Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (2019-nCoV) outbreak

Sang Woo Park[1,*] Benjamin M. Bolker[2,3,4] David Champredon[5] David J. D. Earn[3,4] Michael Li[2] Joshua S. Weitz[6, 7] Bryan T. Grenfell[1,8,9] Jonathan Dushoff[2,3,4,*]

1 Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA
2 Department of Biology, McMaster University, Hamilton, ON, Canada
3 Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada
4 M. G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada
5 Department of Pathology and Laboratory Medicine, University of Western Ontario, London, Ontario, Canada
6 School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
7 School of Physics, Georgia Institute of Technology, Atlanta, GA, USA
8 Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA
9 Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ, USA

*Corresponding authors: swp2@princeton.edu and dushoff@mcmaster.ca

Abstract

Background

A novel coronavirus (2019-nCoV) has recently emerged as a global threat. As the epidemic progresses, many disease modelers have focused on estimating the basic reproductive number $\mathcal{R}_0$, the average number of secondary cases caused by a primary case in an otherwise susceptible population. The modeling approaches and resulting estimates of $\mathcal{R}_0$ vary widely, despite relying on similar data sources.

Aim

We aimed to develop a framework for comparing and combining different estimates of $\mathcal{R}_0$ across a wide range of models.

Methods

We reviewed 7 model-based analyses of the 2019-nCoV outbreak that were published online between January 23–26, 2020. We decompose their $\mathcal{R}_0$ estimates into three key quantities: the exponential growth rate r, the mean generation interval $\bar{G}$, and the generation-interval dispersion κ. We use a Bayesian multilevel model to construct pooled estimates and measure uncertainties associated with these quantities.

Results

We find that most early estimates of $\mathcal{R}_0$ rely on strong assumptions, especially about the generation-interval dispersion. Estimates that rely on narrow generation-interval distributions are overly sensitive to estimates of the exponential growth rate.

Conclusion

Our results emphasize the importance of propagating uncertainties in all components of $\mathcal{R}_0$, including the shape of the generation-interval distribution, in efforts to estimate $\mathcal{R}_0$ at the outset of an epidemic.

Keywords

Basic reproductive number, 2019-nCoV, novel coronavirus, Bayesian multilevel model

Funding

BMB and DJDE were supported by Natural Sciences and Engineering Research Council (NSERC). ML was supported by Canadian Institutes of Health Research (CIHR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of interests

We declare no competing interests.

Acknowledgements

We thank Daihai He for providing helpful comments on the manuscript.

Contribution

SWP and JD developed the statistical framework. SWP reviewed the published literature. SWP performed the analysis. SWP, BMB, and JD created the figures. SWP and JD wrote the first draft. All authors contributed to the writing and approval of the final report.

Data availability

`R` code is available in GitHub (`https://github.com/parksw3/nCoV_framework`).

1 Introduction

Since December 2019 [1], a novel coronavirus (2019-nCoV) has been spreading in China and other parts of the world. Although the virus is believed to have originated from animal reservoirs, the ability of 2019-nCoV to directly transmit between humans has posed a greater threat for its spread [2]. As of February 3, 2020, the World Health Organization (WHO) has confirmed 17391 cases, including 153 confirmed cases in 23 different countries, outside China [3].

As the disease continues to spread, many researchers have published their preliminary analyses of the outbreak as pre-prints [4–10], focusing in particular on estimates of the basic reproductive number $\mathcal{R}_0$ (i.e., the average number of secondary cases generated by a primary case in a fully susceptible population [11]). Estimates of the basic reproductive number are of interest during an outbreak because they provide information about the level of intervention required to interrupt transmission, and about the final size of the outbreak [11]. We commend these researchers for their timely contribution and those who made the data publicly available. However, it can be difficult to compare a disparate set of estimates of $\mathcal{R}_0$ from different research groups when the estimation methods and their underlying assumptions vary widely.

Here, we show that a wide range of approaches to estimating $\mathcal{R}_0$ can be understood and compared in terms of estimates of three quantities: the exponential growth rate r, the mean generation interval $\bar{G}$, and the generation-interval dispersion κ. The generation interval, defined as the interval between the time when an individual becomes infected and the time when that individual infects another individual [12], plays a key role in shaping the relationship between r and $\mathcal{R}_0$ [13]; therefore, estimates of $\mathcal{R}_0$ from different models directly depend on their implicit assumptions about the generation-interval distribution and the exponential growth rate. Early in an epidemic, information is scarce and there is inevitably a great deal of uncertainty surrounding disease transmission. We suggest that disease modelers should make sure their assumptions about these three quantities are clear and reasonable, and that estimates of uncertainty in $\mathcal{R}_0$ should propagate error from all three sources [14].

We compare seven disparate models published online between January 23–26, 2020 that estimated $\mathcal{R}_0$ for the 2019-nCoV outbreak [4–10]. We use our framework to construct pooled estimates for the three key quantities: r, $\bar{G}$, and κ; the pooled estimates reflect the uncertainties present in modeling approaches. We use these pooled estimates to illustrate the importance of propagating different sources of error, particularly uncertainty in both the growth rate and the generation interval. We also use our framework to tease apart which assumptions of these different models led to their different estimates and confidence intervals. Despite the availability of more recent and/or updated estimates of $\mathcal{R}_0$, we restrict ourselves to the estimates above in order to focus on the resolution of uncertainty in the earliest stages of an epidemic.

	Basic reproductive number $\mathcal{R}_0$	Mean generation interval $\bar{G}$ (days)	Generation-interval dispersion κ	
Study 1	1.5–3.5	10	1	Bedford *et al.* [4]
Study 2	2.5 (1.5–3.5)*	8.4	unspecified†	Imai *et al.* [5]
Study 3	2.92 (95% CI: 2.28–3.67)	8.4	0.2	Liu *et al.* [6]
Study 4	3.8 (95% CI: 3.6–4.0)	7.6	0.5	Read *et al.* [8]
Study 5	2.2 (90% CI: 1.4–3.8)	7–14	0.5	Riou and Althaus [10]
Study 6	5.47 (95% CI: 4.16–7.10)‡	7.6–8.4	0.2	Zhao *et al.* [9]
Study 7	2.0–3.1	6–10	0	Majumder and Mandl [7]

Table 1: **Reported estimates of the basic reproductive number and the assumptions about the generation-interval distributions.** Estimates of $\mathcal{R}_0$ and their assumptions about the shape of the generation interval distributions were collected from 7 studies. *We treat these intervals as a 95% confidence interval in our analysis. †We assume $\kappa = 0.5$ in our analysis. ‡The authors presented $\mathcal{R}_0$ estimates under different assumptions regarding the reporting rate; we use their baseline scenario in our analysis to remain consistent with other studies, which do not account for changes in the reporting rate.

2 Methods

2.1 Description of the studies

We gathered information on estimates of $\mathcal{R}_0$ and their assumptions about the underlying generation-interval distributions from 7 articles that were published online between January 23–26, 2020 (Table 1). Five studies [6–10] were uploaded to pre-print servers (bioRxiv, medRxiv, and SSRN); one report was posted on the web site of Imperial College London [5]; and one report was posted on nextstrain.org [4]. Their modeling approaches vary widely: a branching process model [4, 5, 10], a deterministic Susceptible-Exposed-Infected-Recovered (SEIR) model [8], an exponential growth model [9], a Poisson offspring distribution model [6], and the Incidence Decay and Exponential Adjustment (IDEA) model [7]. Four studies estimated $\mathcal{R}_0$ by directly fitting their models to incidence data [6–9]. The remaining three studies estimated $\mathcal{R}_0$ by comparing the predicted number of cases from their models with the estimated number of total cases by January 18 (between 1,000 and 9,7000 [15]) Some of these studies have now been published in peer-reviewed journals [16, 17] or have been updated with better uncertainty quantification [18].

2.2 Gamma approximation framework for linking r and $\mathcal{R}_0$

Early in an outbreak, $\mathcal{R}_0$ is difficult to estimate directly; instead, $\mathcal{R}_0$ is often inferred from the exponential growth rate r, which can be estimated reliably from incidence data [19]. Given an estimate of the exponential growth rate r and an *intrinsic* generation-interval distribution $g(\tau)$ [20], the basic reproductive number can be estimated via the Euler-Lotka equation [13]:

$$1/\mathcal{R}_0 = \int \exp(-r\tau) g(\tau)\, d\tau. \tag{1}$$

In other words, estimates of $\mathcal{R}_0$ must depend on the assumptions about the exponential growth rate r and the shape of the generation-interval distribution $g(\tau)$.

Here, we use the gamma approximation framework [21] to (i) characterize the amount of uncertainty present in the exponential growth rates and the shape of the generation-interval distribution and (ii) assess the degree to which these uncertainties affect the estimate of $\mathcal{R}_0$. Assuming that generation intervals follow a gamma distribution with the mean $\bar{G}$ and the squared coefficient of variation κ, we have

$$\mathcal{R}_0 = \left(1 + \kappa r \bar{G}\right)^{1/\kappa}. \tag{2}$$

This equation demonstrates that a generation-interval distribution that has a larger mean (higher $\bar{G}$) or is less variable (lower κ) will give a higher estimate of $\mathcal{R}_0$ for the same value of r.

2.3 Statistical framework

As most studies do not report their estimates of the exponential growth rate, we first re-calculate the exponential growth rate that correspond to their model assumptions. We do so by modeling reported distributions of the reproductive number $\mathcal{R}_0$, the mean generation interval $\bar{G}$, and the generation-interval dispersion parameter κ with appropriate probability distributions; we used gamma distributions to model values reported with confidence intervals and uniform distributions to model values reported with ranges. For example, Study 3 estimated $\mathcal{R}_0 = 2.92$ (95% CI: 2.28–3.67); we model this estimate as a gamma distribution with a mean of 2.92 and a shape parameter of 67, which has a 95% probability of containing a value between 2.28 and 3.67 (see Table 2 for a complete description). For each study i, we construct a family of parameter sets by drawing 100,000 random samples from the probability distributions (Table 2) that represent the estimates of $\mathcal{R}_{0i}$ and the assumed values of $\bar{G}_i$ and κ_i and calculate the exponential growth rate r_i via the inverse of Eq. 2:

$$r_i = \frac{\mathcal{R}_{0i}^{\kappa_i} - 1}{\kappa_i \bar{G}_i}. \tag{3}$$

This allows us to approximate the probability distributions of the estimated exponential growth rates by each study; uncertainties in the probability distributions that we calculate for the estimated exponential growth rates will reflect the methods and assumptions that the studies rely on.

We construct pooled estimates for each parameter (r, $\bar{G}$, and κ) using a Bayesian multilevel modeling approach, which assumes that the parameters across different studies come from the same gamma distribution. The pooled estimates, which are represented as probability distributions rather than point estimates, allow us to average across different modeling approaches, while accounting for the uncertainties in the assumptions they make:

$$\begin{aligned} r_i &\sim \text{Gamma}(\text{mean} = \mu_r, \text{shape} = \mu_r^2/\sigma_r^2), \\ \bar{G}_i &\sim \text{Gamma}(\text{mean} = \mu_G, \text{shape} = \mu_G^2/\sigma_G^2), \\ \kappa_i &\sim \text{Gamma}(\text{mean} = \mu_\kappa, \text{shape} = \mu_\kappa^2/\sigma_\kappa^2), \end{aligned} \tag{4}$$

	Basic reproductive number $\mathcal{R}_0$	Mean generation interval $\bar{G}$ (days)	Generation-interval dispersion κ
Study 1	Uniform$(1.5, 3.5)$	10	1
Study 2	Gamma(mean $= 2.6, \alpha = 28$)	8.4	0.5
Study 3	Gamma(mean $= 2.92, \alpha = 67$)	8.4	0.2
Study 4	Gamma(mean $= 3.8, \alpha = 1400$)	7.6	0.5
Study 5	Gamma(mean $= 2.2, \alpha = 12$)	Uniform$(7, 14)$	0.5
Study 6	Gamma(mean $= 5.47, \alpha = 54$)	Uniform$(7.6, 8.4)^*$	0.2
Study 7	$\exp(r\bar{G})^\dagger$	Uniform$(6, 10)$	0

Table 2: **Probability distributions for $\mathcal{R}_0$, $\bar{G}$, and κ.** We use these probability distributions to obtain a probability distribution for the exponential growth rate r. The gamma distribution is parameterized by its mean and shape α. Constant values are fixed according to Table 1. *We do not account for this uncertainty during our recalculation of the exponential growth rate r because the reported estimate of $\mathcal{R}_0$ and its uncertainty assumes $\bar{G} = 8$; the original article reports three $\mathcal{R}_0$ (and 95% CIs) estimates using three different values of $\bar{G}$: 7.6 (MERS-like), 8 (average), and 8.4 (SARS-like). We still account for this uncertainty in our pooled estimates (μ_G). †Study 6 uses the IDEA model [22], through which the authors effectively fit an exponential curve to the cumulative number of confirmed cases without propagating any statistical uncertainty. Instead of modeling $\mathcal{R}_0$ with a probability distribution and recalculating r, we use $r = 0.114\,\text{days}^{-1}$, which explains all uncertainty in the reported $\mathcal{R}_0$, when combined with the considered range of $\bar{G}$.

where μ_r, μ_G, μ_κ represent the pooled estimates, and σ_r, σ_G, and σ_κ represent between-study standard deviations. We account for uncertainties associated with r_i, $\bar{G}_i$ and κ_i (and their correlations), by drawing a random set from the family of parameter sets for each study at each Metropolis-Hastings step. Since the gamma distribution does not allow zeros, we use $\kappa = 0.02$ instead for Study 7. We note that this approach does not account for non-independence between the parameter estimates made by different modelers. As we add more models, the pooled estimates can become sharper even when the models no longer add more information. Thus, the pooled estimator should be interpreted with care.

We use weakly informative priors on hyperparameters:

$$\begin{aligned} \mu_r &\sim \text{Gamma}(\text{mean} = 1\,\text{week}^{-1}, \text{shape} = 2) \\ \mu_G &\sim \text{Gamma}(\text{mean} = 1\,\text{week}, \text{shape} = 2) \\ \mu_\kappa &\sim \text{Gamma}(\text{mean} = 0.5, \text{shape} = 2) \\ (\sigma_r, \sigma_G, \sigma_\kappa) &\sim \text{half-normal}(0, 10). \end{aligned} \tag{5}$$

We followed recommendations outlined in Gelman *et al.* [23], parameterizing the top-level gamma distributions in terms of their means and standard deviations and imposing weakly informative prior distributions on between-study standard deviations, i.e., half-normal$(0, 10)$. We had initially used gamma priors with small shape parameters (< 1) on between-study shape parameters ($= \mu^2/\sigma^2$) but found this put too much prior probability on large between-study variances. This phenomenon is a known problem [23]. Alternative choices of prior for the between-study shape parameters are also suboptimal: imposing strong priors (e.g. half-

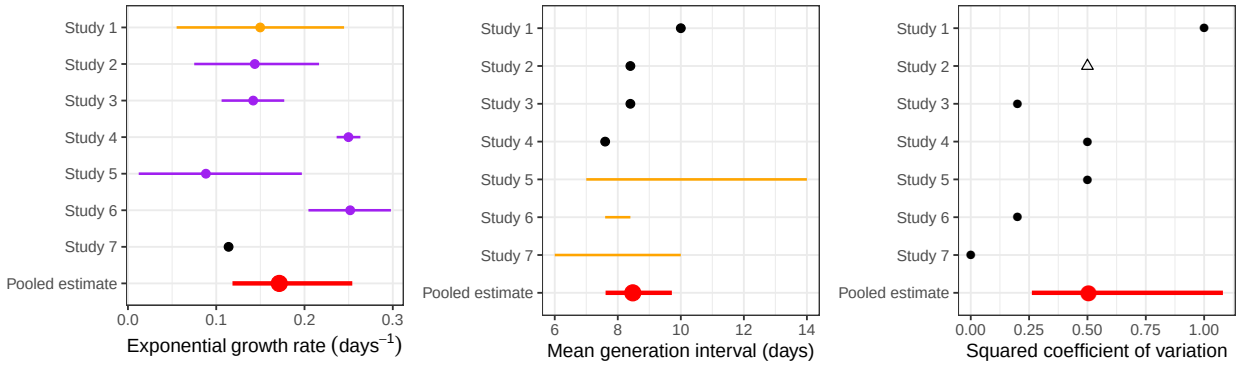

Figure 1: **Comparisons of the reported parameter values with our pooled estimates.** We inferred point estimates (black), uniform distributions (orange) or confidence intervals (purple) for each parameter from each study, and combined them into pooled estimates (red; see text). Open triangle: we assumed $\kappa = 0.5$ for Study 2 which does not report generation-interval dispersion.

$t(\mu = 0, \sigma = 1, \nu = 4)$ assumes *a priori* that between-study variance is large, while weak priors (e.g. half-Cauchy(0,5)) can lead to poor mixing.

We run 4 independent Markov Chain Monte Carlo chains each consisting of 500,000 burnin steps and 500,000 sampling steps. Posterior samples are thinned every 1000 steps. Convergence is assessed by ensuring that the Gelman-Rubin statistic is below 1.01 for all hyperparameters [24]; trace plots and marginal posterior distribution plots are presented in Appendix. 95% confidence intervals are calculated by taking 2.5% and 97.5% quantiles from the marginal posterior distribution for each parameter.

3 Results

Fig. 1 compares the reported values of the exponential growth rate r, mean generation interval $\bar{G}$, and the generation-interval dispersion κ from different studies with the pooled estimates that we calculate from our multilevel model. We find that there is a large uncertainty associated with the underlying parameters; many models rely on stronger assumptions that ignore these uncertainties. Surprisingly, no studies take into account how the variation in generation intervals affects their estimates of $\mathcal{R}_0$: all studies assumed fixed values for κ, ranging from 0 to 1.

Fig. 2 shows how propagating uncertainty in different combinations would affect estimates and CIs for $\mathcal{R}_0$. For illustrative purposes, we use our pooled estimates, which may represent a reasonable proxy for the state of knowledge as of January 23–26 (Fig. 2A). Comparing the models that include only some sources of uncertainty to the "all" model, we see that propagating error from the growth rate (which all but one of the studies reviewed did) is absolutely crucial: the middle bar ("GI mean"), which lacks growth-rate uncertainty, is relatively narrow. In this case, propagating error from the mean generation interval

has negligible effect compared to propagating the uncertainty in r. Uncertainty in the generation-interval dispersion also has important effects as it determines the functional form of the relationship between r and $\mathcal{R}_0$ (compare "growth rate + GI mean" with "all"). For example, reducing the dispersion parameter κ from 1 (assuming exponentially distributed generation intervals) to 0 (assuming fixed generation intervals) changes the r–$\mathcal{R}_0$ relationship from linear to exponential, therefore increasing the sensitivity of $\mathcal{R}_0$ estimates to r and $\bar{G}$.

As uncertainty associated with the exponential growth rate decreases, accounting for uncertainties in generation intervals becomes even more important (Fig. 2B). Propagating error only from the growth rate gives very narrow confidence intervals in this case. Likewise, propagating errors from the growth rate and the mean generation interval gives wider but still too narrow confidence intervals. We expect this hypothetical example to better reflect more recent scenarios, as increased data availability will allow researchers to estimate r with more certainty.

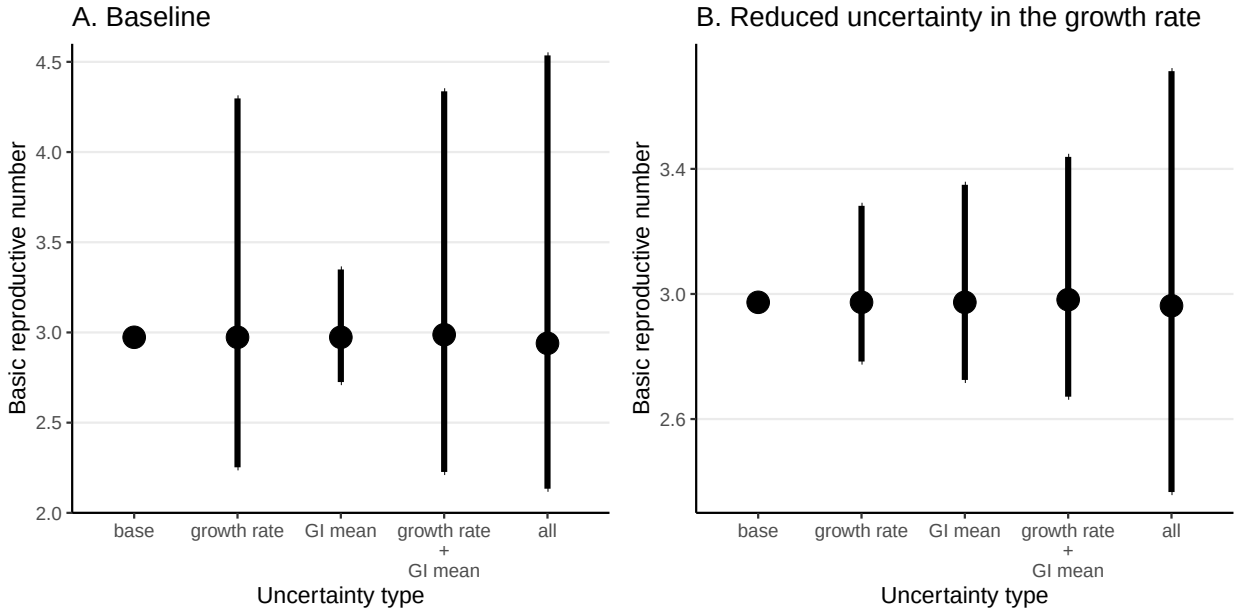

Figure 2: **Effects of r, $\bar{G}$, and κ on the estimates of $\mathcal{R}_0$.** We compare estimates of $\mathcal{R}_0$ under five scenarios that propagate different combinations of uncertainties (A) based on our pooled estimates (μ_r, μ_G, and μ_κ) and (B) assuming a 4-fold reduction in uncertainty of our pooled estimate of the exponential growth rate (using $(\mu_r + 3 \times \text{median}(\mu_r))/4$, instead). **base**: $\mathcal{R}_0$ estimates based on the median estimates of μ_r, μ_G, and μ_κ. **growth rate**: $\mathcal{R}_0$ estimates based on the the posterior distribution of μ_r while using median estimates of μ_G and μ_κ. **GI mean**: $\mathcal{R}_0$ estimates based on the the posterior distribution of μ_G while using median estimates of μ_r and μ_κ. **growth rate + GI mean**: $\mathcal{R}_0$ estimates based on the the joint posterior distributions of μ_r and μ_G while using a median estimate of μ_κ. **all**: $\mathcal{R}_0$ estimates based on the joint posterior distributions of μ_r, μ_G, and μ_κ. Vertical lines represent the 95% confidence intervals.

We also compare the estimates of $\mathcal{R}_0$ across different studies by replacing their values

of r, $\bar{G}$, and κ with our pooled estimates (μ_r, μ_G, and μ_κ, respectively) one at a time and recalculating the basic reproductive number $\mathcal{R}_0$ (Fig. 3). This procedure allows us to assess the sensitivity of the estimates of $\mathcal{R}_0$ across appropriate ranges of uncertainties. We find that incorporating uncertainties one at a time increases the width of the confidence intervals in all but 7 cases. We estimate narrower confidence intervals for Study 3, Study 6, and Study 7 when we account for proper uncertainties in the generation-interval dispersion because they assume a narrow generation-interval distribution (compare "base" with "GI variation"); when higher values of κ are used, their estimates of $\mathcal{R}_0$ become less sensitive to the values of r and $\bar{G}$, giving narrower confidence intervals. We estimate narrower confidence intervals for Study 5 and Study 7 when we account for proper uncertainties in the mean generation interval (compare "base" with "GI mean") because the range of uncertainty in the mean generation interval $\bar{G}$ they consider is much wider than the pooled range (Fig. 1). Substituting the reported r or $\bar{G}$ from Study 1 with our pooled estimates give narrower confidence intervals for similar reasons.

We find that accounting for uncertainties in the estimate of r has the largest effect on the estimates of $\mathcal{R}_0$ in most cases (Fig. 3). For example, recalculating $\mathcal{R}_0$ for Study 7 by using our pooled estimate of r gives $\mathcal{R}_0 = 3.9$ (95% CI: 2.3–8.6), which is much wider than the uncertainty range they reported (2.0–3.1). There are two explanations for this result. First, even though the exponential growth rate r and the mean generation interval $\bar{G}$ have identical mathematical effects on $\mathcal{R}_0$ in our framework (Eq. 2 in Methods), r is more influential in this case because it is associated with more uncertainty (Fig. 1). Second, assuming a fixed generation interval ($\kappa = 0$) makes the estimate of $\mathcal{R}_0$ too sensitive to r and $\bar{G}$. One exception is Study 1: we find this estimate of $\mathcal{R}_0$ is most sensitive to generation-interval dispersion κ. This is because Study 1 assumes an exponentially distributed generation interval ($\kappa = 1$): estimates that rely on this assumption make $\mathcal{R}_0$ relatively insensitive and thus tend to have particularly narrow confidence intervals.

Finally, we incorporate all uncertainties by using posterior samples for μ_r, μ_G, and μ_κ to recalculate $\mathcal{R}_0$ and compare it with the reported $\mathcal{R}_0$ estimates. Our estimated $\mathcal{R}_0$ from the pooled distribution has a median of 2.9 (95% CI: 2.1–4.5). While the point estimate of $\mathcal{R}_0$ is similar to other reported values from this date range, the confidence intervals are wider than all but one study. This result does not imply that assumptions based on the pooled estimate are too weak; we believe that this confidence interval more accurately reflects the level of uncertainties present in the information that was available when these models were fitted. In fact, because the pooled estimate does not account for overlap in data sources used by the models, we feel that it is more likely to be over-confident than under-confident. Our median estimate averages over the various studies, and therefore particular studies have higher or lower median estimates. We note in particular that, while the baseline example we used from Study 6 may appear to be an outlier, the authors of this study also explore different scenarios involving changes in reporting rate over time, under which their estimates of $\mathcal{R}_0$ are similar to other reported estimates. Here, our focus is on estimating uncertainty, not on identifying potential explanations for these discrepancies.

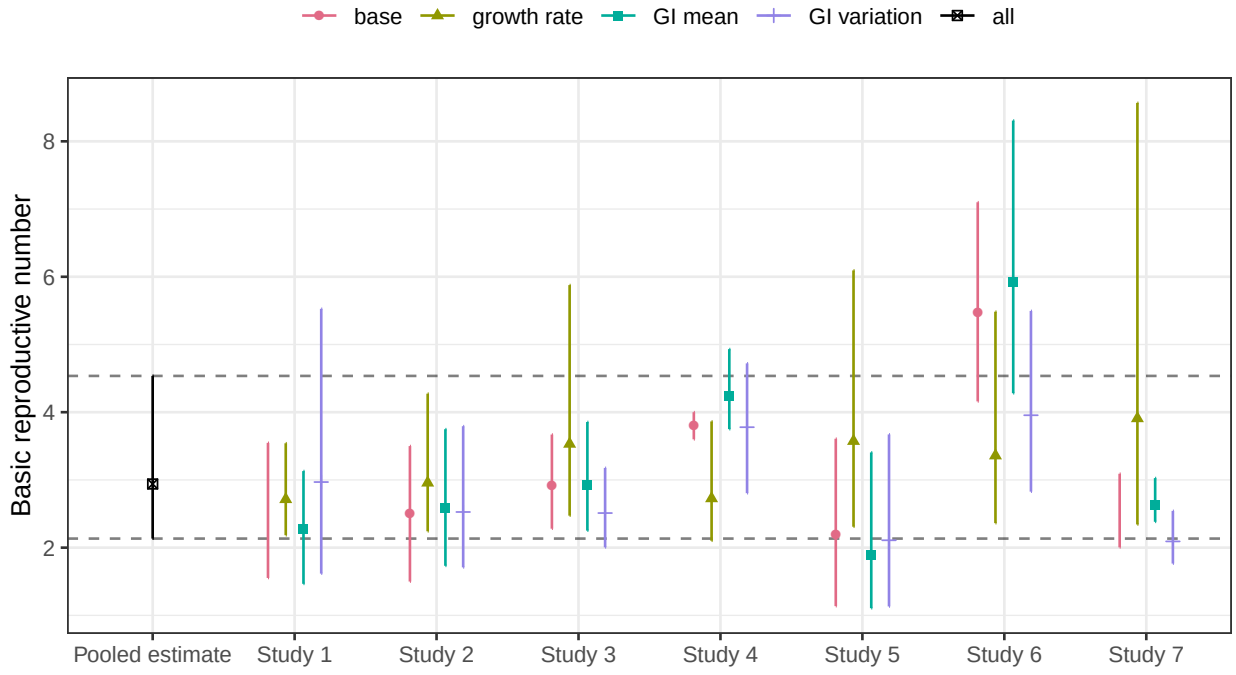

Figure 3: **Sensitivity of the reported $\mathcal{R}_0$ estimates with respect to our pooled estimates of the underlying parameters.** We replace the reported parameter values (growth rate r, GI mean $\bar{G}$, and GI variation κ) with our corresponding pooled estimates (μ_r, μ_G, and μ_κ) one at a time and recalculate $\mathcal{R}_0$ (**growth rate**, **GI mean**, and **GI variation**). The pooled estimate of $\mathcal{R}_0$ is calculated from the joint posterior distribution of μ_r, μ_G, and μ_κ (**all**); this corresponds to replacing all reported parameter values with our pooled estimates, which gives identical results across all studies. Horizontal dashed lines represent the 95% confidence intervals of our pooled estimate of $\mathcal{R}_0$. The reported $\mathcal{R}_0$ estimates (**base**) have been adjusted to show the approximate 95% confidence interval using the probability distributions that we defined if they had relied on different measures for parameter uncertainties.

4 Discussion

Estimating the basic reproductive number $\mathcal{R}_0$ is crucial for predicting the course of an outbreak and planning intervention strategies. Here, we use a gamma approximation [21] to decompose $\mathcal{R}_0$ estimates into three key quantities (r, $\bar{G}$, and κ) and apply a multilevel Bayesian framework to compare estimates of $\mathcal{R}_0$ for the novel coronavirus outbreak. Our results demonstrate the importance of accounting for uncertainties associated with the underlying generation-interval distributions, including uncertainties in the amount of dispersion in the generation intervals: our analysis of individual studies shows that assuming too narrow a generation-interval distribution can make the estimate of $\mathcal{R}_0$ overly sensitive to the estimates of the exponential growth rate r.

Of the seven studies that we reviewed, two of them directly fit their models to cumulative number of confirmed cases. This approach can be appealing because of its simplicity and apparent robustness, but fitting a model to cumulative incidence instead of raw incidence can both bias parameters and give overly narrow confidence intervals, if the resulting non-independent error structures are not taken into account [19, 25]. Naive fits to cumulative incidence data should therefore be avoided.

Many sources of noise affect real-world incidence data, including both dynamical, or "process", noise (randomness that directly or indirectly affects disease transmission); and observation noise (randomness underlying how many of the true cases are reported). Disease modelers face the choice of incorporating one or both of these in their data-fitting and modeling steps. This is not always a serious problem, particularly if the goal is inferring parameters rather than directly making forecasts [19]. Modelers should however be aware of the possibility that ignoring one kind of error can give overly narrow confidence intervals [25, 26].

There are other important phenomena not covered by our simple framework. Examples that seem relevant to this outbreak include: changing reporting rates, reporting delays (including the effects of weekends and holidays), and changing generation intervals. For emerging pathogens such as 2019-nCoV, there may be an early period of time when the reporting rate is very low due to limited awareness or diagnostic resources; for example, Zhao *et al.* [9] (Study 6) demonstrated that estimates of $\mathcal{R}_0$ can change from 5.47 (95% CI: 4.16–7.10) to 3.30 (95% CI: 2.73–3.96) when they assume 2-fold changes in the reporting rate between January 17, when the official diagnostic guidelines were released [27], and January 20. Delays between key epidemiological timings (e.g., infection, symptom onset, and detection) can also shift the shape of an observed epidemic curve and, therefore, affect parameter estimates as well as predictions of the course of an outbreak [28]. Finally, generation intervals can become shorter throughout an epidemic as intervention strategies, such as quarantine, can reduce the infectious period [29]. Accounting for these factors is crucial for making accurate inferences.

Here, we focused on the estimates of $\mathcal{R}_0$ that were published within a very short time frame (January 23–26). During early phases of an outbreak, it is reasonable to assume that the epidemic grows exponentially [11]. However, as the number of susceptible individuals decreases, the epidemic will saturate, and estimates of r used for $\mathcal{R}_0$ should account for the possibility that r is decreasing through time. Although our analysis only reflects a snapshot of a fast-moving epidemic, we expect certain lessons to hold: confidence intervals must combine different sources of uncertainty. In fact, as epidemics progress and more data becomes available, it is likely that inferences about exponential growth rate (and other epidemiological parameters) will become more precise; thus the risk of over-confidence when uncertainty about the generation-interval distribution is neglected will become greater.

We strongly emphasize the value of attention to accurate characterization of the transmission chains via contact tracing and better statistical frameworks for inferring generation-interval distributions from such data [30]. A combined effort between public-health workers and modelers in this direction will be crucial for predicting the course of an epidemic and controlling it. We also emphasize the value of transparency from modelers. Model estimates

during an outbreak, even in pre-prints, should include code links and complete explanations. We suggest using methods based on open-source tools allow for maximal reproducibility.

In summary, we have provided a basis for comparing exponential-growth based estimates of $\mathcal{R}_0$ and its associated uncertainty in terms of three components: the exponential growth rate, mean generation interval, and generation interval dispersion. We hope this framework will help researchers understand and reconcile disparate estimates of disease transmission early in an epidemic.

References

[1] World Health Organization. Pneumonia of unknown cause – China. 2020. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/. Accessed January 30, 2020.

[2] Jasper Fuk-Woo Chan, Shuofeng Yuan, Kin-Hang Kok, Kelvin Kai-Wang To, Hin Chu, Jin Yang, Fanfan Xing, Jieling Liu, Cyril Chik-Yan Yip, Rosana Wing-Shan Poon, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*, 2020.

[3] World Health Organization. Novel Coronavirus (2019-nCoV) Situation Report - 14. 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200203-sitrep-14-ncov.pdf?sfvrsn=f7347413_2. Accessed February 3, 2020.

[4] Trevor Bedford, Richard Neher, James Hadfield, Emma Hodcroft, Misja Ilcisin, and Nicola Müller. Genomic analysis of nCoV spread. Situation report 2020-01-23. 2020. https://nextstrain.org/narratives/ncov/sit-rep/2020-01-23. Accessed 24, January, 2020.

[5] Natsuok Imai, Anne Cori, Ilaria Dorigatti, Marc Baguelin, Chirstl A. Donelly, Steven Riley, and Neil M. Ferguson. Report 3: Transmissibility of 2019-nCoV. 2020. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-2019-nCoV-transmissibility.pdf. Accessed 26, January, 2020.

[6] Tao Liu, Jianxiong Hu, Min Kang, Lifeng Lin, Haojie Zhong, Jianpeng Xiao, Guanhao He, Tie Song, Qiong Huang, Zuhua Rong, Aiping Deng, Weilin Zeng, Xiaohua Tan, Siqing Zeng, Zhihua Zhu, Jiansen Li, Donghua Wan, Jing Lu, Huihong Deng, Jianfeng He, and Wenjun Ma. Transmission dynamics of 2019 novel coronavirus (2019-nCoV). 2020. https://www.biorxiv.org/content/10.1101/2020.01.25.919787v1. Accessed 27, January, 2020.

[7] Maimuna Majumder and Kenneth D. Mandl. Early transmissibility assessment of a novel coronavirus in Wuhan, China. 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3524675. Accessed 27, January, 2020.

[8] Jonathan M. Read, Jessica R.E. Bridgen, Derek A.T. Cummings, Antonia Ho, and Chris P. Jewell. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. 2020. https://www.medrxiv.org/content/10.1101/2020.01.23.20018549v1. Accessed 26, January, 2020.

[9] Shi Zhao, Jinjun Ran, Salihu S. Musa, Gaungpu Yang, Yijun Lou, Daozhou Gao, Lin Yang, and Daihai He. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. 2020.

https://www.biorxiv.org/content/10.1101/2020.01.23.916395v1. Accessed 26, January, 2020.

[10] Julien Riou and Christian L Althaus. Pattern of early human-to-human transmission of wuhan 2019-nCoV. 2020. https://www.biorxiv.org/content/10.1101/2020.01.23.917351v1. Accessed 26, January, 2020.

[11] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1991.

[12] Åke Svensson. A note on generation times in epidemic models. *Math Biosci*, 208(1): 300–311, 2007.

[13] Jacco Wallinga and Marc Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc R Soc Lond B Biol Sci*, 274 (1609):599–604, 2007.

[14] Bret D. Elderd, Vanja M. Dukic, and Greg Dwyer. Uncertainty in predictions of disease spread and public health responses to bioterrorism and emerging diseases. *Proc Natl Acad Sci USA*, 103(42):15693 –15697, October 2006. doi: 10.1073/pnas.0600816103. URL `http://www.pnas.org/content/103/42/15693.abstract`.

[15] Natsuok Imai, Ilaria Dorigatti, Anne Cori, Chirstl A. Donelly, Steven Riley, and Neil M. Ferguson. Report 3: Transmissibility of 2019-nCoV. 2020. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/2019-nCoV-outbreak-report-22-01-2020.pdf. Accessed 3, February, 2020.

[16] Julien Riou and Christian L Althaus. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro Surveill*, 25(4):2000058, 2020.

[17] Shi Zhao, Qianyin Lin, Jinjun Ran, Salihu S Musa, Guangpu Yang, Weiming Wang, Yijun Lou, Daozhou Gao, Lin Yang, Daihai He, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int J Infect Dis*, 2020.

[18] Jonathan M. Read, Jessica R.E. Bridgen, Derek A.T. Cummings, Antonia Ho, and Chris P. Jewell. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. 2020. https://www.medrxiv.org/content/10.1101/2020.01.23.20018549v2. Accessed 5, February, 2020.

[19] Junling Ma, Jonathan Dushoff, Benjamin M Bolker, and David JD Earn. Estimating initial epidemic growth rates. *Bull Math Biol*, 76(1):245–260, 2014.

[20] David Champredon and Jonathan Dushoff. Intrinsic and realized generation intervals in infectious-disease transmission. *Proc R Soc Lond B Biol Sci*, 282(1821):20152026, 2015.

[21] Sang Woo Park, David Champredon, Joshua S Weitz, and Jonathan Dushoff. A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics*, 27:12–18, 2019.

[22] David N Fisman, Tanya S Hauck, Ashleigh R Tuite, and Amy L Greer. An IDEA for short term outbreak projection: nearcasting using the basic reproduction number. *PloS One*, 8(12), 2013.

[23] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534, 2006.

[24] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Stat Sci*, 7(4):457–472, 1992.

[25] Aaron A King, Matthieu Domenech de Cellès, Felicia MG Magpantay, and Pejman Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc R Soc Lond B Biol Sci*, 282(1806):20150347, 2015.

[26] Bradford P Taylor, Jonathan Dushoff, and Joshua S Weitz. Stochasticity and the limits to confidence when estimating $\mathcal{R}_0$ of Ebola and other emerging infectious diseases. *J Theor Biol*, 408:145–154, 2016.

[27] World Health Organization. Laboratory testing for 2019 novel coronavirus (2019-nCoV) in suspected human cases. 2020. https://www.who.int/publications-detail/laboratory-testing-for-2019-novel-coronavirus-in-suspected-human-cases-20200117. Accessed February 4, 2020.

[28] A Tariq, K Roosa, K Mizumoto, and G Chowell. Assessing reporting delays and the effective reproduction number: The Ebola epidemic in DRC, May 2018–January 2019. *Epidemics*, 26:128–133, 2019.

[29] Herbert Hethcote, Ma Zhien, and Liao Shengbing. Effects of quarantine in six endemic models for infectious diseases. *Math Biosci*, 180(1-2):141–160, 2002.

[30] Tom Britton and Gianpaolo Scalia Tomba. Estimation in emerging epidemics: Biases and remedies. *J R Soc Interface*, 16(150):20180670, 2019.

Appendix

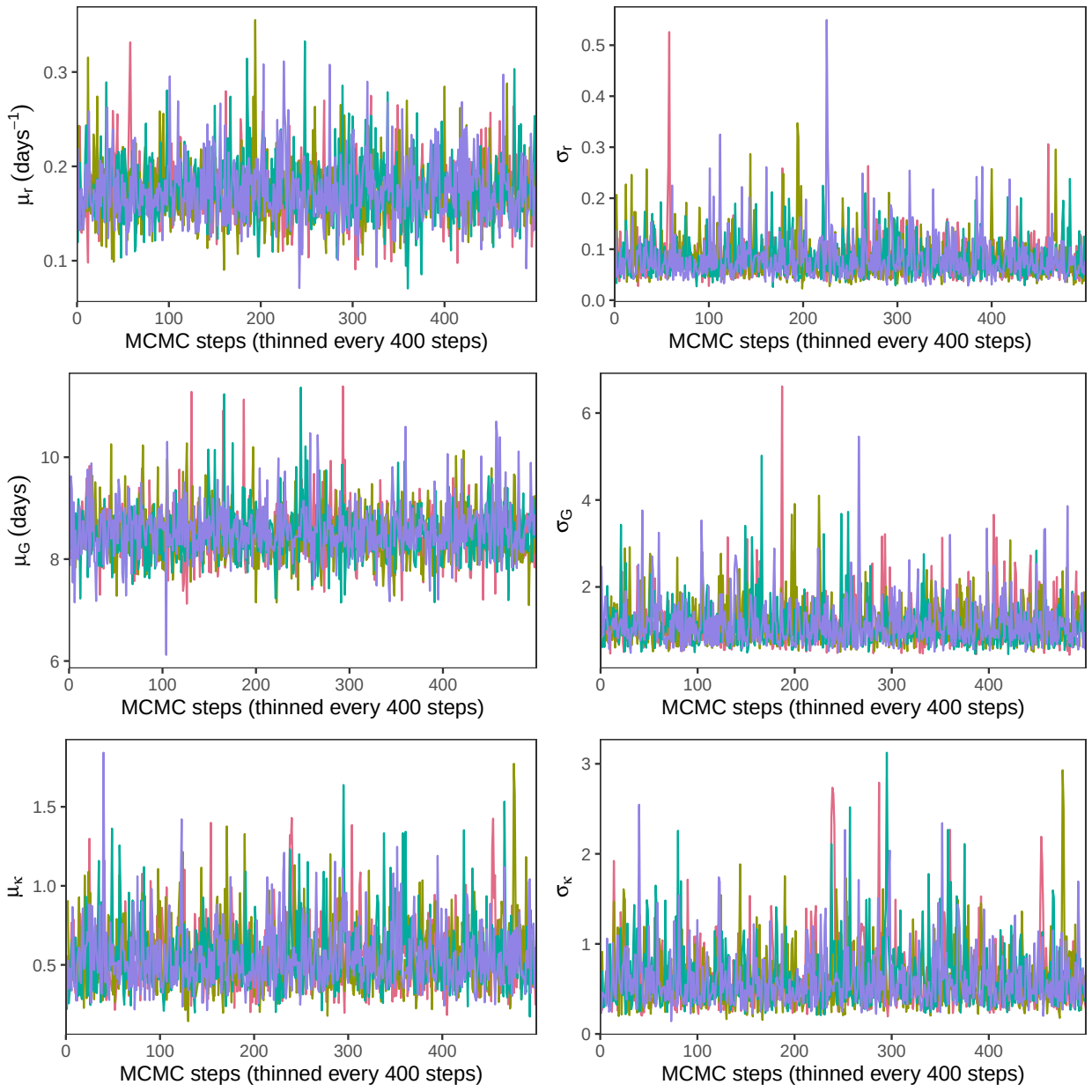

Figure A1: **Trace plots of the multilevel model.** Each chain is represented by a different color.

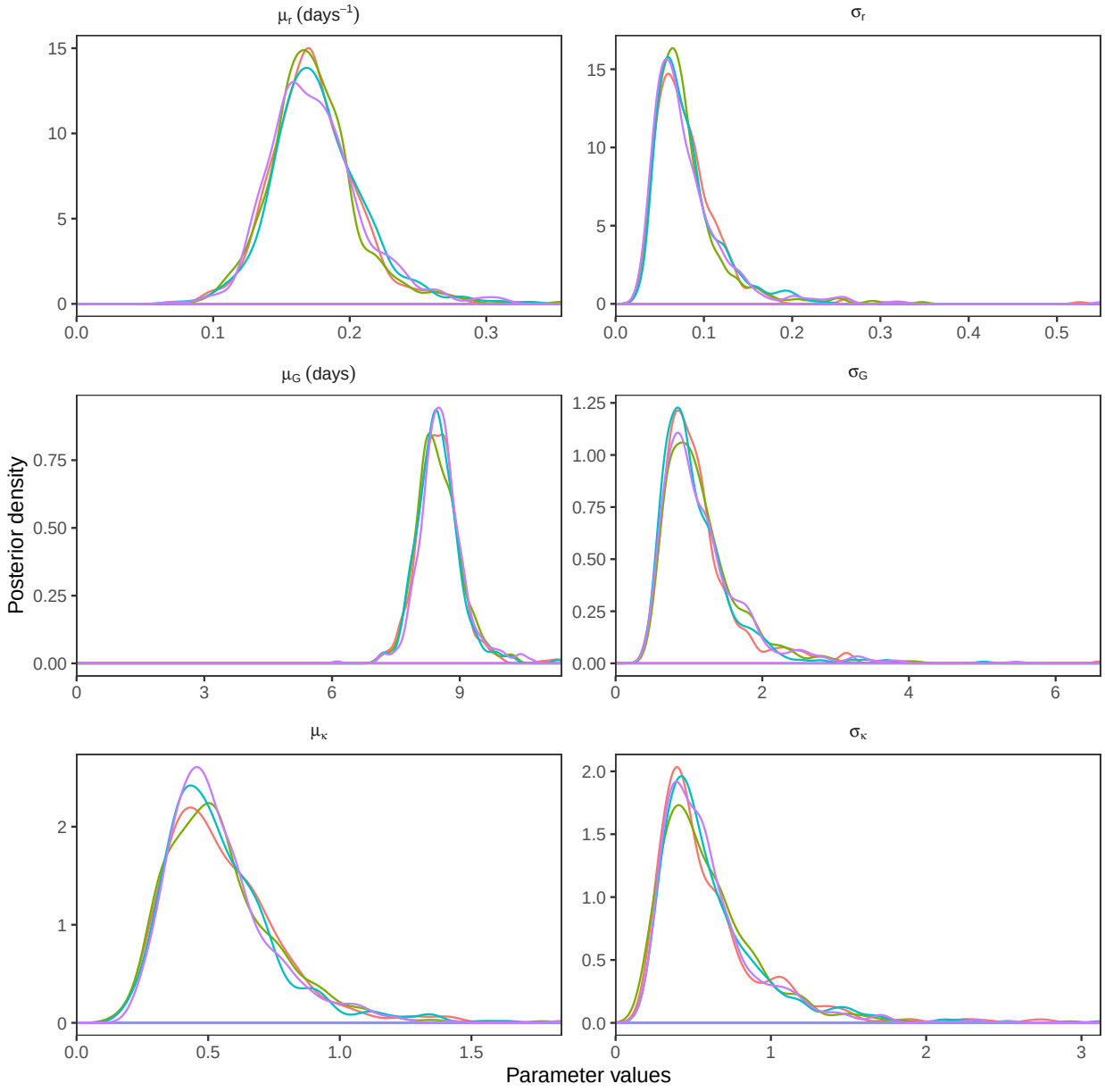

Figure A2: **Marginal posterior distributions of the multilevel model.** Each chain is represented by a different color.