school and university closures robustly. Therefore, for both waves, we equally distribute the com-

bined reduction of closing all educational institutions across primary schools, secondary schools, and universities. We also considered assigning the full effect to school closures—school pupils significantly outnumber university students. Compared to the results below, this change further increases the advantage of the second wave estimates.

1.2.3. Results

Across the 6 countries included, the analysed period contains 24 dates on which NPIs status changes ("shock dates" in methods description above), with an average of 1.6 NPIs changing per date, 39 NPI changes in total.

The changes in R_t upon implementing/lifting NPIs predicted by first-wave estimates are on average 18 percentage points larger than the observed changes in R_t. In contrast, our second-wave estimates only overestimate the observed changes by 2 percentage points. They are thus largely unbiased estimates for observed changes in R_t upon implementing/lifting NPIs in the third wave.

The second wave estimates are not only less biased, but also more accurate in predicting changes in R_t in the third wave. As shown in Supplementary Figure S2, using second wave estimates results in smaller average prediction errors than using first-wave estimates in each of the 6 countries included.

Anecdotally, these results match our observations. For example, the reopenings of schools in the UK and Ireland in March 2021 were not associated with large increases in transmission, in contrast to predictions from first-wave studies (*9, 12, 13*).

We also considered an alternative prediction strategy—fitting a smoothly varying R_t using a random walk in transmission. Such a model cannot account for large, instanteneous changes in transmission. Nevertheless, this approach can be used to observe the change in R_t 5 days before and after implemented NPIs change. We compare the observed change to a prediction that uses the specific NPIs that changed, and their effect estimates. This approach lead to similar results as described above.

1.2.4. Limitations

The instantaneous reduction in R_t is estimated individually for each shock and thus may be influenced by country specific unobserved factors such as changing behaviour or un-

recorded NPIs. For example, the model may estimate an increase in transmission for a shock at NPI implementation when B.1.1.7. is growing in a particular country, even if the NPI implementation actually reduced transmission. Further, we assumed that the Infection-Ascertainment Rate (IAR) is constant over time, but increased vaccination coverage may change the IAR.

Relative Average Prediction Error

Second wave MSE as a percentage of first wave MSE
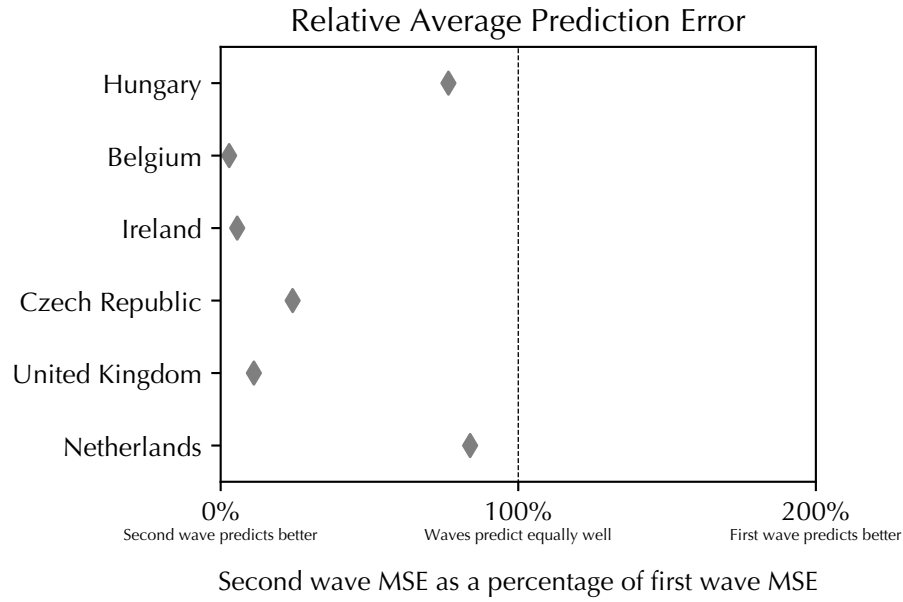
Supplementary Figure S2: Ratio of errors for predictions of R_t with results from the first and second waves. First-wave results from Brauner et al. (9). Observed short-term reductions in R_t were estimated using a model with random walk and shocks based on national case data across 6 European countries. Prediction error is measured by calculating the mean squared error across NPI changes for both the first wave and second wave NPIs, and averaging by country. The relative error is the ratio between the prediction errors.

Supplementary Note 2. Validation

2.1. Univariate sensitivity analysis

Sensitivity analysis reveals the extent to which results depend on uncertain parameters and modelling choices, and can diagnose model misspecification and excessive collinearity in the data (*14*). We vary many of the components of our model and recompute the NPI effectiveness estimates[a]. Overall, we perform 17 sensitivity analyses with 86 experimental conditions. Supplementary Table S1 summarises our sensitivity analyses and their categories (matching Fig. **3**).

**Supplementary Table S1: Sensitivity analysis categories.**

Category Name	Analyses
Delay distributions	Generation interval mean (Supplementary Figure S3)
	Infection to confirmation delay (Supplementary Figure S4)
	Infection to reported-death delay (Supplementary Figure S5)
Data	Bootstrap (Supplementary Figures S6,S7, and S8)
	Maximum VOC permitted (Supplementary Figure S9)
	Schools delayed (Supplementary Figure S10)
Prior distributions	Infection noise scale (Supplementary Figure S11)
	σ_R prior (Supplementary Figure S12)
	Output noise scale prior (Supplementary Figure S13)
	Intervention prior (Supplementary Figure S14)
	$\tilde{R}_{0,l}$ mean (Supplementary Figure S15)
	$\tilde{R}_{0,l}$ scale (Supplementary Figure S16)
	Seeded infections prior (Supplementary Figure S17)
Model structure	Random walk period (Supplementary Figure S18)
	Seeding days (Supplementary Figure S19)
Unobserved factors	NPI leaveout (Supplementary Figures S20,S21,S22, and S23)
	Time-varying IFR/IAR adjustment (Supplementary Figure S24)

---

[a] As in the main text, NPI effectiveness is displayed as the percentage reduction in R_t, computed as $1 - \exp(-\beta_i)$.

2.1.1. Sensitivity to delay distributions

Supplementary Figures S3, S4 and S5 show the sensitivity of the NPI effectiveness estimates to the means of the distributions of the generation interval, the delay between infection and case reporting, and the delay between infection and death reporting.

Consistent with Flaxman et al. (*15*), Brauner et al. (*9*) and Sharma et al. (*12*), we find that a shorter mean generation interval implies smaller values of R_t when the epidemic is growing, and therefore lower effectiveness estimates. Furthermore, the sensitivity to the delay between infection and case reporting appears to be larger than the sensitivity to the delay between infection and death reporting.
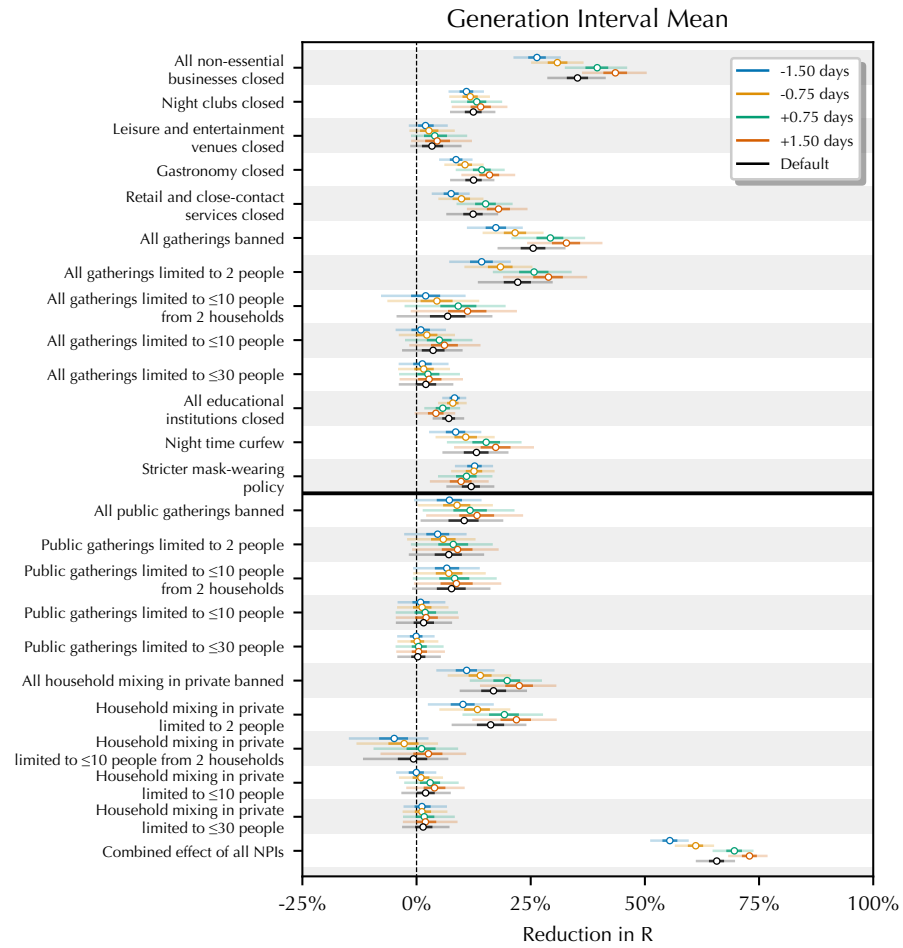
Supplementary Figure S3: Sensitivity of NPI effect estimates to the mean of the generation interval. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

Mean Delay from Infection to Case Reporting
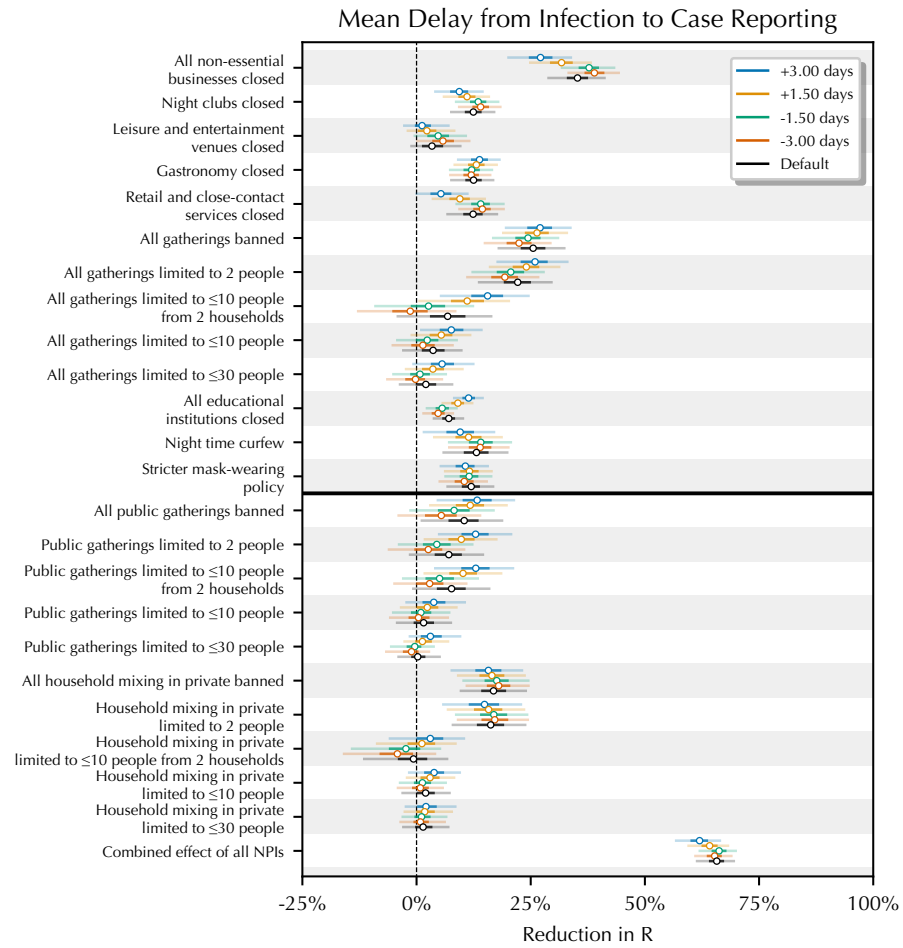
Supplementary Figure S4: Sensitivity of NPI effect estimates to the mean of the delay from infection to case reporting. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
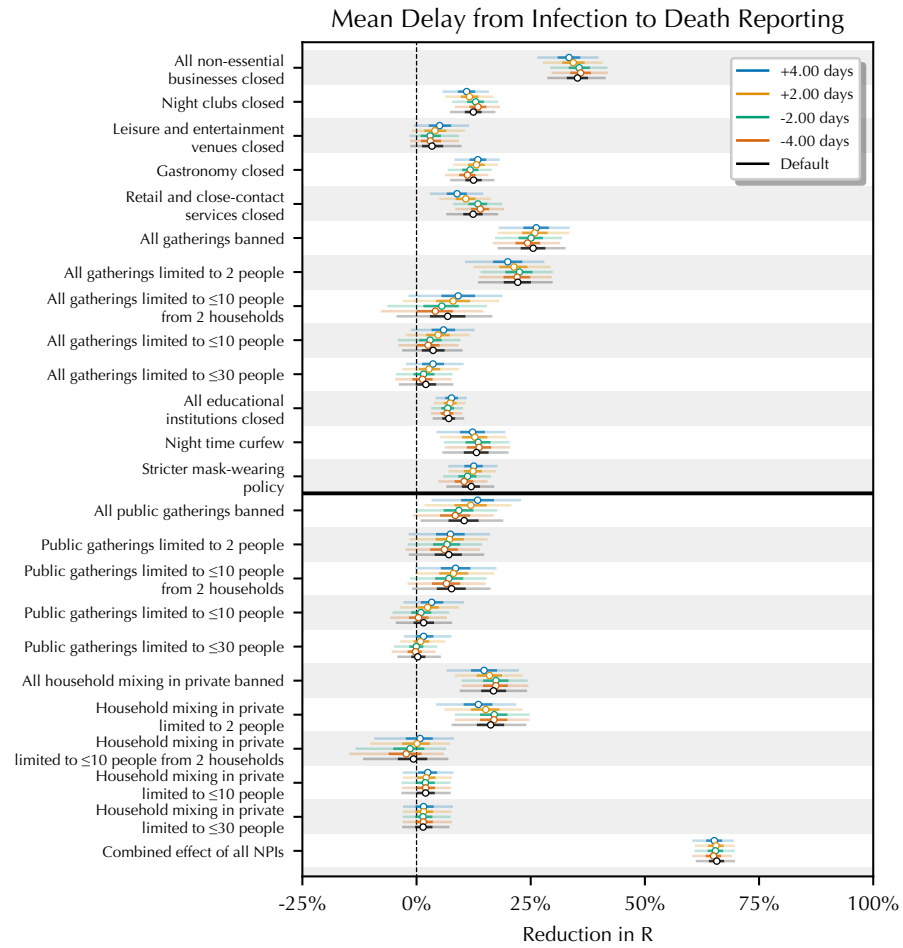
Supplementary Figure S5: Sensitivity of NPI effect estimates to the mean of the delay from infection to death reporting. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

15

2.1.2. Sensitivity to data

Supplementary Figures S6, S7 and S8 show the sensitivity of NPI effect estimates when bootsrapping over the locations in our dataset. For each random seed, we sample 114 locations **with replacement** from our pool of locations. Bootsrapping assesses the extent to which our NPI effect estimates depend on the specific locations that were included. Each bootstrap typically samples around 75 *unique* locations.
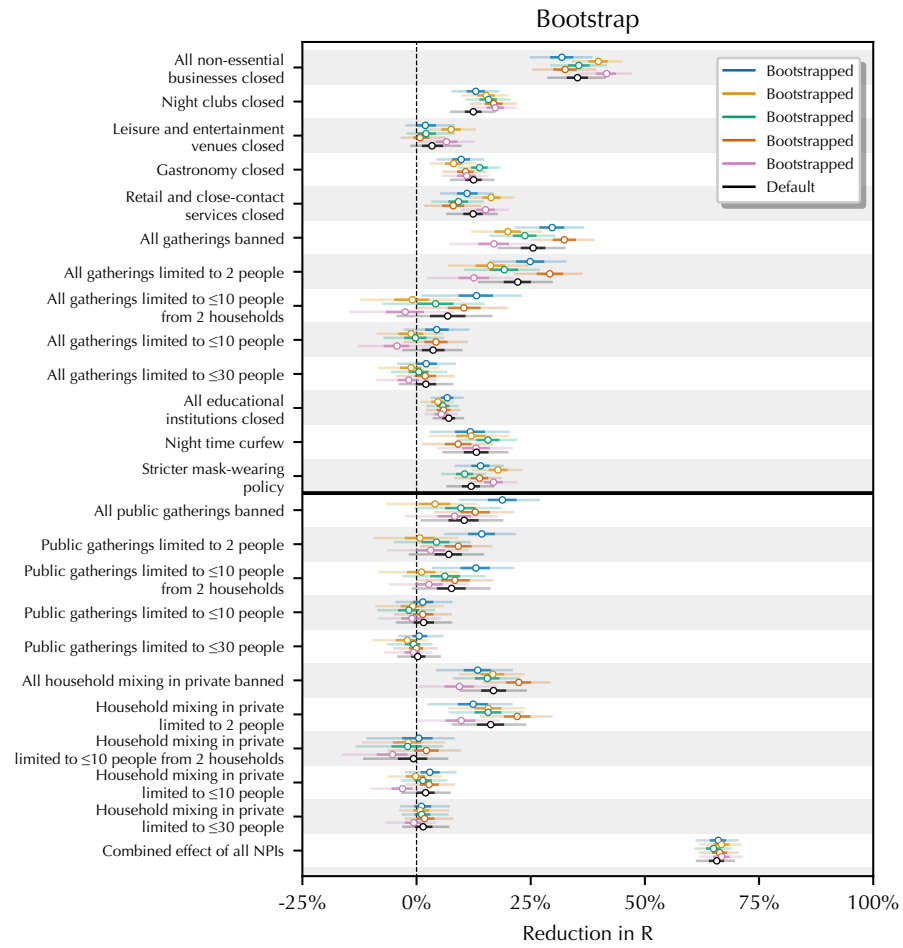
Supplementary Figure S9 shows the sensitivity of NPI effect estimates when the maximum permitted fraction of cases attributed to the b.1.1.7 Variant of Concern (VOC) is changed. This sensitivity analysis only affects locations within England.
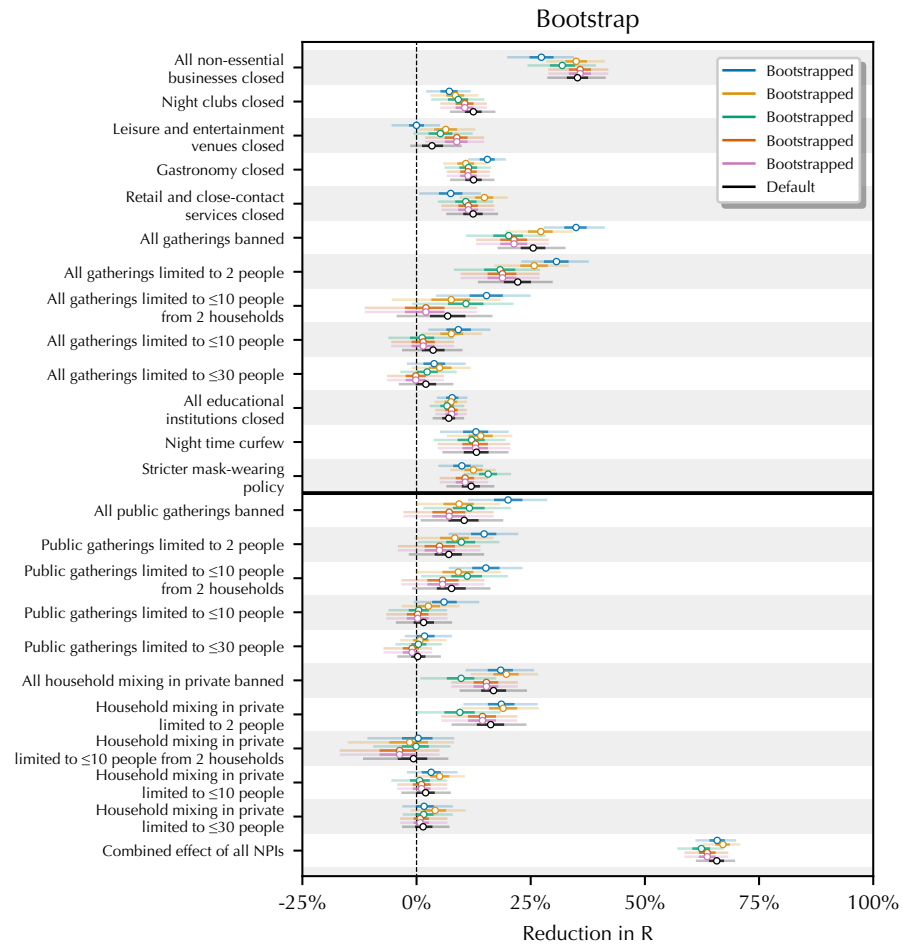
Supplementary Figure S6: Sensitivity of NPI effect estimates using random bootstrapped sets of regions. Each experimental condition above uses a different random seed to generate the bootstrapped dataset. In other words, each condition uses a different subset of regions to fit the model. This figure shows seeds 1-5. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
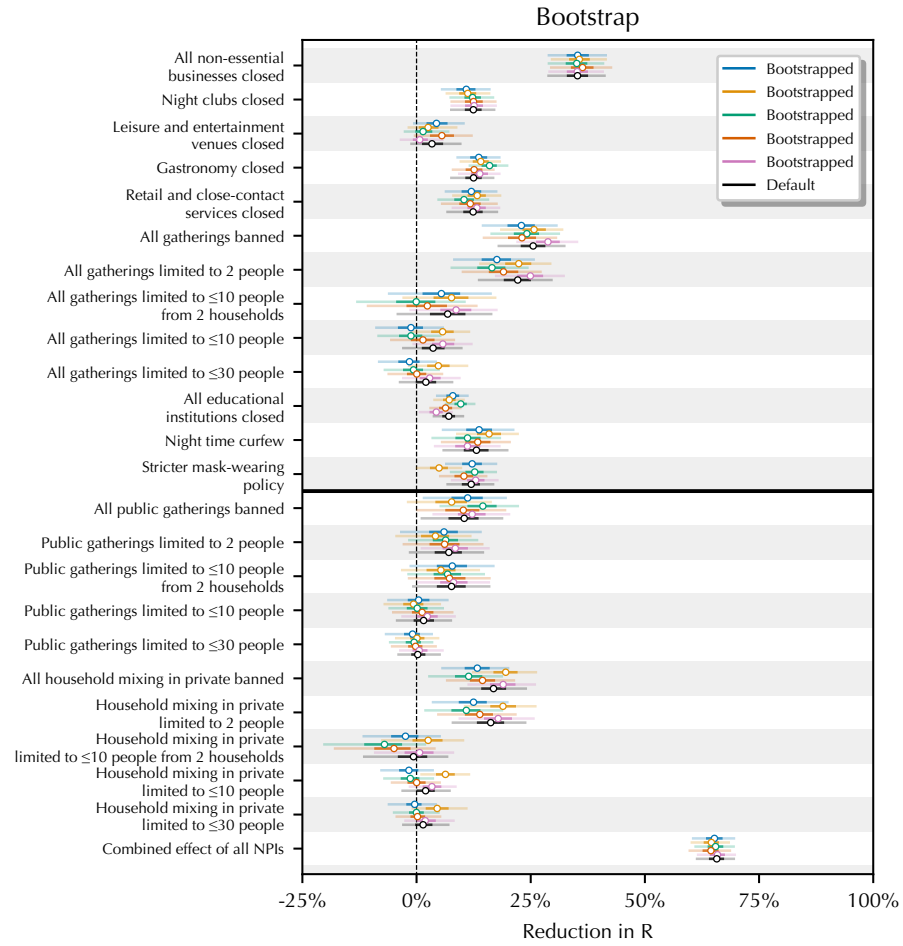
Supplementary Figure S7: Sensitivity of NPI effect estimates using random bootstrapped sets of regions. Each experimental condition above uses a different random seed to generate the bootstrapped dataset. In other words, each condition uses a different subset of regions to fit the model. This figure shows seeds 6-10. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
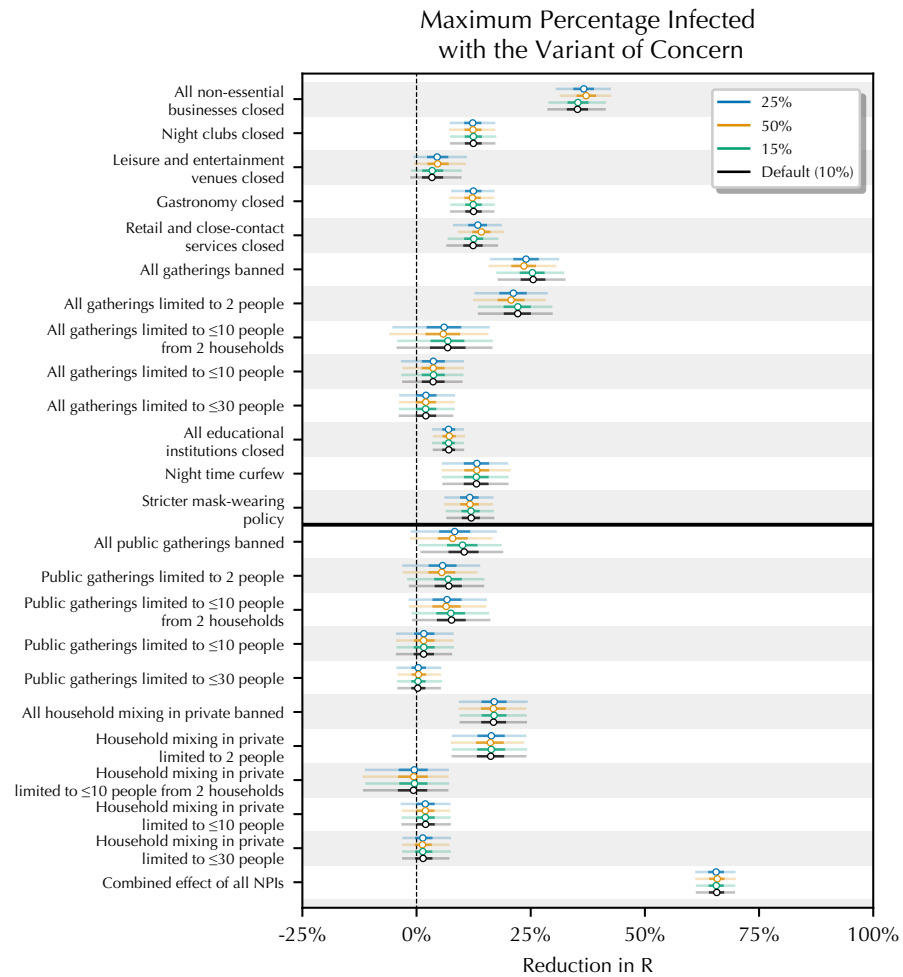
Supplementary Figure S8: Sensitivity of NPI effect estimates using random bootstrapped sets of regions. Each experimental condition above uses a different random seed to generate the bootstrapped dataset. In other words, each condition uses a different subset of regions to fit the model. This figure shows seeds 11-15. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

Maximum Percentage Infected with the Variant of Concern

Supplementary Figure S9: Sensitivity of NPI effect estimates to changing the maximum proportion of the variant of concern allowed before the window of analysis is ended. Note: this sensitivity analysis only affects England. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

2.1.3. The individual effects of school and university closures are sensitive

Some evidence points towards underreporting of COVID-19 cases amongst school-age children. For example, children may be more likely to be asymptomatic, and thus more likely to remain untested (*16*). An important consequence of this is that the effect of school openings and closures may be lagged. For example, suppose COVID-19 infections amongst school-age children remain undetected, and that school opening increases transmission. Then, if schools were to open, the number of reported cases in a location would only rise once children, who were infected when schools opened, subsequently infect members of the adult population. To investigate this, we delay the intervention time series for school closure NPIs by approximately one generation interval–the average time between successive infections in a transmission chain.

Supplementary Figure S10 shows the sensitivity of the school closure and universities closed NPIs when we delayed the school NPI time series by one mean generation interval. While the total effect of closing *all* education institutions remains comparatively stable, the relative effectiveness of school and university closures shifts clearly. If we believed that delay between school closures and reported cases and deaths is similar to other NPIs, we would conclude that transmission in universities is more significant than in schools. However, if we believed that school closures had an *additionally* delayed effect on reported cases and deaths in comparison to other NPIs, we would conclude the opposite. Instead, we conclude that we cannot robustly disentangle the *individual* effects of school closures and university closures; as such, we report their *combined* effect, which is more stable.
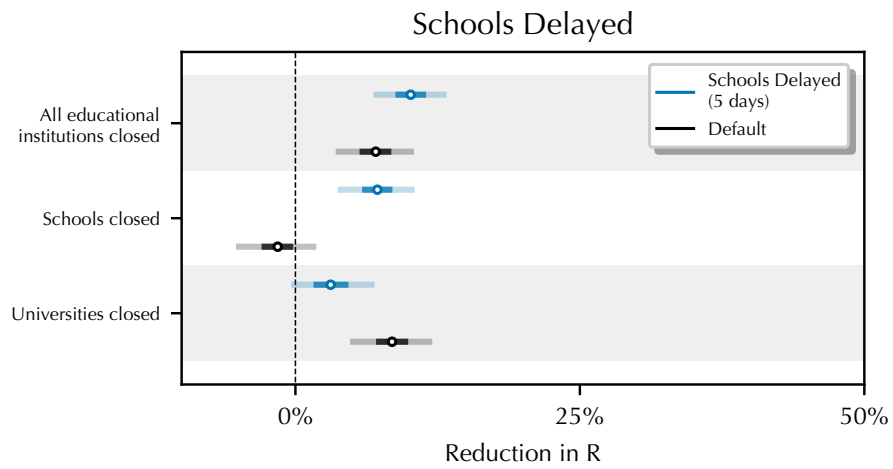
Supplementary Figure S10: Sensitivity of effect estimates for education NPIs when delaying school openings and closures by one generation interval (5 days). Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

2.1.4. Sensitivity to prior distributions

Supplementary Figure S11 shows the sensitivity of NPI effect estimates to the infection noise scale.

Supplementary Figure S12 shows the sensitivity of NPI effect estimates to the prior over the random walk noise scale, σ_R.
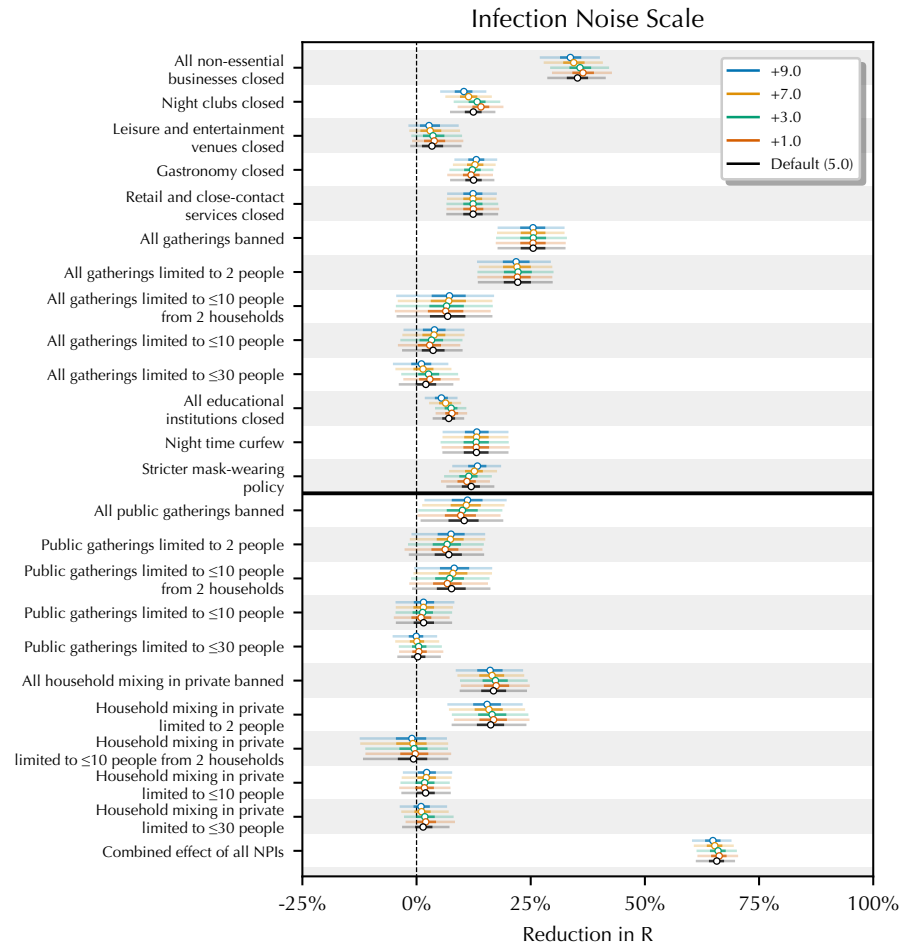
Supplementary Figure S13 shows the sensitivity of NPI effect estimates to the prior over the country-specific output noise scales, $\Psi_c^{(C)}$ and $\Psi_c^{(D)}$.

Supplementary Figure S14 shows the sensitivity of NPI effect estimates to the prior over the NPI effects.

Supplementary Figure S15 shows the sensitivity of NPI effect estimates to the mean of the prior over $\tilde{R}_{0,l}$. Recall that $\tilde{R}_{0,l}$ is the reproduction number at the start of the window of analysis, supposing no NPIs are active.

Supplementary Figure S16 shows the sensitivity of NPI effect estimates to the scale/standard deviation of the prior over $\tilde{R}_{0,l}$. Recall that $\tilde{R}_{0,l}$ is the reproduction number at the start of the window of analysis, supposing no NPIs are active.

Supplementary Figure S17 shows the sensitivity of NPI effect estimates to the scale of the prior over the seeded infections. All priors are of the form Lognormal$(0, \sigma)$.

Supplementary Figure S11: Sensitivity of NPI effect estimates to the infection noise scale. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
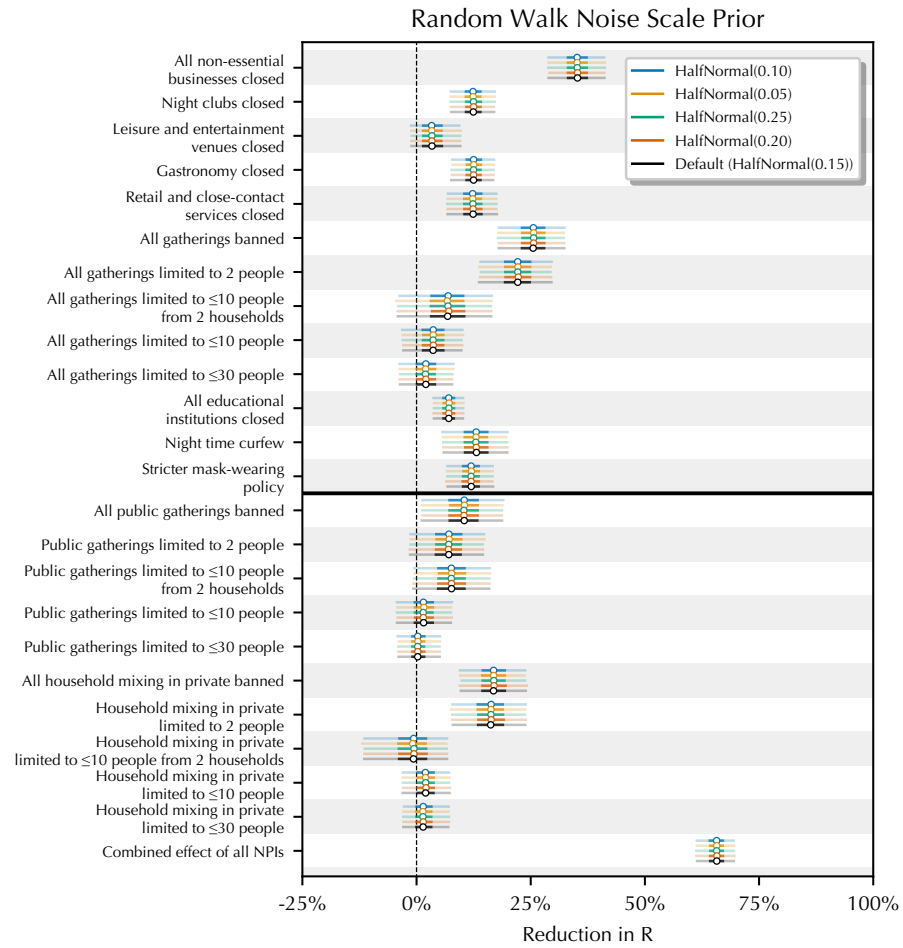
Supplementary Figure S12: Sensitivity of NPI effect estimates to the prior over the scale of the latent random walk. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
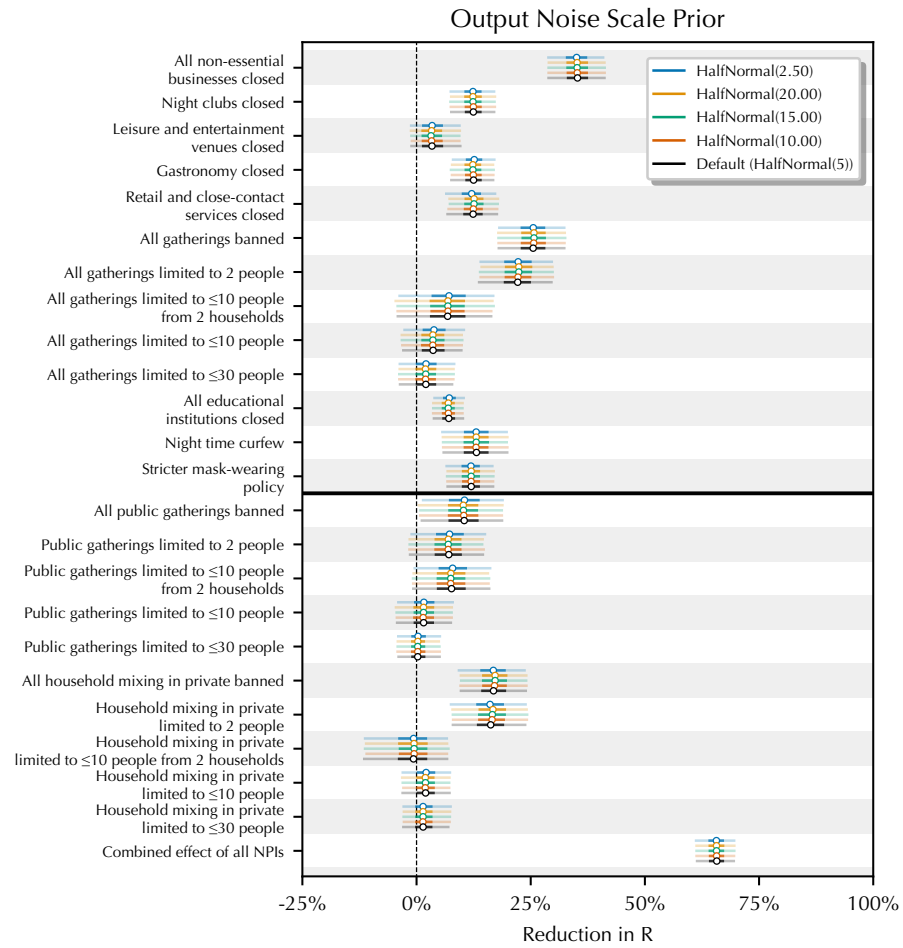
Supplementary Figure S13: Sensitivity of NPI effect estimates to the prior over the output noise scale. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
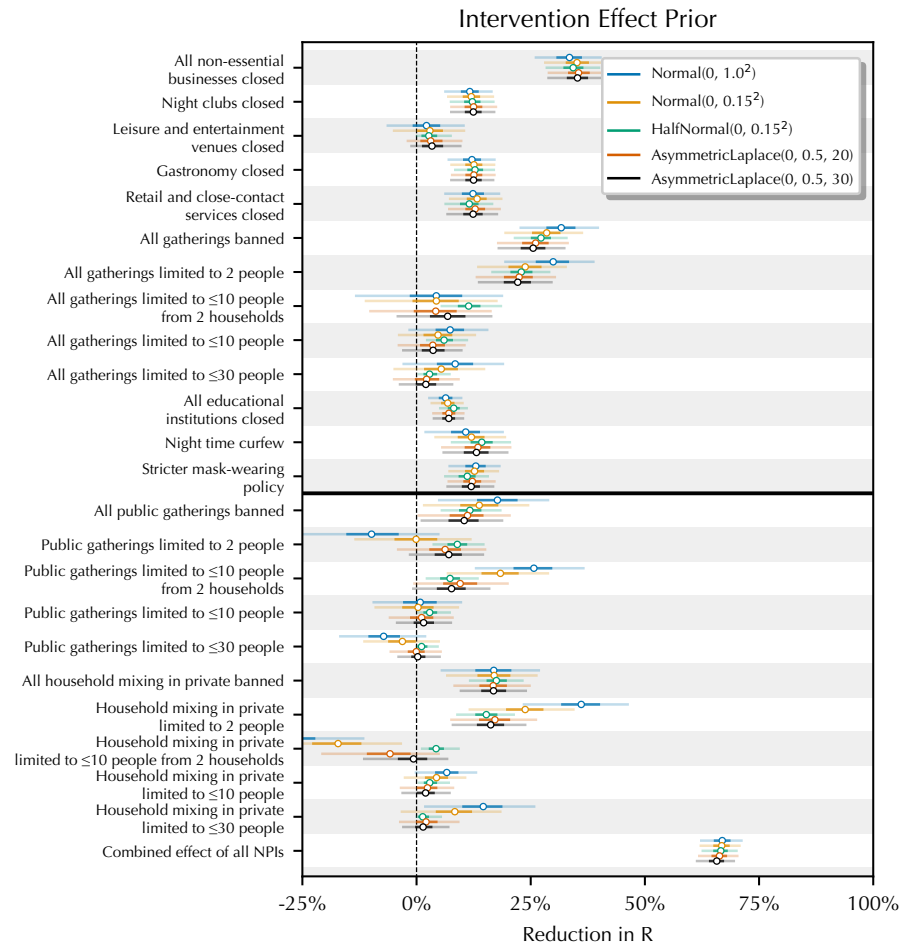
Supplementary Figure S14: Sensitivity of NPI effect estimates to the prior over the intervention effect. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

Supplementary Figure S15: Sensitivity of NPI effect estimates to the mean value of the prior over $\tilde{R}_{0,l}$ i.e., the value of the reproduction number at the start of the window of analysis, supposing no NPIs were active. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
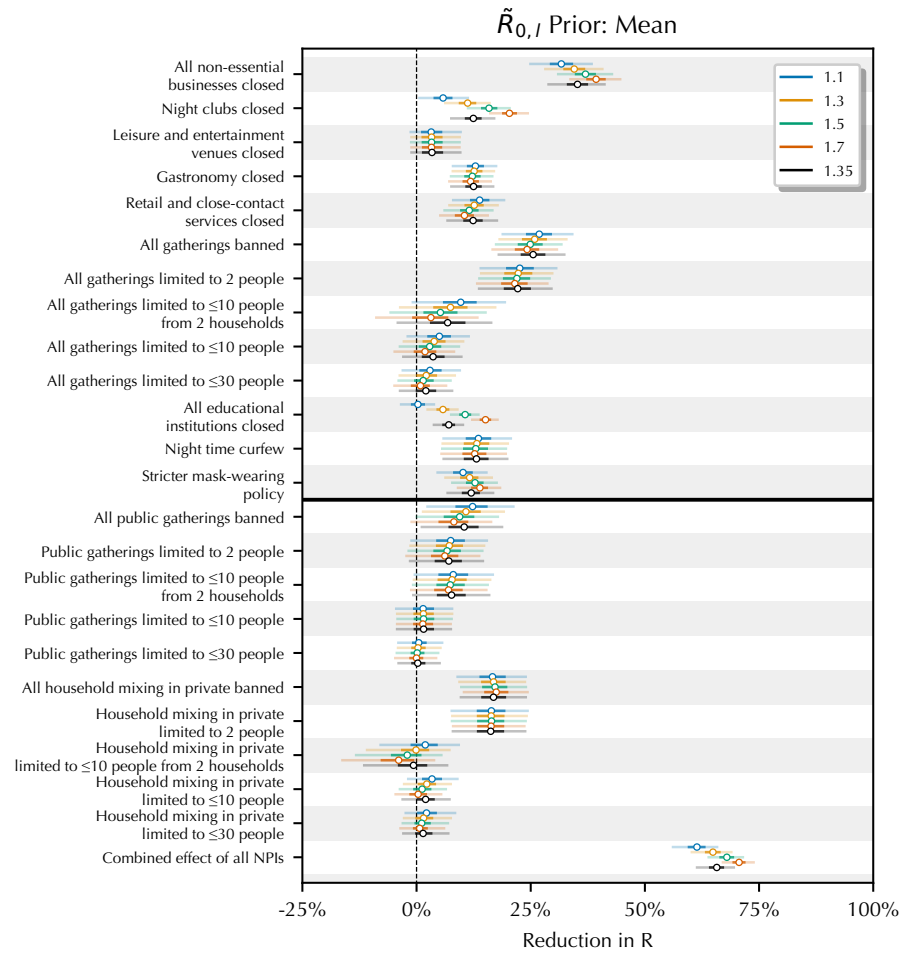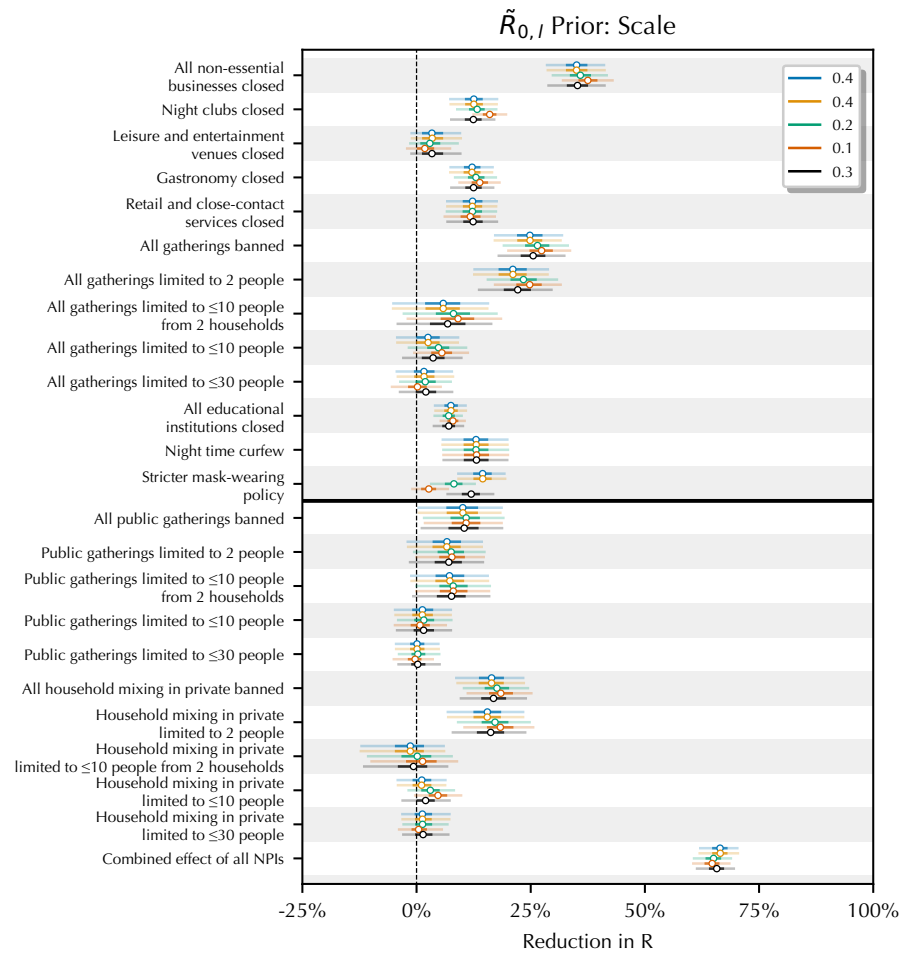
Supplementary Figure S16: Sensitivity of NPI effect estimates when the variability/scale of the prior over $\tilde{R}_{0,l}$ **i.e., the value of the reproduction number at the start of the window of analysis, supposing no NPIs were active. Posterior percentage reductions in** R_t **shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.**

Supplementary Figure S17: Sensitivity of NPI effect estimates to the prior over over the seeded infections. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
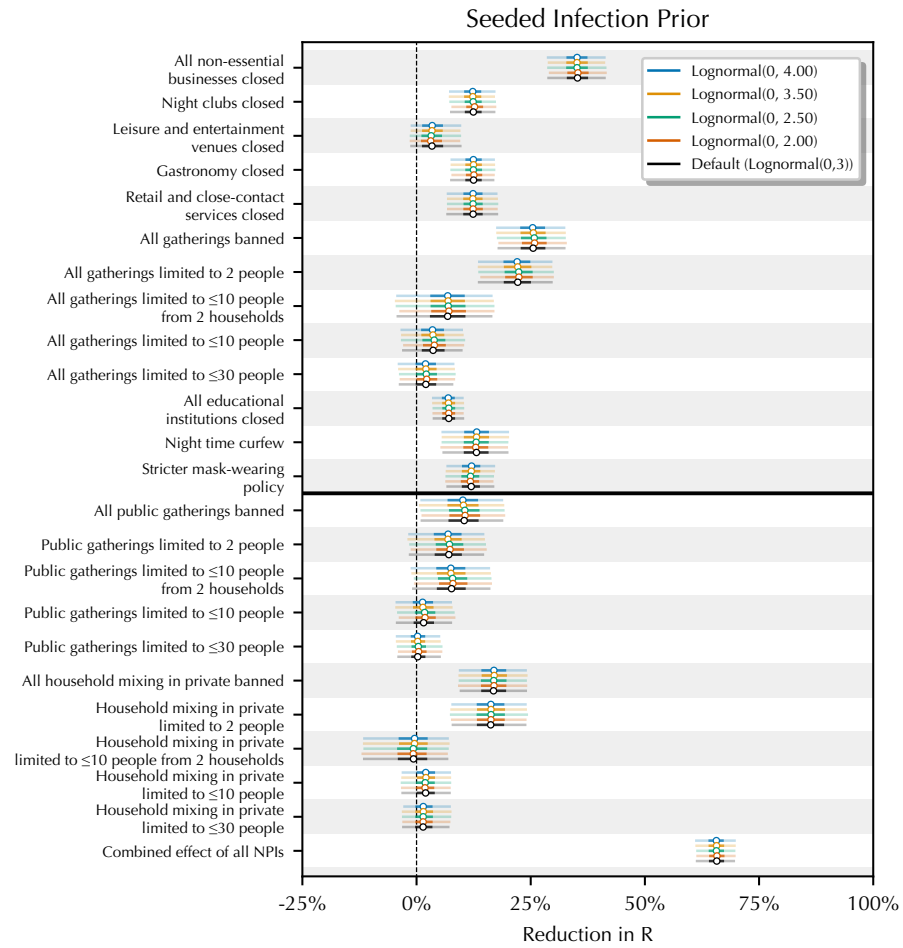
2.1.5. Sensitivity to model structure

Supplementary Figure S18 shows the sensitivity of NPI effect estimates to the period of the latent random walk. If the period is 5 days, for example, then the value of $R_{t,l}$ may change (without a change of interventions) every 5 days.

Supplementary Figure S19 shows the sensitivity of NPI effect estimates to the number of days of seeded infections.
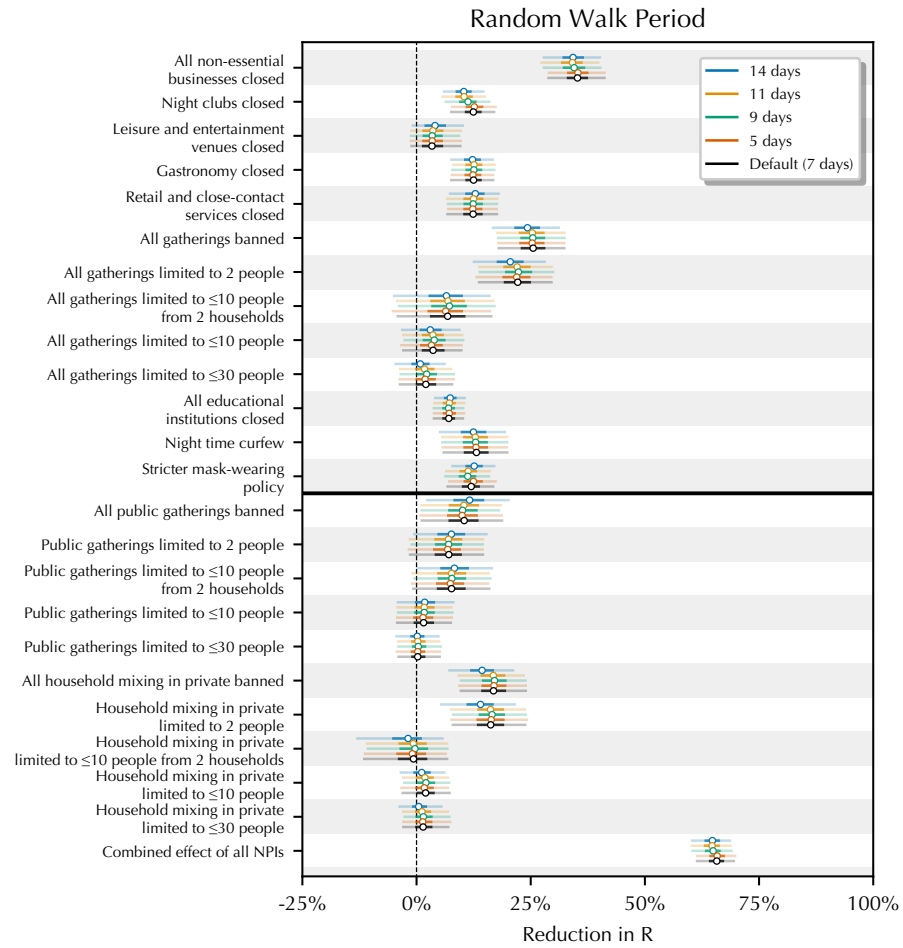
Supplementary Figure S18: Sensitivity of NPI effect estimates to the period of the random walk. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

Supplementary Figure S19: Sensitivity of NPI effect estimates to the number of days of seeded infections. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

2.1.6. Sensitivity to unobserved factors

Our data neither captures all of the NPIs that were implemented nor directly measures voluntary behaviour changes. Since these factors influence R_t, we must be wary of their effect being attributed to the observed NPIs. Unobserved factors can bias results if their timing is correlated with the timing of the observed NPIs (*17*).

We investigate this by assessing how much effectiveness estimates change when previously when observed factors are excluded. This is best practice for addressing unobserved factors, such as confounders (*18,19*), and follows Sharma et al. (*12*). Note: if we remove a component of a hierarchical NPI, we do not show the same hierarchical NPI, since its value may no longer be defined. For example, if we were to exclude retail closures, the effect of closing all non-essential businesses is not displayed.

Supplementary Figures S20, S21, S22, and S23 shows NPI effectiveness estimates when previously observed NPIs are excluded in turn. Considering that some of the excluded NPIs have strong estimated effects when included, and are correlated with other NPIs, this degree of robustness is encouragingly high.

Supplementary Figure S20: Sensitivity of NPI effect estimates to leaving out recorded interventions, simulating unobserved confounding effects on transmission. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.
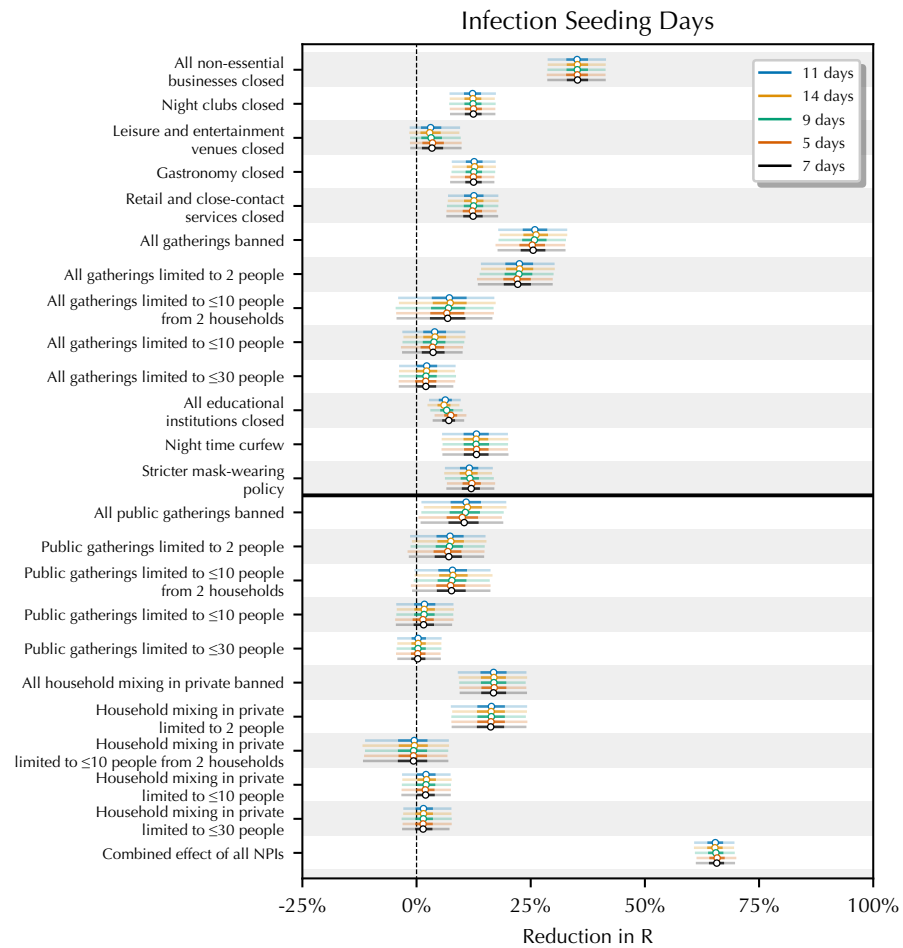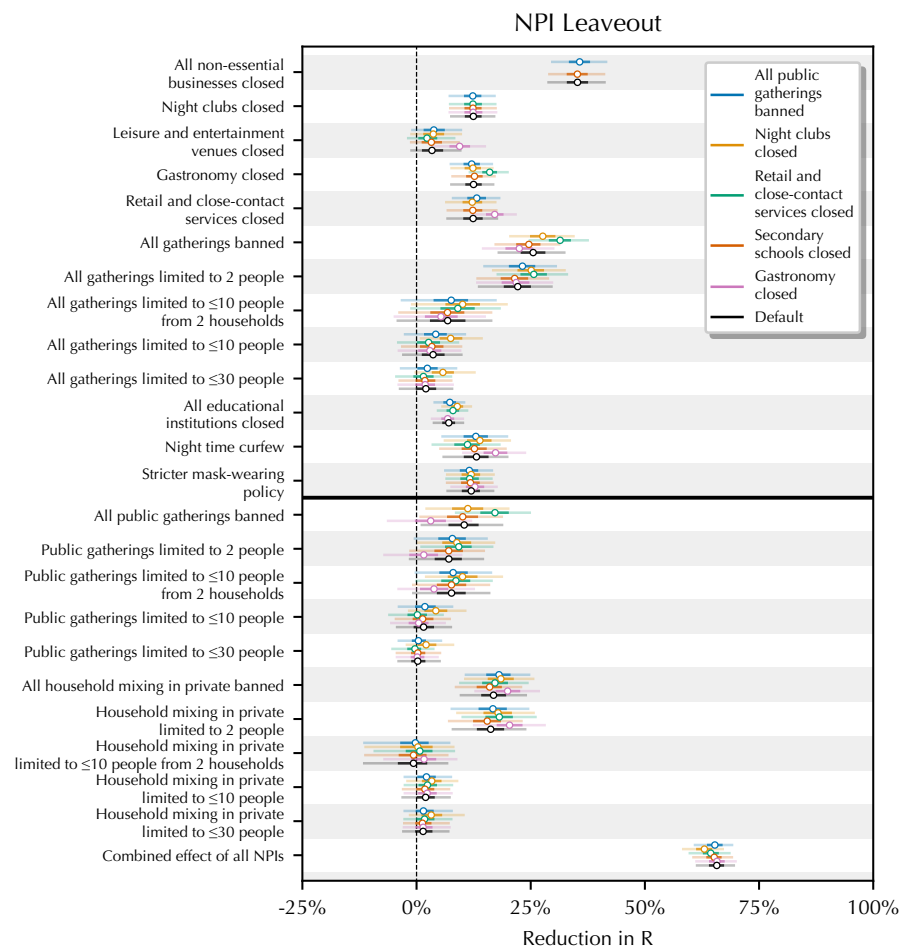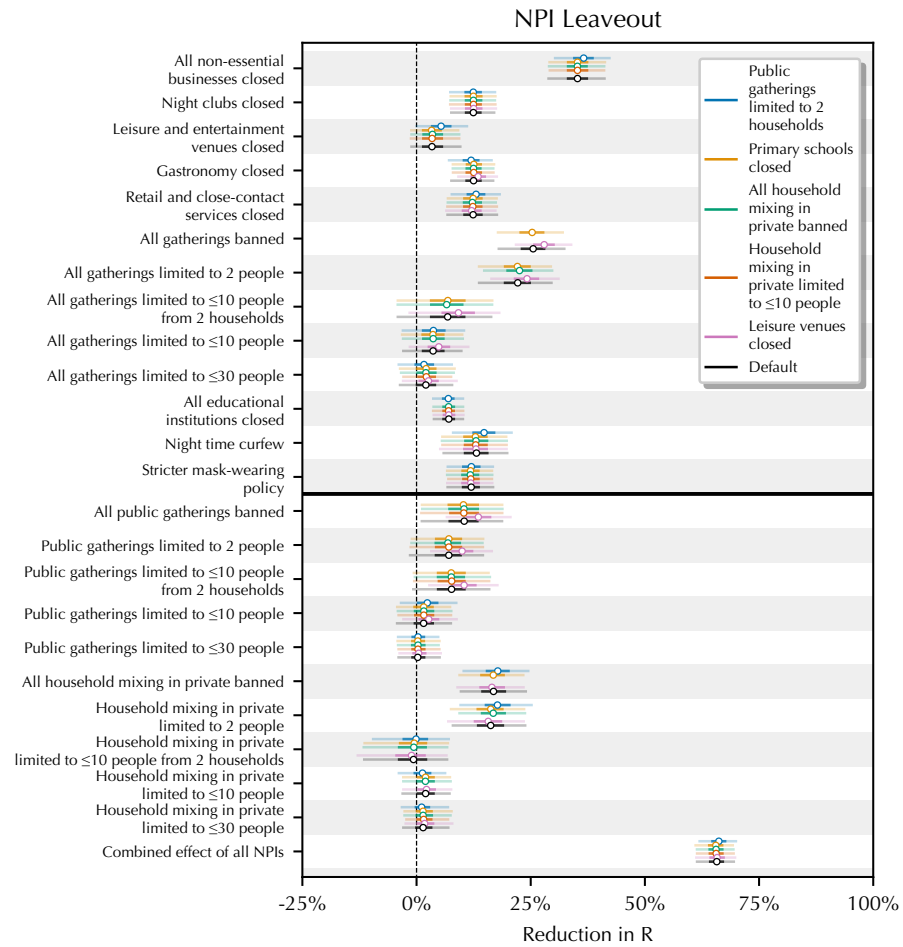
Supplementary Figure S21: Sensitivity of NPI effect estimates to leaving out further recorded interventions, simulating unobserved confounding effects on transmission. Continuation of the previous figure.
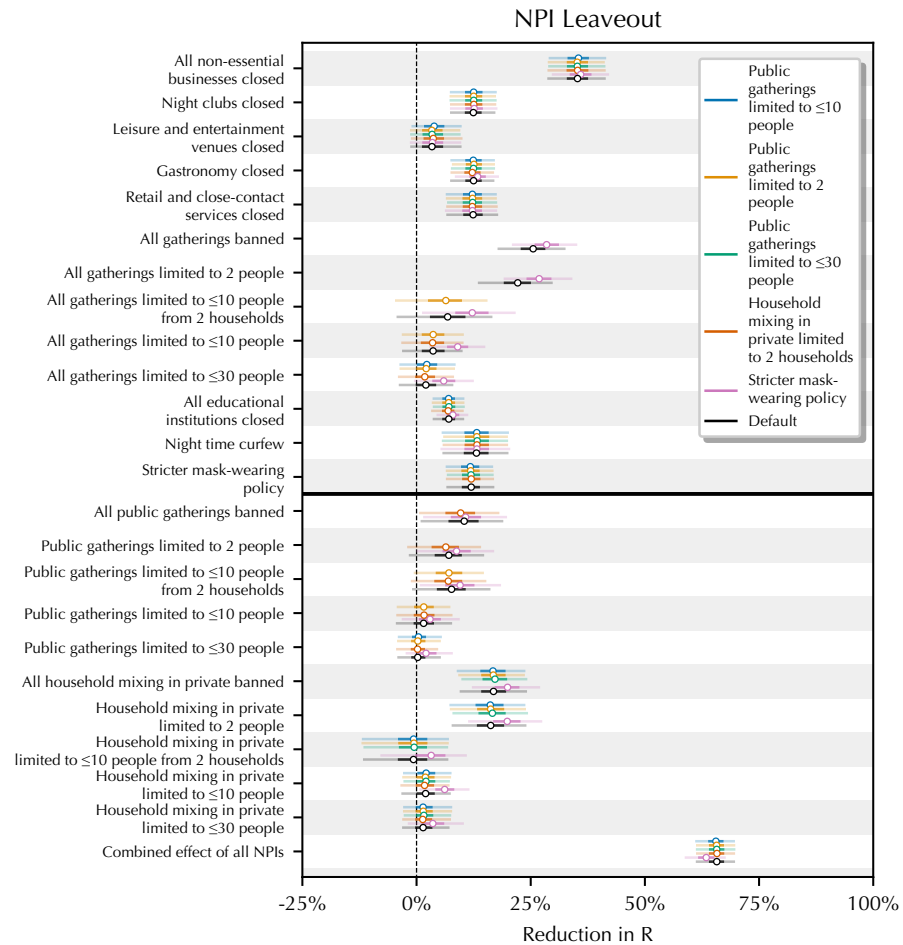
Supplementary Figure S22: Sensitivity of NPI effect estimates to leaving out further recorded interventions, simulating unobserved confounding effects on transmission. Continuation of the previous figure.
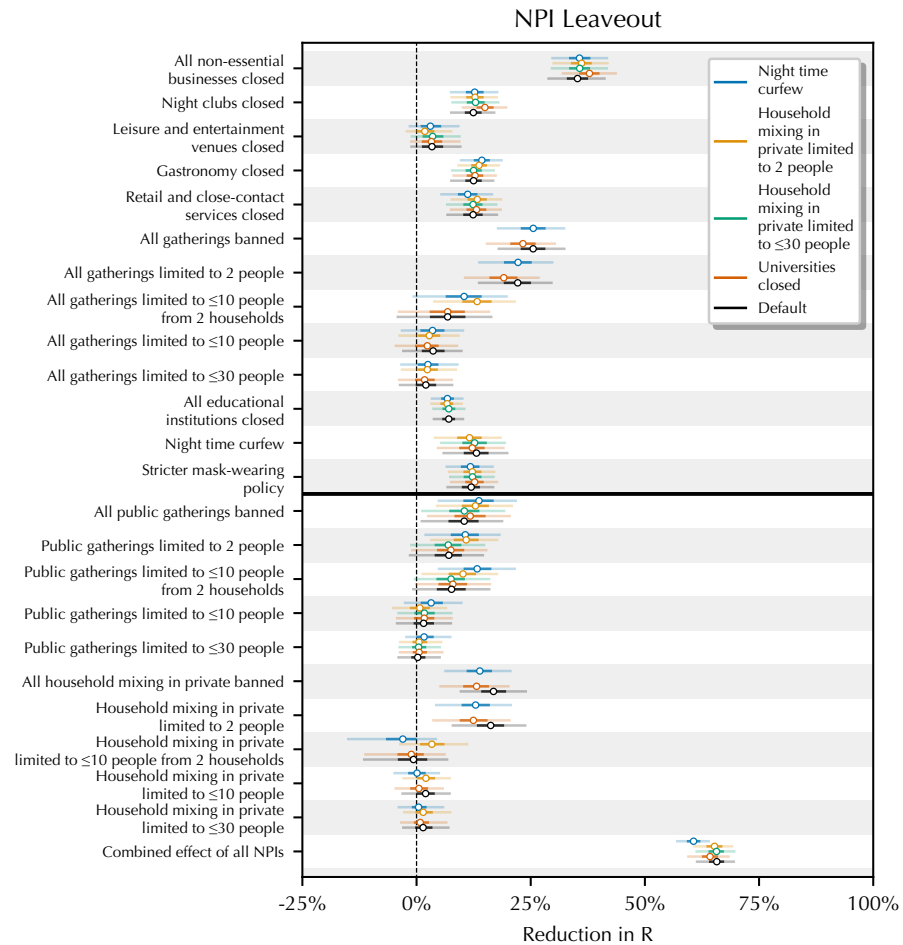
Supplementary Figure S23: Sensitivity of NPI effect estimates to leaving out further recorded interventions, simulating unobserved confounding effects on transmission. Continuation of the previous figure.

2.1.7. Sensitivity to Changing IFR and IAR

Recall that our model assumes a time-constant Infection-Fatality Rate (IFR)–the proportion of infections that result in fatalities–and a time-constant Infection-Ascertainment Rate (IAR)–the proportion of infections that are confirmed as COVID-19 cases. In reality, the IFR and IAR may change over time. For example, improvements in COVID-19 treatment and increasing testing capacity may cause the IFR to decrease and the IAR to increase. Therefore, we assess the sensitivity of our NPI effectiveness estimates when accounting for the time-varying IFR and IAR.

In particular, we use estimates of the IFR and the IAR in England, taken from Mishra et al. (*20*). The time-varying estimates of IFR/IAR are estimated using seroprevalence data from ONS (*21*) and REACT (*22*) along with case and death time series for England. While the trends in the IAR and IFR are unique to every country, the trend in England is indicative of how, and how much the IFR and IAR may change in other countries. Thus, to investigate the time-constant IFR and IAR assumption, we apply the estimates from England to *all* of our locations. Specifically, instead of modelling the number of infections that are later confirmed/result in death as

$$N_{t,l}^{(C)} = N_{t,l}, \text{ and } N_{t,l}^{(D)} = \text{IFR}_l \cdot N_{t,l},\tag{B.1}$$

as in the main model, we instead use:

$$N_{t,l}^{(C)} = \widehat{\text{IAR}}_{t,\text{england}} \cdot N_{t,l}, \quad \text{and } N_{t,l}^{(D)} = \widehat{\text{IFR}}_{t,\text{england}} \cdot N_{t,l}.\tag{B.2}$$

$\widehat{\text{IAR}}_{t,\text{england}}$ is the time-varying estimate for the ascertainment rate in England, while $\widehat{\text{IFR}}_{t,\text{england}}$ is the corresponding infection fatality rate estimate.

Supplementary Figure S24 shows the NPI effectiveness estimates when adjusting the time-varying IFR/IAR. We find low sensitivity to this change, suggesting that the time-constant IFR and IAR is not a key limitation of this work.

Undercounting Adjustment
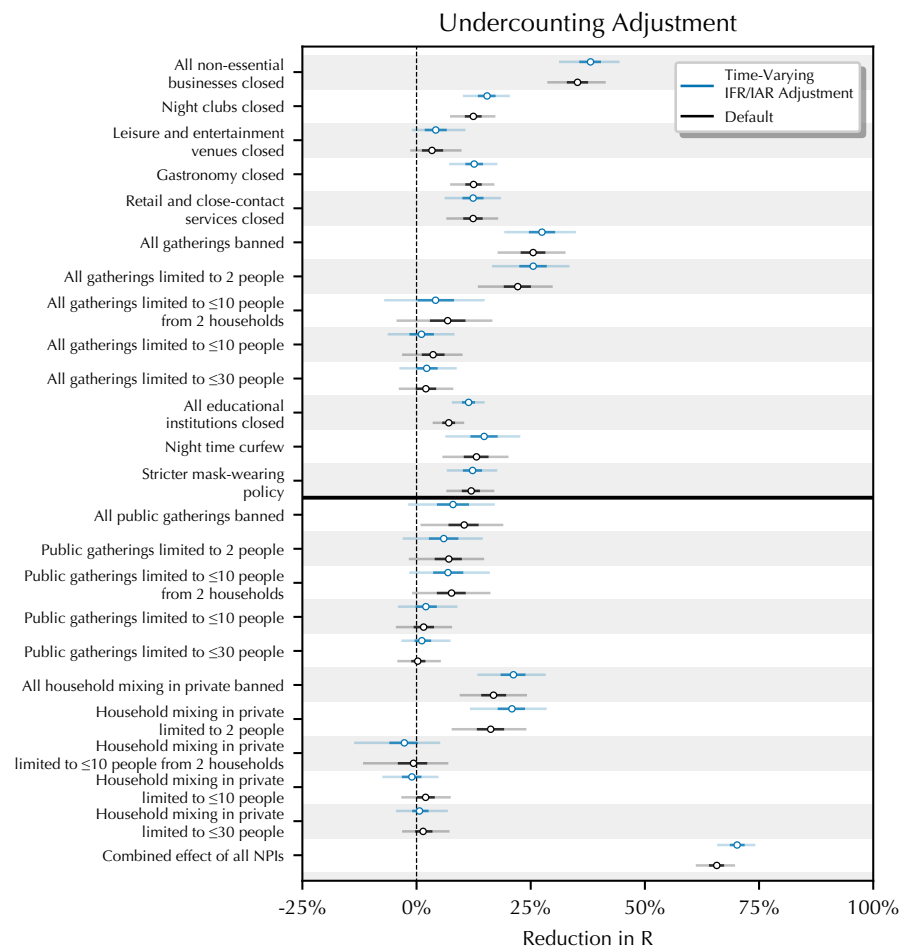
Supplementary Figure S24: Sensitivity of NPI effect estimates when adjusting for time-varying case under-counting (i.e., a time-varying ascertainment rate), as well as a time-varying infection-fatality rate. Posterior percentage reductions in R_t shown. Markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

2.2. Multivariate sensitivity analysis

Sensitivity analysis reveals the extent to which results depend on uncertain parameters and modelling choices, and can diagnose model misspecification and excessive collinearity in the data (*14*). By performing a *multivariate* sensitivity analysis—i.e., by varying several assumed model parameters jointly—we can understand the extend to which our NPI effectiveness estimates vary under plausible assumptions and parameter values, and thereby better understand the uncertainty of our results. Note that the variation in our results under a multivariate sensitivity analysis depends significantly on the assumed variation in underlying parameters.

To perform our multivariate sensitivity analysis, we place sampling distributions over several assumed model parameters. We rerun our analysis for 120 random seeds. For each random seed, a draw from the sampling distribution is taken, and we run our analysis using the sampled parameter values.

The sampling distributions that we use are detailed in Supplementary Table S2. Note that for the TruncatedNormal distribution, l denotes the low truncation value and u the upper truncation value. The probability density outside $[l, u]$ is 0. Further, the DiscreteUniform distribution with N items assigns equal probability mass to each item. For example, DiscreteUniform(1, 2, 3) would assign probability 1/3 to 1, 2, and 3.

Fig S26 shows how posterior median estimates vary under 30 experimental conditions, for the seeds where the generation interval was in the central 25th percentile of the sampled values. Specifically, we sampled 120 parameter values. The plot shows only the settings where the generation inteval is in the central 25th percentile, giving 30 conditions. We show this plot as the changing the generation interval typically shifts intervention effectiveness in a systematic way without pertubing NPI effectiveness ranking—see Supplementary Figure S3. This variational in NPI effectiveness is not particularly policy relevant; current estimated R values would similarly depend on the generation interval value.

Fig S26 shows how posterior median estimates vary under all 120 experimental conditions. We find that while exact effectiveness estimates vary under the experimental conditions, a broad picture of NPI effectiveness emerges.

Supplementary Table S2: Multivariate sensitivity analysis sampling distribution.

Parameter	Default value	Sensitivity analysis sampling distribution
Generation interval mean	4.83	$\text{TruncatedNormal}(\mu = 4.83, \sigma^2 = 0.75^2, l = 3.33, u = 6.33)$
Onset to case reporting delay mean	5.28	$\text{TruncatedNormal}(\mu = 5.28, \sigma^2 = 1.5^2, l = 2.28, u = 8.28)$
Onset to death reporting delay mean	5.28	$\text{TruncatedNormal}(\mu = 18.6, \sigma^2 = 1.5^2, l = 15.6, u = 21.6)$
$\tilde{R}_0$ prior mean	1.35	$\text{TruncatedNormal}(\mu = 1.35, \sigma^2 = 0.15^2, l = 1, u = 1.7)$
$\tilde{R}_0$ prior scale	0.3	$\text{TruncatedNormal}(\mu = 0.3, \sigma^2 = 0.1^2, l = 0.1, u = 0.5)$
Maximum VOC permitted	10%	$\text{DiscreteUniform}(10\%, 15\%, 25\%, 50\%)$
Random walk period	7	$\text{DiscreteUniform}(5, 7, 9, 11, 14)$
Number of days seeding	7	$\text{DiscreteUniform}(5, 7, 9, 11, 14)$
Seeded infections scale	3	$\text{DiscreteUniform}(2, 2.5, 3, 3.5, 4)$
Infection noise scale	5	$\text{DiscreteUniform}(1, 3, 5, 7, 9)$
Random walk noise scale prior scale	0.15	$\text{DiscreteUniform}(0.05, 0.1, 0.15, 0.2, 0.25)$
Output noise scale prior scale	0.15	$\text{DiscreteUniform}(2.5, 5, 10, 15, 20)$
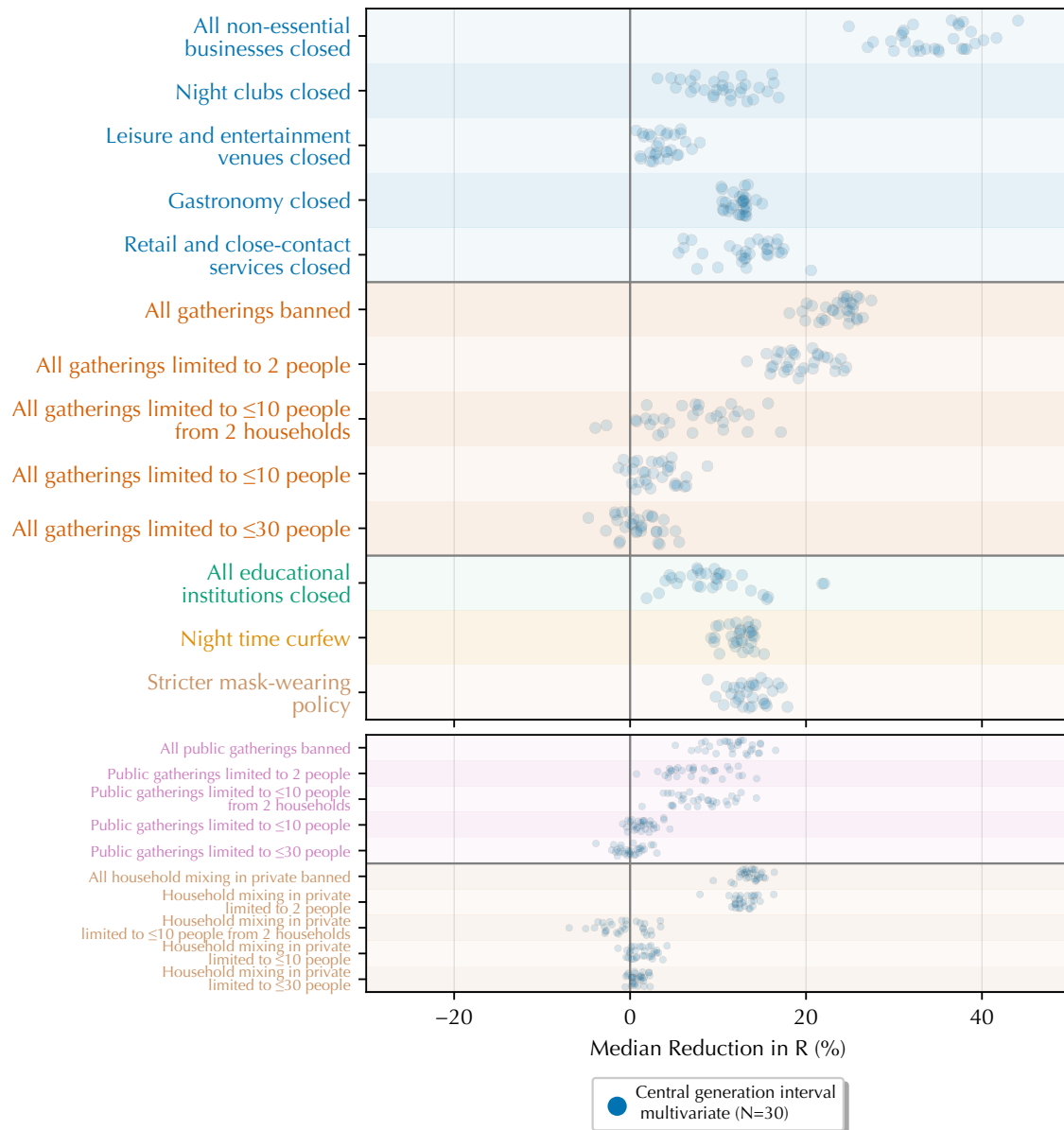
Supplementary Figure S25: Multivariate sensitivity analysis. This plot shows 30 experimental conditions, where the generation interval mean shift is in the central 25% percentile. Each dot represents the posterior median intervention effectiveness under a particular experimental condition. The parameter settings for each experimental condition are drawn according to a sampling distribution—see Supplementary Table S2.
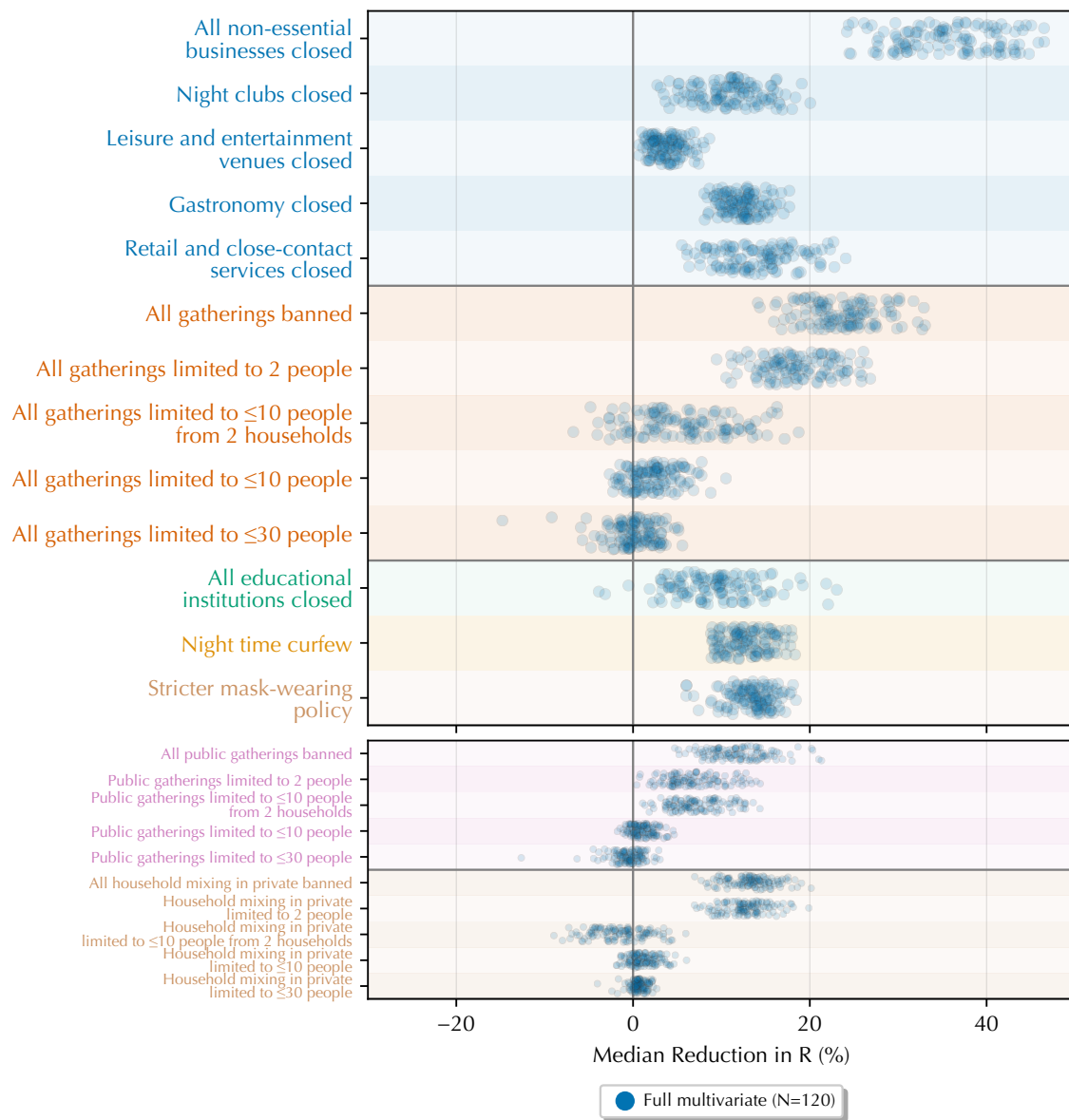
Supplementary Figure S26: Multivariate sensitivity analysis. Each dot represents the posterior median intervention effectiveness under a particular experimental condition. The parameter settings for each experimental condition are drawn according to a sampling distribution—see Supplementary Table S2.

2.3. Simulation study

We now assess whether our model can correctly recover the ground truth NPI effects. We use our model to simulate synthetic epidemics for all 114 regions, keeping the original timing and ordering of interventions and assigning our estimated effect sizes to each intervention. If we are able to recover the assigned effect sizes, we may be able to recover the "ground truth" NPI effectiveness estimates. Specifically, for simulation we assign each NPI an effect size equal to the median effect size estimated for that particular NPI by our model on observed data (main text Fig 2). Note that we use the real intervention dates to ensure that our simulation is reflective of reality, and the simulated NPIs have the same collinearity as the real NPIs.

As shown in Supplementary Figure S27, the estimates from the fitted model are close to the simulated effect size. This analysis provides evidence that our model can recover "ground truth" NPI effectiveness estimates.
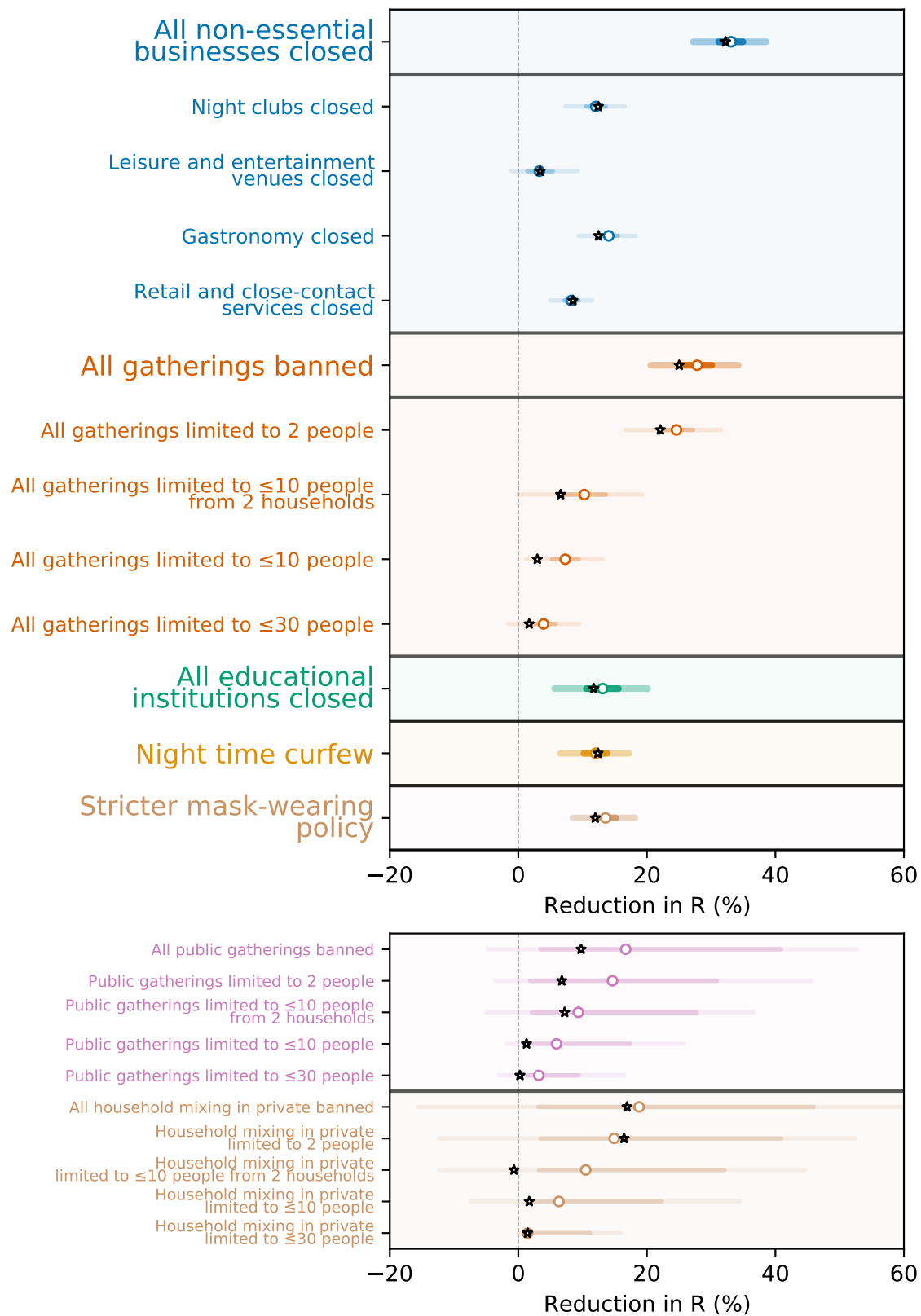
Supplementary Figure S27: Estimated effect sizes from simulated data. Black star shaped markers indicate the effect size we assigned while simulating the data. Circle shaped markers indicate posterior median estimates from 5000 posterior samples across 4 chains. Lines indicate the 50% and 95% posterior credible intervals.

2.4. Meta analysis of individual runs

To explore issues around the identifiability of effect sizes at individual local areas versus across all areas, in Supplementary Figure S28, we provide a meta-analysis of the effects estimated by running an individual model for each local area. In other words, we run our model for each location separately and aggregate the results by performing a meta-analysis. Specifically, we first estimate the effectiveness of interventions in each location. Then, we compute the median effectiveness of each intervention for each location. We then summarise the median effect sizes across the 114 individual runs by computing the median of the median effects and the 95% interval of the median effects.

We find that this approach gives NPI effects that are broadly similar to performing joint inference in all locations. Although the overall median effects for the closure of non-essential businesses and the banning of gatherings are lower in the meta-analysis, they remain the NPIs with the strongest effects, as in the main model run.

We note that the relative impact of different interventions cannot be disentangled for a single local area treated in isolation. However, when looking across multiple countries jointly, an identifiable and robust effect of the interventions can be estimated.
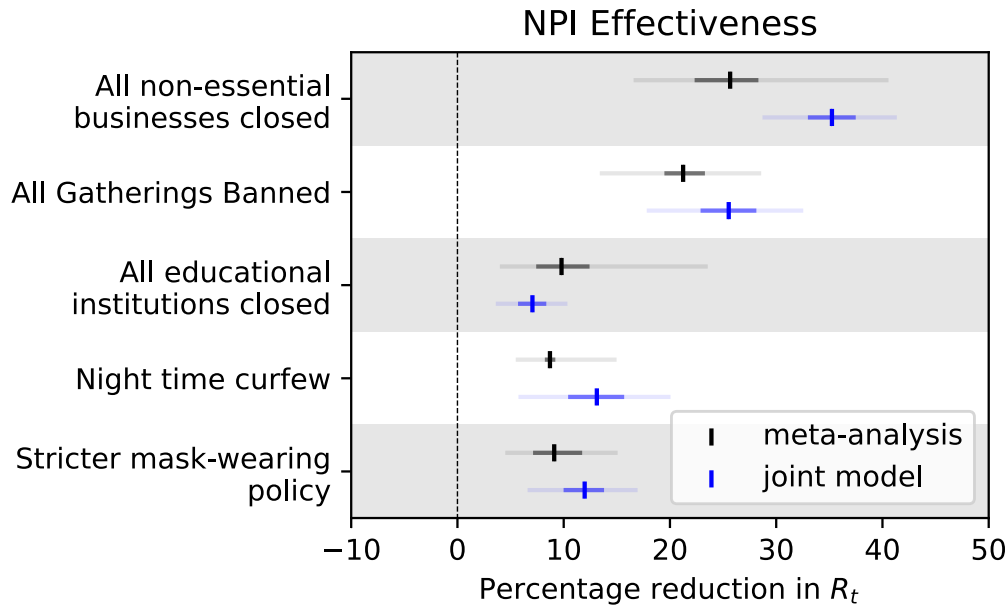
Supplementary Figure S28: Inferred intervention effect sizesn using a meta-analysis across 114 individual models; one seprate model for each local area. Percentage reductions in R_t shown. Markers indicate median estimates across n=114 regions. The bold line indicates the range from the 25th to the 75th percentile, the narrow line indicates the range from the 2.5th to the 97.5th percentile.

2.5. Prior predictive checks

Graphical prior predictive checks help us to understand whether our data generation process produces data that obeys physical constraints and matches our intuition (*23*). Since we have included a latent random walk, it is important to understand the prior predictive distribution on R_t.

Supplementary Figure S29 shows prior predictive distributions for R_t, produced using our NPI intervention timing dataset. Note that the prior on $R_{t,l}$ depends on the active interventions. Under the prior, the uncertainty in $R_{t,l}$ increases over time as noise in the latent random walk accumulates. However, since we assume that the latent random walk is equally likely to increase and decrease transmission, the *median* value of $R_{t,l}$ changes only when the active interventions change in each particular location. Since we have assumed that interventions are more likely to decrease transmission than increase transmission, under the prior distribution, we find that $R_{t,l}$ decreases when NPIs are implemented–see e.g., Abruzzo at the end of October.

These prior checks show the appropriateness of our model specification with regards to prior assumptions. The simulated values are likely to lead to a second wave. Following interventions from October, $R_{t,l}$ is suppressed to minimum median values around 0.7, matching our data-informed observed posteriors.
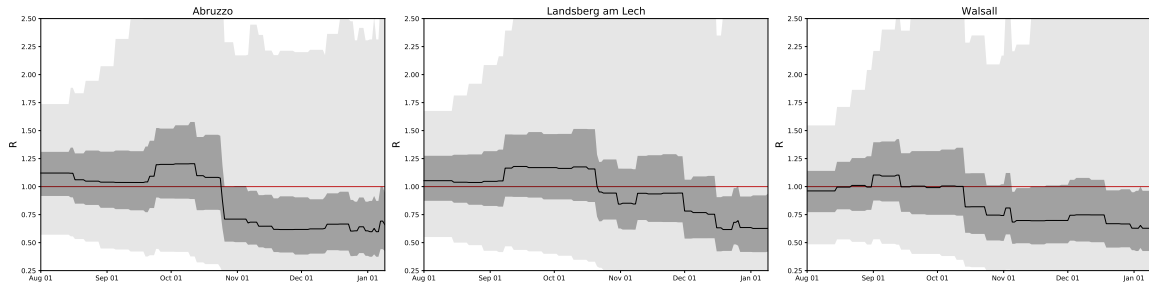
Supplementary Figure S29: Prior predictive distributions for R_t of three randomly chosen regions. Solid lines indicate prior medians while shaded areas represent 50% (dark) and 95% (light) credible intervals.

2.6. Posterior predictive distributions

The posterior predictive distribution (Supplementary Figure S30) shows the predicted number of cases and deaths after observing the data. The fit is generally tight as the latent random walk allows the model to account for location-specific changes in transmission. However, note that the NPI effect parameters are shared across all locations. Therefore, our model has posterior a predictive distribution that matches the observed data, and NPI effect estimates that apply across *all* regions.
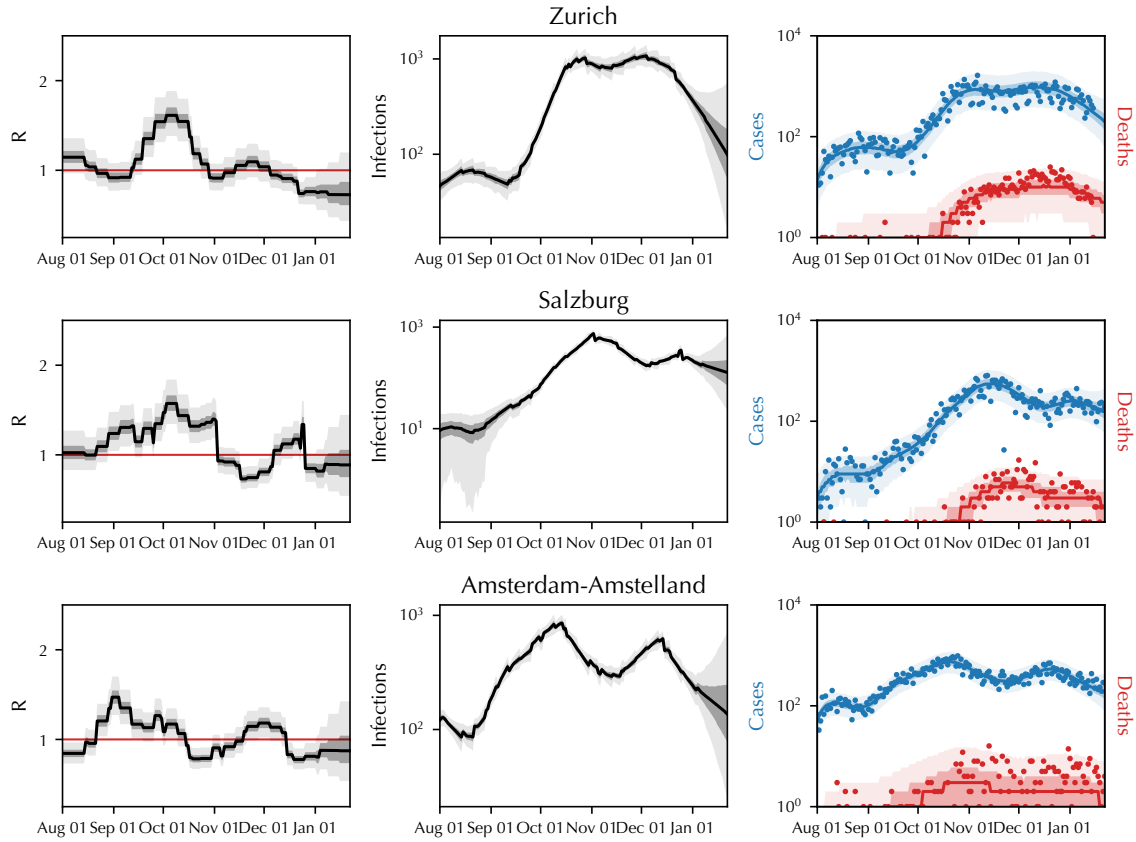
Supplementary Figure S30: *Left:* **Posterior predictive distributions for three randomly chosen regions.** *Left:* **Inferred R_t.** *Middle:* **Inferred N_t i.e., the daily number of infections.** *Right:* **Predicted and observed cases and deaths. Solid lines indicate posterior medians while shaded areas represent 50% (dark) and 95% (light) credible intervals.**

2.7. MCMC convergence statistics

Supplementary Figure S31 shows the rank-normalized split $\hat{R}$ score (*24*) and Effective Sample Size (ESS) MCMC convergence statistics.

Values of $\hat{R}$ close to 1 indicate convergence of the MCMC sampling algorithm. Vehtari et al. suggest a sample should only be used if $\hat{R} \leq 1.01$ (*24*). Therefore, Supplementary Figure S31 (left) indicates that our posterior has converged and may be used to draw inferences.

We used 250 warmup samples and 1250 iterations per chain, giving 5000 posterior samples. Inspecting Supplementary Figure S31 (right), the relative effective sample size exceeds 0.5 for the vast majority of parameters, indicating low autocorrelation and quick mixing.
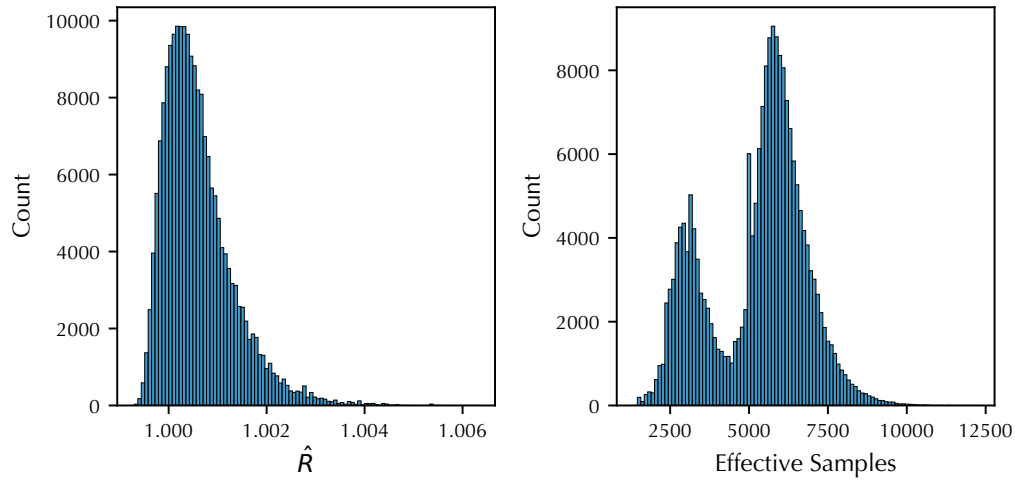
Supplementary Figure S31: MCMC convergence statistics, taken from a run using the default model with default settings and values for all parameters. *Left:* $\hat{R}$ **values are close to 1, indicating convergence.** *Right:* **Effective sample size.**

Supplementary Note 3. Additional results

3.1. Co-occurrence of NPIs

Supplementary Figure S32 shows the total number of days across all regions available to distinguish NPI effects. For every pair of NPIs (row - column), the entry shows the number of days on which only one of the NPIs was implemented (but not both or neither). Note that we do not show the traditional collinearity statistics (variance inflation factors and data correlations) since their applicability to time series data is limited. In particular, the value of these statistics in our data increases as data for a longer period becomes available, which would misleadingly suggest that we could address problems from collinearity by using less data.
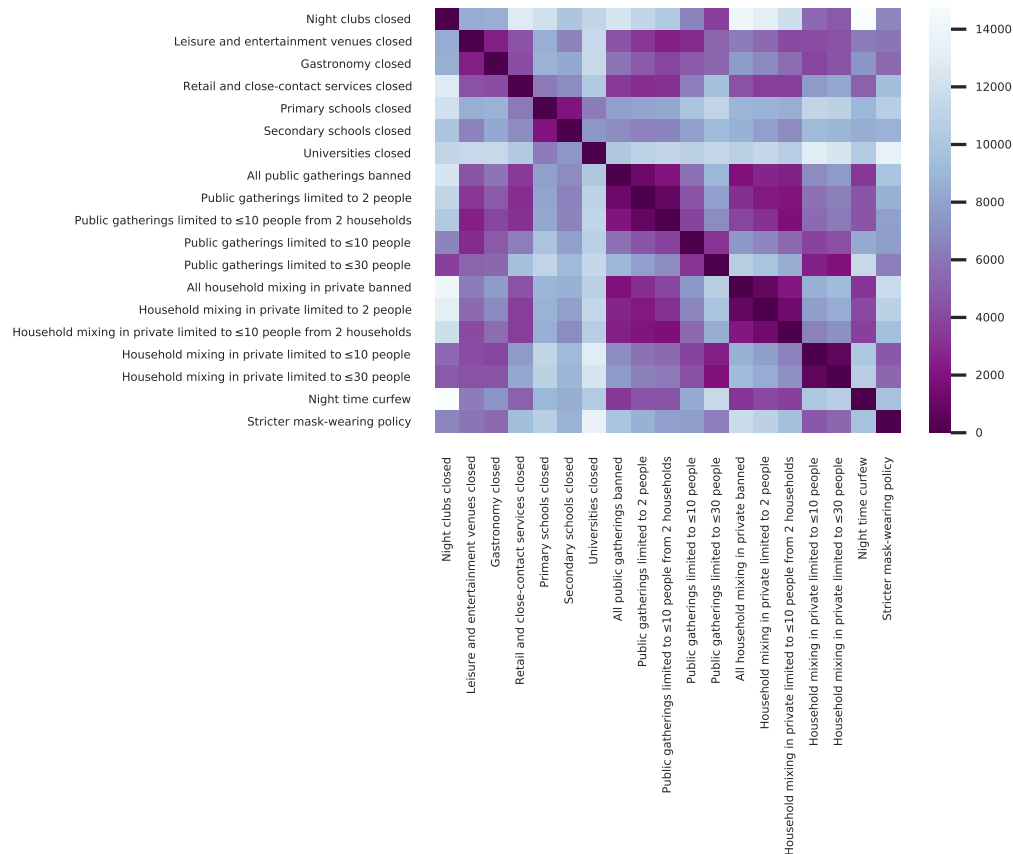
Supplementary Figure S32: Number of days across all regions available to distinguish NPI effects. For every pair of NPIs (row - column), the entry shows the number of days on which exactly one of the two NPIs was active.

51

3.2. Correlations between effectiveness estimates

The mean effect parameters β_i of NPIs that are often used together are typically negatively correlated with each other, reflecting uncertainty about which NPI reduced R. Excessive collinearity in the data would result in wide posterior credible intervals with strong correlations (*14*). We, however, generally find weak posterior correlations between effectiveness estimates with some exceptions–see Supplementary Figure S33. These weak correlations are one indicator that collinearity is manageable with our dataset. Note that we present the correlations between *marginal* effect parameters. For example, the feature household mixing in private limited to ≤10 people is made up of two marginal features: household mixing in private limited to ≤30 people and the additional effect of limiting household mixing in private to ≤10 people.

We find strong negative correlations between: Public gatherings limited to 2 people and Public gatherings limited to 2 households (-0.44); All household mixing in private banned and Household mixing in private limited to 2 people (-0.45); Household mixing in private limited to 2 households and Household mixing in private limited to 2 people (-0.54), and Household mixing in private limited to ≤10 people and Household mixing in private limited to ≤30 people (-0.42). All other correlations between pairs of NPIs have an absolute value less than 0.4.
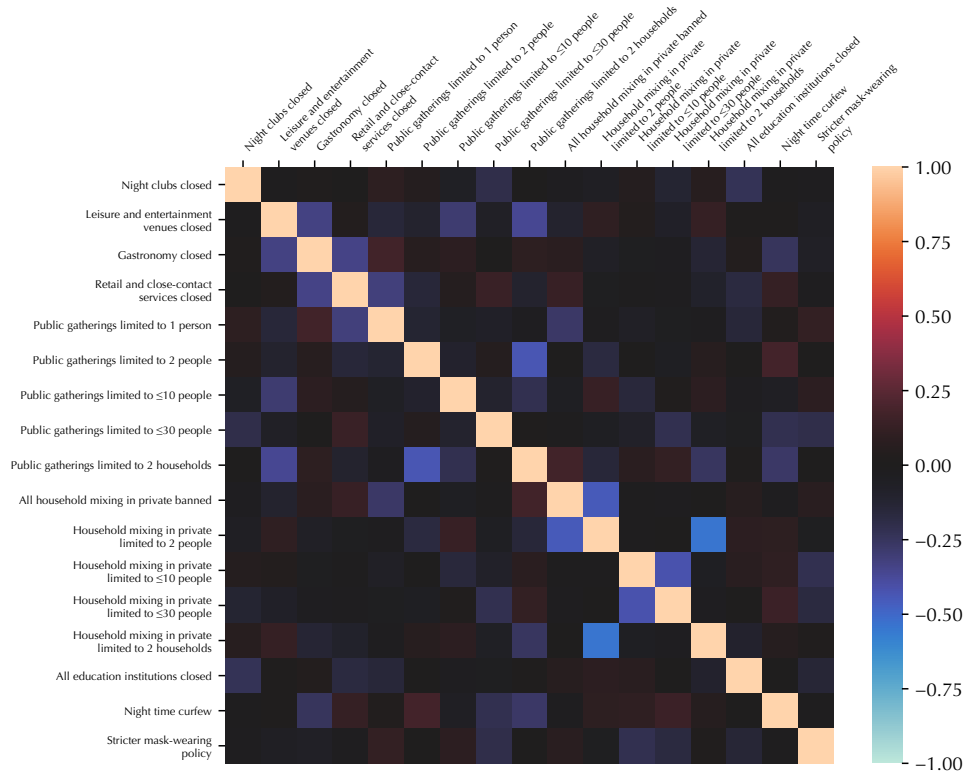
Supplementary Figure S33: Posterior correlations between effect parameters β_i.

3.3. Random walk noise scale posterior

Supplementary Figure S34 shows the posterior (using a kernel density estimate) of the random walk noise scale, σ_R. We observe that the posterior is peaked despite the wide prior, suggesting that our data is informative about this parameter.
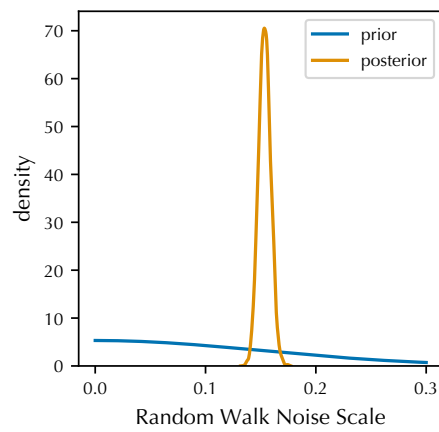
Supplementary Figure S34: Prior and posterior distributions over the random walk noise scale parameter, σ_R.

Supplementary Note 4. Data details

4.1. Feature creation

In this section, we describe the preprocessing steps used to create the NPI features used in our model from the raw data. All NPIs not mentioned below are extracted from the raw data in a straight-forward manner.

Universities closed. Some of the regions of analysis did not contain universities. For these, the "universities closed" feature was 1 ("on") throughout the period of analysis.

Restrictions on gatherings and household mixing. The dataset contains the exact number of people and households allowed at various gathering types over time. To create binary features from the limitations on gatherings, we thresholded gathering limits into "gatherings limited to ≤30 people", "gatherings limited to ≤10 people", "gatherings limited to 2 people or fewer", "all gatherings banned" (below, we call these features G30, G10, G2, and G1). These thresholds were chosen based on the histogram of gathering limits to reduce collinearity between the thresholded features. Concretely, we chose these thresholds as each of the following were active for a similar number of days across all regions:

- "gatherings with >30 people are banned, but gatherings with >10 people are allowed" (G30 is on; G10, G2, G1 are off)
- "gatherings with >10 people are banned, but gatherings with >2 people are allowed" (G30, G10 are on; G2, G1 are off)
- "gatherings with >2 people are banned, but gatherings with 2 people are allowed" (G30, G10, G2 are on; G1 is off)
- "all gatherings are banned" (G30, G10, G2, G1 are on)

The dataset records restrictions for 4 types of gatherings separately: public indoor gatherings, public outdoor gatherings, household mixing in private indoors, and household mixing in private outdoors. For both public gatherings and household mixing in private, there was very high collinearity between rules for indoors and outdoors. We thus aggregated the indoor and outdoor data into one NPI feature used for modelling. In the rare cases in which there were differences between rules for indoor and outdoor settings, we counted the rules for indoor settings, as indoor gatherings are expected to carry a higher risk of virus transmission.

Supplementary Table S3: Stringency levels of mandatory mask-policies used in the dataset. The stringency level definitions were taken from the Oxford COVID Government Response Tracker (OXCGRT) dataset (*10*) (OXCGRT codebook <u>here</u>).

Stringency level	Definition
0	No policy.
1	Recommended.
2	Required in some specified shared/public spaces outside the home with other people present, or some situations when social distancing is not possible.
3	Required in all shared/public spaces outside the home with other people present or all situations when social distancing is not possible.
4	Required outside the home at all times regardless of location or presence of other people.

For each entry, the dataset contains not only limits on the number of people at gatherings, but also any limits on the number of different households that these people were allowed to be from. The only such limit that was common across several countries was a limit of 2 households for gatherings of 6 people, and therefore this is the only household limit we include as a feature and whose effect we estimate.

Mask-wearing policy. During the data collection, we categorised mask-wearing policies into five stringency levels (Supplementary Table S3). All regions of analysis in our dataset had policies with at least a stringency level of 2 at all times during our analysis period. However, many regions switched from a level 2 policy to a policy of stringency level 3 (or higher) during our analysis period. We thus defined our mask-wearing policy feature to be 1 ("on") if the stringency level was 3 or higher, and 0 ("off") otherwise. Our inferred effect thus corresponds to the added benefit of a policy of level 3 (or higher) over a level 2 policy.

Business closures. We collected data on the closures of four major types of face-to-face businesses (night clubs and related, leisure time and entertainment venues, gastronomy, retail and close-contact services). To verify that we had not missed any major business type, we also collected an NPI called "all non-essential businesses closed". The data collectors were not informed that this NPI would be used for this validation purpose. The NPIs had the following definition: *"A country has suspended the operations of many face-to-face businesses. By default, face-to-face businesses are suspended unless they are designated as essential (whitelist). The whitelist is determined by each country and is different from country to country."* After the data collection, we evaluated the consistency between the business NPIs. Indeed,

"all non-essential businesses" (as per the dedicated NPI) are closed whenever the four specified business types are closed, and vice versa.

4.2. Data tables

Supplementary Table S4: The regions of analysis in each country and sources for daily case and death data. All links are clickable.

Country	Regions of analysis	Number of regions of analysis	Sources for case and death data
Austria	States	9 (whole country)	https://www.data.gv.at/katalog/dataset/covid-19-zeitliche-darstellung-von-daten-zu-covid19-fallen-je-bezirk/resource/9eb08d45-ff99-40f1-90cd-7b3659b0bc8d
The Czech Republic	Administrative regions	14 (whole country)	https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19 ; Spreadsheet name: COVID-19: Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu (engl: COVID-19: Overview of the epidemiological situation according to the reports of regional hygienic stations by district)
England	NUTS 3 statistical regions	15 (stratified random sample)	https://coronavirus.data.gov.uk/ ; Lookup table for conversion from LAU-1 to NUTS3 here.
Germany	Districts [including "rural districts" (Landkreise) and "urban districts" (kreisfreie Städte/Stadtkreise)]	15 (stratified random sample)	https://github.com/jgehrcke/covid-19-germany-gae
Italy	Administrative regions (With the exception of the Trentino-Alto Adige region, which was split into the autonomous provinces Trentino and Alto Adige).	21 (whole country)	https://github.com/pcm-dpc/COVID-19/blob/master/dati-regioni/dpc-covid19-ita-regioni.csv
The Netherlands	Safety regions	25 (whole country)	https://coronadashboard.government.nl/veiligheidsregio/VR13/positief-geteste-mensen
Switzerland	Cantons	15 (stratified random sample)	https://github.com/openZH/covid_19

Supplementary Table S5: Regions of analysis from each country. Local language names are given in parentheses if they differ from the English names. Abbreviations: CC: county council, RD: rural district, UD: urban district

Country	Regions of analysis
Austria	Burgenland, Carinthia (Kärnten), Lower Austria (Niederösterreich), Salzburg, Styria (Steiermark), Tyrol (Tirol), Upper Austria (Oberösterreich), Vienna (Wien), Vorarlberg
The Czech Republic	Central Bohemian (Středočeský), Hradec Králové (Královéhradecký), Karlovy Vary (Karlovarský), Liberec (Liberecký), Moravian-Silesian (Moravskoslezský), Olomouc (Olomoucký), Pardubice (Pardubický), Plzeň (Plzeňský), Prague (Hlavní město Praha), South Bohemian (Jihočeský), South Moravian (Jihomoravský), Ústí nad Labem (Ústecký), Vysočina (Vysočina), Zlín (Zlínský)
England	Brighton and Hove, Buckinghamshire CC, Coventry, East Derbyshire, Enfield, Essex Haven Gateway, Gloucestershire, Greater Manchester South West, Lincolnshire, North Yorkshire CC, Portsmouth, Redbridge and Waltham Forest, Southampton, Southend-on-Sea, Walsall
Germany	Aschaffenburg (RD), Breisgau-Hochschwarzwald (RD), Donau-Ries (RD), Ennepe-Ruhr-Kreis (RD), Enzkreis (RD), Fürth (UD), Gifhorn (RD), Hildesheim (RD), Landsberg am Lech (RD), Nuremberg (Nürnberg, SD), Minden-Lübbecke (RD), Mönchengladbach (SD), Münster (SD), Rems-Murr-Kreis (RD), Rhein-Kreis Neuss (RD)
Italy	Abruzzo, Aosta Valley (Valle d'Aosta), Apulia (Puglia), Basilicata, Calabria, Campania, Emilia-Romagna, Friuli Venezia Giulia, Lazio, Liguria, Lombardy (Lombardia), Marche, Molise, Piedmont (Piemonte), Sardinia (Sardegna), Sicily (Sicilia), South Tyrol (Alto Adige), Trentino, Tuscany (Toscana), Umbria, Veneto
The Netherlands	Amsterdam-Amstelland, Brabant-Noord, Brabant-Zuidoost, Drenthe, Flevoland, Friesland (Fryslân), Gelderland-Midden, Gelderland-Zuid, Gooi en Vechtstreek, Groningen, Haaglanden, Hollands Midden, IJsselland, Kennemerland, Limburg-Noord, Midden- en West-Brabant, Noord- en Oost-Gelderland, Noord-Holland-Noord, Rotterdam-Rijnmond, Twente, Utrecht, Zaanstreek-Waterland, Zeeland, Zuid-Holland-Zuid, Zuid-Limburg
Switzerland	Aargau, Basel-City (Basel-Stadt), Basel-Country (Basel-Landschaft), Friburg (Freiburg), Geneva (Genéve), Grisons (Grischun), Jura, Lucerne (Luzern), Neuchâtel, St. Gallen, Thurgau, Valais, Vaud, Zug, Zurich (Zürich)

Supplementary Note 5. Additional explanations and judgement calls in data collection

Throughout the data collection process, our researchers consistently applied the definitions for each of the NPIs when determining how to code each location and period. In certain cases, however, the way NPIs were defined or implemented in a country left room for some ambiguity about how to correctly code a data point. In such situations, our researchers discussed the issue internally to ensure that all similar cases were treated similarly across countries. In certain cases, where the ambiguity could not be immediately resolved, we consulted professional or personal contacts from the respective country.

This appendix outlines some common questions and edge cases that arose during the data collection process, including detailed descriptions about how all such cases were handled.

5.1. Non-mandatory recommendations

As it is very hard to consistently define various kinds of advice and non-mandatory recommendations, we only record mandatory interventions, not recommendations. Given the intractability of determining the extent to which mandatory NPIs were enforced, we did not actively research the level of enforcement. However, some regions do not have the legislative capacity to regulate gatherings in private homes. As an exception, we therefore also recorded prohibitions against private gatherings that were not legally binding. This was applicable in Italy, Austria, and the German state Northrhine-Westphalia. In each case, we confirmed with locals that these restrictions were followed by the vast majority of the population.

5.1.1. Exceptions

In Italy, due to the government's inability to introduce a mandatory limit on gatherings in private homes, a strong recommendation for a limit of 6 non-cohabiting individuals gatherings in private homes was instead issued on 13th October 2020. Private parties in bars, restaurants and other public places were also prohibited, and weddings and other ceremonies had guest limits reduced to 30. According to Prime Minister Conte, *"We can't introduce a binding rule, but the message we must send at all costs is this: do not organise private events with lots of people."* It seems they could not legislate for this rule of 6 in private homes but the strong messaging and quote seems that it would have altered public

behaviour to try and stick to this rule, at least in the most part. Also: *"The wording of the decree suggests that you can't invite more than six guests, not that there can't be more than six people in total."* (https://www.thelocal.it/20201019/covid-19-what-to-know-about-italys-new-rule-of-six) However, *"It's not unambiguous, though, and for simplicity's sake the advice has been widely described in the Italian media as 'no more than six to dinner."* Keeping this in mind, we judged that it was best to code the limit on people as 6, and this judgment was validated in conversation with an Italian epidemiologist.

5.2. Exceptions to rules on gatherings

Sometimes gathering restrictions don't apply to certain types of gatherings (e.g. not to funerals) or types of areas (e.g., parks in the UK), or apply only to certain areas (e.g. parks in Switzerland). Since our NPIs are defined in terms of "most or all" of the population/venues/events being affected, we used our best judgment to determine whether this was the case. E.g., when only funerals are exempt from a gathering ban, then certainly the limit applies to most gatherings, as funerals are quite rare. However, when all "social and cultural events" are exempt, the limit cannot be said to apply to "most gatherings", and we would not record it.

5.2.1. Examples

In Austria, the state Burgenland seems to have had stricter measures than the national measures between 2020/10/23 and 2020/11/02 (national measures were significantly tightened on 2020/11/03). However, these seem to have only applied to sporting events. (See the State ordinance, archived corona-ampel.gv.at, the press release on the State authority's website.) Since the measures only applied to sporting events, not to all or most events, we did not record a state-level deviation in any NPIs related to this.

In England, from the 3rd of July onwards, gatherings were only allowed for up to thirty people. There were quite a few exceptions for public gatherings. Exceptions (note: all three of the bullet points below must be satisfied):

- (i)the gathering has been organised by a business, a charitable, benevolent or philanthropic institution, a public body, or a political body,
- (ii)the person responsible for organising the gathering ("the gathering organiser") has carried out a risk assessment which would satisfy the requirements of regulation

3 of the Management of Health and Safety at Work Regulations 1999(1), whether or not the gathering organiser is subject to those Regulations, **and**

- (iii)the gathering organiser has taken all reasonable measures to limit the risk of transmission of the coronavirus, taking into account the risk assessment carried out under paragraph (ii).

These exceptions would have to be approved by the government and were granted larger gatherings with the numbers decided on a case-by-case basis. As most public gatherings did not warrant applying for an exception, we recorded this as a maximum of 30 for both public and private outdoor spaces.

In Enzkreis, Baden-Württemberg, Germany: on 24/10/2020 gatherings and private events were limited to 5 individuals except in cases of meetings with family members or just one other household. This was coded as a public/private as well as indoor/outdoor restriction on gatherings of 5 individuals, with no household limit.

In Germany, North Rhine Westphalia had a long list of restrictions for meetings. Meeting with $\sharp$10 people was banned, but events with safety restrictions (including either fixed seating or 1.5m distance) could get around this. There was a requirement to apply for permission for events with $\sharp$300, to verify that hygiene protocol was being followed. The rule stated:

- **§ 1, paragraph 2: Several people may only meet in public space if they are**
  1. exclusively to relatives in a straight line, siblings, spouses, life partners, 2. exclusively about people from a maximum of two different domestic communities, 3. to accompany minors and persons in need of support, 4. for absolutely necessary meetings for reasons relevant to care or 5. in all other cases by **a group of no more than ten people**. ... Other gatherings and get-togethers of people in public spaces are not permitted until further notice; **with exception of**:
- 1. unavoidable accumulations during the intended use of permitted facilities (in particular when using passenger transport services and its facilities),
- 2. **Participation in events and meetings permitted under this Ordinance**,
- 3. Permissible sporting activities and permissible youth work and youth social work,
- 4. Compulsory meetings for professional practice in public spaces. ...
- **At events and meetings** that do not fall under the special provisions of this ordinance, **suitable precautions must be taken to ensure hygiene, to control access, to ensure a minimum distance of 1.5 meters (also in queues) between people who do not belong to the requirements specified in § 1, paragraph 2, and, if necessary, to implement**

an obligation to wear a face-to-face mask (Section 2, paragraph 3). Apart from outdoors, easy traceability must also be ensured in accordance with Section 2a (1). **If the participants sit in fixed places during the event or meeting, the requirement of a minimum distance of 1.5 meters between people can be replaced by ensuring the special traceability** according to § 2a paragraph 2.

- (2) **Events and assemblies that do not fall under the special provisions of this ordinance with more than 300 participants require a special hygiene and infection protection concept** in accordance with Section 2b, which at least ensures the requirements of Paragraph 1. ...
- (4) Notwithstanding paragraphs 1 and 2, **large festive events are prohibited** ...
- (5) Paragraphs 1 and 2 do not apply to **parties** (events with a primarily social character). These **are only permitted for an outstanding occasion (e.g. anniversary, wedding, christening, birthday, graduation party) and with a maximum of 150 participants. The distance requirement and an obligation to wear a face-to-face mask do not apply** as long as suitable precautions for hygiene and easy traceability are ensured

We classified this as a 10 person limit on public gatherings and a 150 person limit on private gatherings (as nearly all private events with 150 people will be due to an "outstanding occasion" and so nearly no private events with ∭150 people were actually banned). As the 150 person limit was reduced, we reduced the limit on private gatherings accordingly. When similar situations came up in other regions, we followed a similar solution.

In the Czech Republic, a national [measure](link) says basically that ~all events and gatherings are limited to 1000 people outdoors and 500 indoors, except for ~events in "structurally partitioned facilities." The full quote is *"this prohibition does not apply to... large events held in structurally partitioned facilities (sports stadiums, exhibition grounds, etc.) with the presence of more than 1,000 people at any one time if the event is held largely outdoors, or more than 500 people in each of a structurally separate section (by means of a mobile barrier, e.g. portable fence) of a facility if the event is largely held indoors, provided the entire facility is divided into a maximum of five structurally partitioned sections, each having access from outdoors and the attendees being unable to move between the sections; the distance between people in adjacent sections must be at least four metres; in premises with seating (e.g. sports stadium stands, etc.), attendees of the large event must be seated in only every second row and must be separated from other attendees in these rows by at least one empty seat"* and, also excluded are *"meetings, conventions and similar events held by constitutional bodies, public authorities, courts and other public entities, which are held by law,"* and *"gatherings pursuant to Act No. 84/1990 Coll., as amended, on*

the Right to Gather." While this was a borderline case, we concluded that "most" events above that size were prohibited (and the bigger events that were allowed to occur with barriers could be considered multiple events of only 1000/500 (outdoor/indoor) people), and recorded maximum gathering sizes of 1000/500/1000/500 for public outdoor/public indoor/private outdoor/private indoor gatherings, respectively.

In Switzerland, until October, there was a limit on the total number of attendees at an event as well as a limit on divisions ("Sektoren") within events where it might not be possible to social distance for the entire event. We used these division limits as gathering restrictions in the absence of other more specific restrictions.

5.2.2. Exceptions

In Austria, the [Government advice page](#) states that private Christmas parties of 10 people from 10 different households were allowed on December 24 and 25. [The regulation attached to this advice](#) allows meetings of 10 people from 10 different households. Therefore, this was taken to mean that public outdoor/indoor and private outdoor events were limited to 10 people on the 24th and 25th of December.

In Germany, over Christmas (24/12-26/12), one was allowed to meet up to four family members regardless of the number of households. This was coded as a limit of 5 with no household restriction (5,no).

5.3. Special arrangements for educational institutions

In some cases, special teaching arrangements (e.g., online teaching or an option to keep children home from school) made it ambiguous whether schools should be recorded as open or closed. As a general rule, we aimed to determine whether it was the case that "most or all" (i.e., $\S 50\%$) of children at a given educational level (i.e., primary, secondary, or university) were in school. Below is a list of examples from various countries and explanations of our decision criteria.

5.3.1. Examples

In Lower Saxony, Germany, parents were allowed to keep their children home from the 14th to the 18th (the last week of school before holidays). This was coded as no closure on schools.

In North Rhine Westphalia, Germany, during the last week before christmas, parents of school children in grades 1-7 (ages 7-13) could choose to keep their children home, while grades 8-13 (ages 14-19) had to do online schooling. We coded this as a restriction on secondary schools but not on primary schools, after consulting an online document which records what fraction of pupils actually attended school `https://www.schulministerium.nrw.de/system/files/media/document/file/Zeitreihe-Kreise-KW51.pdf`.

In Minden-Lubbecke, Germany, the largest University started its normal face-to-face courses on the 2nd November. However, they hint that there might have been mandatory face-to-face introductory events earlier. In addition, a few of the courses involved working for part of the year at companies and followed a different schedule. (See here: `https://www.fh-bielefeld.de/minden/informationen-fuer-studierende-und-erstsemester`). Consequently, the end of autumn vacation was coded as 1st November.

In Furth, Germany, a small university's "exam time" extended into August (see [here](#)). However, this was counted as part of the summer holidays, as it is not clear how many students had exams in August.

In Portsmouth, UK, the University of Portsmouth had a "Consolidation Week" during term. While classes were not in session, students most likely remained on campus and continued using the facilities (to the extent that they had been doing this before). Consequently, we did not consider this as a closure of the university.

In Portsmouth, UK, the University of Portsmouth required international students to commence one week before the rest of the students. However, as it seems like internationals make up just ~4000 of the 25,000 students, we used the date when most (i.e., including nationals) students began term.

In the Czech Republic, national rules for primary schools involved a long period where 1st and 2nd grade pupils (1st and 2nd years of primary school) were going into school, and the older grades (3rd, 4th, and 5th years of primary school) were not. As attendance was roughly 40% (2 of 5 years attending), we did not record an NPI for such periods.

Similarly, in the Czech Republic, national rules for secondary schools involved periods where grades 10-13 (the last 4 grades of secondary school) were closed, and grades 6-9 (the first 4 grades) went into school on alternate weeks (i.e., if they followed the rules exactly, you would expect 50% attendance). Since about 4/8 * 0.5 = 25% of secondary school students were attending school at any one time, we counted an NPI for periods where grades 10-13

were closed and grades 6-9 attended on alternate weeks, as most students were likely not in attendance.

In Austria, the school situation was somewhat complicated. There were plenty of provisions for students to go into school after the schools had been "closed" or switched to "distance learning". The key periods are: Lockdown Light 1 (2020/11/03-2020/11/16), Lockdown Light 2 (2020/12/07-2020/12/23), Strict Lockdown 1 (2020/11/17-2020/12/06), and Strict Lockdown 2 (2020/12/26-post our data window). During the 'light' lockdowns, only ~"higher schools" were distance learning, which roughly corresponds to Sekundarstufe II on this page. During the strict lockdowns, all schools were distance learning. We found two relevant sets of data regarding school attendance: a comprehensive survey on 2020/11/17 as reported here, and another survey on 2020/11/27 reported here. The 2020/11/17 survey claims an average attendance of 15% across all school years nationwide, ~23% (inferred) average attendance in primary schools nationwide, and 25% attendance across all school years for regions with the highest attendance. The 2020/11/27 survey has similar numbers. It claims that the all-age nationwide average attendance rate increased by 2%, from 15% to 17%. Based on this, we infer that in Strict Lockdown 1 our secondary school NPI was in effect in all regions, while our primary school NPI was more borderline: Oberoesterreich and Niederoesterreich had an overall attendance of 25%, and the numbers from the other regions suggest that this would have been concentrated in primary schools. We estimated that there was ~32% attendance in primary schools in Oberoesterreich and Niederoesterreich. Based on this, we decided to code the primary school NPI as active, since assuming any specific regional variation would be based on speculation and because we expect the regional variation to be small (all regions had attendance fairly close to 30%). For Strict Lockdown 2, we did the same as for Strict Lockdown 1. We did this because the rules were similar and because this lockdown was only relevant for a period of 3 days, since schools were on holiday until 2021/01/07. For the light lockdowns, it was clear that we should not record an NPI for primary schools, as they operated as normal. We did not record an NPI for secondary schools either, as Sekundarstufe II was distance learning and the rest of secondary school-age students, Sekundarstufe I, were going to school as normal (Sekundarstufe II covers 5 grades, Sekundarstufe I covers 4 grades. If we assume that all of the Sekundarstufe I students went into school, all of the Sekundarstufe II students stayed at home, and that all year groups had the same number of students, then 4/9 = 44% of students attended school). In summary, we coded both school NPIs as active for the strict lockdowns and inactive for the light lockdowns and any other non-holiday periods.

5.4. Distinctions between types of gatherings (indoor/outdoor, private/public, number of households)

In some cases, countries did not clearly distinguish between indoor/outdoor and private/public gatherings. In others, their rules did not map neatly onto our distinctions. We evaluated such situations on a case-by-case basis while striving to ensure consistent handling across all countries and cases.

5.4.1. Examples

In England, public indoor gatherings are not defined distinctly from private indoor gatherings in the legislature. The UK seems to use "indoors" for all categories but has clear exceptions where things are allowed (prison, education, etc). While some public buildings were open in the second lockdown, it does not appear that people were able/supposed to organise gatherings in them. This was vague in the sense that there was leeway for students to meet and study in libraries together in groups, and it was not illegal for families to go to shopping malls together. We concluded that legislation on "indoor gatherings" applies to both public and private. *Ref:* (pages 9-16) `https://www.legislation.gov.uk/uksi/2020/1611/pdfs/uksi_20201611_en.pdf`.

Regarding household limits in England prior to the introduction of the tier system on the 14th of October: this (as well as a key amendment that introduced the rule of six) was the key legislation that affected gatherings. Neither legislation mentions anything about a maximum number of households. However, some sources had vague messages about government guidance (e.g., a government news release that explained the rule of six stated this: *The rule applies across England and replaces the existing ban on participating in gatherings of more than 30 and the current guidance on allowing two households to meet indoors.*"). Upon further investigation, it appears that government guidance and legislation differed:

- (published on September 14) *"Since July, the Government guidance has stated that no more than six people should meet outside, while only two households are allowed to meet indoors. However, this was different from the actual law, which stipulated that any gathering over 30 people was illegal and could be broken up by police."*
- (A BBC article referencing the same thing, published September 9) *"At present, the guidance says two households of any size are allowed to meet indoors or outdoors, or up to six people from different households outdoors. Until now the police have had no powers to*

stop gatherings unless they exceeded 30 [referring to the amendment which instated the rule of six]."

From this, it seems that the Government was encouraging a limit on households and meetings indoors/outdoors, but this limit was not mandatory. The guidance would dictate that we code "any number of people from up to two households (no,2) for all indoor and outdoor gatherings", and "six people from any number of households (6,no) for outdoor gatherings", but the legislation dictates that we only code (30,no). As the guidance was not mandatory, we coded in accordance with the legislation: (30,no).

In North Rhine Westphalia, Germany, there were never any laws about meeting inside homes. However, there was a strong nationwide policy that all gatherings, including those inside homes, should be limited to a maximum of 5 people and a maximum of 2 households. Consequently, this was coded as a (5,2) restriction on private outdoor and indoor spaces.

In Italy, a private indoor and outdoor gathering restriction was introduced on December 24 where 2 people were allowed to travel to another home to meet other people. Quote: *'It is possible to visit one other private residence within the same region between 5am and 10pm, no more than once a day. Curfew must be respected. A maximum of two persons may make such journeys, plus any children in their household under the age of 14.'* Assuming that the average household has 2 people, we recorded the limit at 4.

In Switzerland, in the absence of a private gathering limit, there was a private event limit of 100 for events with masks aside from when seated and eating, and a limit of 15 for events without masks. We coded this as a limit of 100 people, as we considered mask-wearing at such private events (especially where eating and drinking happens) unlikely to be strictly enforced.

In Bavaria, Germany, between August and October, one was allowed to meet up to 10 individuals or individuals from only one further household in public places. This was coded as up to 10 with no household restriction (10,no).

Regarding private gathering sizes during stay at home orders in the Czech Republic: there were periods where people were ordered to stay at home except for essential journeys. There were meeting limits for public places (at some times the limit was 2, and at others it was 6). There were limits for private places. On the face of it, you might think that meeting in private was prohibited except for essential reasons. However, according to a Czech national we spoke to, people took the private gathering limit to be the same as the

public counterpart. And public statements from ministers were also quite confusing (e.g. in this article a spokesperson for the Ministry of Health seems to say that private meetings have no size limit). We recorded the private gathering limit to be the same as the public one in these cases.

5.5. Other edge cases

There was a range of special cases that did not neatly fall into any of the above categories. As before, we evaluated each of these situations on a case-by-case basis, while striving to ensure consistent handling across all countries and cases.

5.5.1. Examples

In Switzerland, cantonal authorities have autonomy to make decisions that overrule federal law. However, during the pandemic, they largely followed federal law, thus we deferred to federal law in cases where cantonal law did not state otherwise. Note that the current federal laws state: *"The following overview shows the rules and bans that currently apply nationwide. In other words, these measures, at the very minimum, apply throughout Switzerland. Stricter measures may apply in certain cantons. You can find information about these in the section entitled Cantonal measures"*. Consequently, we only deferred to cantonal law where it was (a) stricter than federal law (b) explicitly stated in press releases that it would not be following federal law.

In Italy, while traffic light zones were mostly implemented by region, there was an unusual situation in Trentino-Alto Adige/Südtirol. This is because *'The region is divided into two autonomous provinces: Trentino (Autonomous Province of Trento) and South Tyrol (Autonomous Province of Bolzano)'*. Both of these have fairly similar population sizes of ~0.5m. There were cases where Bolzano was in the red or orange zone while Trento was in the yellow zone, resulting in many more NPIs being active in Bolzano compared to Trento in the same period. We contacted an Italian epidemiologist who agreed with this assessment, but who noted that *"the only exception is Trentino and Alto-Adige which are one region (Trentino Alto-Adige), but in the data are often reported as separate provinces. It is best to keep these two separate because measures are applied at the provincial level here."* Consequently, we coded these areas by the two provinces. In Baden-Wurttemberg, stay-at-home-orders applied from the 12th December. When they applied, public gatherings were generally forbidden, with the main exception being that you were allowed to exercise in the fresh air with your own household or with a single person from another household (this is very

clear in [the orders that applied from the 16th](#)). We coded this as allowing public outdoor gatherings of 2 people, and public indoor gatherings of only 1.

5.6. Sub-area NPI implementation

Sometimes, decisions about NPIs might be made on a level even more fine-grained than the local areas for which we collect data. For example, if a local area contains two towns, it might be the case that only one of is a hotspot and ends up introducing certain NPIs (although the other may follow suit later). In these situations, we only recorded the NPI as active if and when they affected the whole local area.

5.6.1. Examples

For example, in the area of Essex Haven Gateway, the local administrative unit of Braintree changed tiers on different dates than the two other administrative units in that area. In this case, we only coded changes in NPIs when all of Essex Haven Gateway had changed tiers.

5.7. Mandatory safety measures

In some cases, it became required to implement certain safety measures for events and gatherings to occur. E.g., in Germany, cinemas operated under safety restrictions such as only allowing 25% capacity; in Italy, it was prohibited to loiter outside bars. In such cases, we did not record the safety measures and only considered actual closures and prohibitions, as nearly everything had some kind of safety measure.

5.7.1. Examples

In Austria, national rules for the maximum number of people at public events over the summer of 2020 allowed events to apply for permits to increase their capacity. For example, the default maximum number of people as of 2020/08/01 for public outdoor events with fixed seating was 750, and a permit could be granted to increase this to 1250. (See [the national rules saved on web.archive.org on 2020/08/02](#).) Since "most or all" events were not constrained by the without-special-permission numbers (our impression is that a permit was not especially difficult to obtain), we coded the higher numbers that were allowed with special permission.

Similarly, in Austria, national rules for the maximum number of people allowed at public events over the summer of 2020 depended on whether the events used fixed seating. (E.g. see the national rules saved on web.archive.org on 2020/08/02.) Again, we used the higher, fixed seating numbers because it is not the case that "most or all" events bigger than the non-fixed-seating limit were constrained by the non-fixed-seating limit (i.e. they probably mostly had fixed seating rather than being cancelled).

In Lower Saxony, Germany, there was a rule that you could meet 10 people without any safety precautions, and as long as you keep a 1.5m distance, private gatherings and celebrations could consist of: 25 people in your home, 50 people in your garden, 100 people in any publicly accessible location *"including in rooms provided outside of one's own home and in catering establishments."* Additionally, for purely seated non-private events (e.g. cinemas), and for at least temporarily standing events with a safety concept (e.g. congresses), you could have up to 500 visitors. (Some of these numbers were later lowered due to high incidence.) Since we expect people to comply with the safety precautions in public spaces, public outdoor and indoor were coded as 10 people. Private spaces were coded as 100, since rented venues are counted as private.

5.8. The timing of implementation

We record the date when an NPI is expected to have affected behaviour. If an NPI is announced to be in place immediately, the starting day is the day of the announcement, unless it was announced in the evening and clearly only in effect on the next day. One exception to this rule is nightclubs, which we counted as the day of (as most nightclub activity happens at night).

For schools, we considered the fact that such institutions are closed on weekends. Consequently, if it is announced on Thursday that schools will stay closed starting from Monday, the actual date when schools were closed is Saturday, and that is the date we recorded.

5.8.1. Examples

In Switzerland, since some laws came into place at 5pm on a certain day, we counted the next day as the first day of restriction in all of these cases.

References

1. YouGov, Personal measures taken to avoid COVID-19, `https://yougov.co.uk/topics/international/articles-reports/2020/03/17/personal-measures-taken-avoid-covid-19` (2020). Accessed: 2021-5-20.

2. COVID-19 consumer sentiment tracker, `https://www.visitbritain.org/covid-19-consumer-sentiment-tracker` (2020). Accessed: 2021-5-24.

3. HOPE dashboard, `https://hope-project.dk/dashboard/`. Accessed: 2021-5-24.

4. R. Laxminarayan, *et al.*, Epidemiology and transmission dynamics of COVID-19 in two indian states. *Science* **370**, 691–697 (2020).

5. J. E. Lemieux, *et al.*, Phylogenetic analysis of SARS-CoV-2 in boston highlights the impact of superspreading events. *Science* **371** (2021).

6. D. Lewis, COVID-19 rarely spreads through surfaces. so why are we still deep cleaning? *Nature* **590**, 26–28 (2021).

7. Federal Office of Public Health FOPH, Coronavirus: Measures and ordinances, `https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/massnahmen-des-bundes.html`. Accessed: 2021-5-20.

8. Fedlex, `https://www.fedlex.admin.ch/eli/cc/2020/439/en`. Accessed: 2021-5-20.

9. J. M. Brauner, *et al.*, Inferring the effectiveness of government interventions against covid-19. *Science* (2020).

10. T. Hale, *et al.*, Oxford COVID-19 Government Response Tracker, Blavatnik School of Government (2020).

11. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* **20**, 533–534 (2020). doi:10.1016/s1473-3099(20)30120-1.

12. M. Sharma, *et al.*, How robust are the estimated effects of nonpharmaceutical interventions against COVID-19? arXiv:2007.13454 [stat.AP] (27 July 2020).

13. N. Banholzer, *et al.*, Impact of non-pharmaceutical interventions on documented cases of COVID-19. MedRxiv 2020.04.16.20062141v3 [Preprint]. 28 April 2020, doi:10.1101/2020.04.16.20062141.

14. C. F. Dormann, *et al.*, Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27–46 (2013). doi:10.1111/j.1600-0587.2012.07348.x.

15. S. Flaxman, *et al.*, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020). doi:10.1038/s41586-020-2405-7.

16. Y. Dong, *et al.*, Epidemiology of covid-19 among children in china. *Pediatrics* **145** (2020).

17. A. Gelman, J. Hill, Causal inference using regression on the treatment variable in *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, 2007), 167–198. doi:10.1017/CBO9780511790942.012.

18. P. R. Rosenbaum, 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**, 212–218 (1983). doi:10.1111/j.2517-6161.1983.tb01242.x.

19. J. M. Robins, A. Rotnitzky, D. O. Scharfstein, *Statistical models in epidemiology, the environment, and clinical trials* (Springer New York, 2000), 1–94. doi:10.1007/978-1-4612-1284-3_1.

20. S. Mishra, *et al.*, A COVID-19 Model for Local Authorities of the United Kingdom. *medRxiv* 2020.11.24.20236661 (2020).

21. Coronavirus (COVID-19) Infection Survey - Office for National Statistics (2020). [Online; accessed 21. Nov. 2020].

22. Real-time Assessment of Community Transmission findings (2020). [Online; accessed 21. Nov. 2020].

23. J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, A. Gelman, Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **182**, 389–402 (2019).

24. A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, P.-C. Bürkner, Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *Bayesian Analysis* (2020). doi:10.1214/20-ba1221.