medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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

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

1 Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA
2 Department of Pathology and Laboratory Medicine, University of Western Ontario, London, Ontario, 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 Biology, McMaster University, Hamilton, ON, 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

A novel coronavirus (2019-nCoV) has recently emerged as a global threat. As the epidemic progresses, many disease modelers have prioritized estimating the basic reproductive number $\mathcal{R}_0$, defined as the average number of secondary cases caused by a primary case. While these efforts are extremely valuable, their modeling approaches and the resulting estimates vary widely. Here, we present a framework for comparing different estimates of $\mathcal{R}_0$ across a wide range of models by decomposing it into three key quantities (the exponential growth rate r, the mean generation interval $\bar{G}$, and the generation-interval dispersion κ) and apply our framework to early estimates of $\mathcal{R}_0$ for the 2019-nCoV outbreak. Our results emphasize the importance of propagating uncertainties in all three quantities, in particular in the shape of the generation-interval distribution. While rapid response during an outbreak is important, careful consideration of methodology is also worthwhile. Modelers should work with field-workers to develop better methods for characterizing generation intervals.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

1 Introduction

Since December 2019, a novel coronavirus (2019-nCoV) has been spreading in China and other parts of the world (World Health Organization, 2020c). Although the virus is believed to have originated from animal reservoirs (Centers for Disease Control and Prevention, 2020), its ability to directly transmit between humans has posed a greater threat for its spread (Huang et al., 2020; World Health Organization, 2020a). As of January 30th, 2020, the World Health Organization (WHO) has confirmed 7818 cases, including 82 confirmed cases in 18 different countries, outside China (World Health Organization, 2020b). WHO has now declared the outbreak a Public Health Emergency of International Concern (World Health Organization, 2020d).

As the disease continues to spread, many researchers have published analyses of the outbreak, 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 (Anderson and May, 1991; Diekmann et al., 1990)). Estimating the basic reproductive number is of interest during an outbreak because it provides information about the level of intervention required to interrupt transmission (Anderson and May, 1991), and about the final size of the outbreak (Anderson and May, 1991; Ma and Earn, 2006). We commend these researchers for their timely contribution and those who made the data publicly available. However, it can be difficult to assess the estimates of $\mathcal{R}_0$ (as well as the associated degrees of uncertainty) when the estimation methods and their underlying assumptions vary widely, especially since these assumptions can affect the estimates.

Here, we show that a wide range of approaches to estimating $\mathcal{R}_0$ can be understood and compared in terms of estimates for three quantities: the exponential growth rate r, the mean generation interval $\bar{G}$, and the generation-interval dispersion κ. The generation interval, which is defined as the time between when an individual becomes infected and when that individual infects another individual (Svensson, 2007), plays a key role in shaping the relationship between r and $\mathcal{R}_0$ (Wearing et al., 2005; Roberts and Heesterbeek, 2007; Wallinga and Lipsitch, 2007; Park et al., 2019); therefore, estimates of $\mathcal{R}_0$ from different models directly depend on their implicit assumptions about the shape of the generation-interval distribution and the exponential growth rate. Early in an epidemic, information is scarce and, inevitably, there is large uncertainty around case reports (affecting the estimates of the exponential growth rate) and contact tracing (affecting the estimates of the generation-interval distribution). We suggest that disease modelers should make sure their assumptions about these three quantities are clear and reasonable, and that estimates of uncertainty should propagate error from all three sources.

We evaluate six disparate models published online between January 24–26, 2020 that estimated $\mathcal{R}_0$ for the 2019-nCoV outbreak (Imai et al., 2020; Liu et al., 2020; Majumder and Mandl, 2020; Read et al., 2020; Riou and Althaus, 2020; Zhao et al., 2020). We use our framework to construct pooled estimates for the three key quantities: r, $\bar{G}$, and κ. 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 unravel which assumptions of these different models led to their different

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

	Basic reproductive number $\mathcal{R}_0$	Mean generation time $\bar{G}$ (days)	Generation-interval dispersion κ	
Study 1	2.5 (1.5–3.5)*	8.4	unspecified†	Imai et al. (2020)
Study 2	2.92 (95% CI: 2.28–3.67)	8.4	0.2	Liu et al. (2020)
Study 3	3.8 (95% CI: 3.6–4.0)	7.6	0.5	Read et al. (2020)
Study 4	2.2 (90% CI: 1.4–3.8)	7–14	0.5	Riou and Althaus (2020)
Study 5	5.47 (95% CI: 4.16–7.10)‡	7.6–8.4	0.2	Zhao et al. (2020)
Study 6	2.0–3.1	6–10	0	Majumder and Mandl (2020)

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 6 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; we use their baseline scenario in our analysis.

estimates and confidence intervals.

2 Results

We gathered information on estimates of $\mathcal{R}_0$ and their assumptions about the underlying generation-interval distributions from 6 articles that were published online between January 24th, 2020 and January 26th, 2020 (Table 1). As most studies do not report their estimates of the exponential growth rate or the associated confidence intervals, we first recalculate the exponential growth rate that correspond their model assumptions. We do so by modeling explicitly or implicitly 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 2 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). We then constructed a family of parameter sets – which include r, $\bar{G}$, and κ – for each study and used these in a Bayesian multilevel model to build a distribution of pooled estimates (see Methods).

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 (μ_r, μ_G, and μ_κ) 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 0.5. Assuming fixed parameter values can lead to overly strong conclusions (Elderd et al., 2006). It is also interesting that none of the six studies explicitly or implicitly assumed an exponentially distributed generation interval (i.e., $\kappa = 1$),

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

	Basic reproductive number $\mathcal{R}_0$	Mean generation time $\bar{G}$ (days)	Generation-interval dispersion κ
Study 1	Gamma(mean $= 2.6, \alpha = 28$)	8.4	0.5
Study 2	Gamma(mean $= 2.92, \alpha = 67$)	8.4	0.2
Study 3	Gamma(mean $= 3.8, \alpha = 1400$)	7.6	0.5
Study 4	Gamma(mean $= 2.2, \alpha = 12$)	Uniform$(7, 14)$	0.5
Study 5	Gamma(mean $= 5.47, \alpha = 54$)	Uniform$(7.6, 8.4)$	0.2
Study 6	$\exp(rG)^*$	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. *Study 6 uses the IDEA model (Fisman et al., 2013), 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 $\mathcal{R}_0$ they report, when combined with the range of $\bar{G}$ they consider.

an assumption which used to be extremely common, particularly implicitly.

Fig. 2 shows how propagating uncertainty (μ_r, μ_G, and μ_κ) 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 26 January. 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 far too narrow. Propagating error from the generation interval also has important effects. Once these two are included, the impact of leaving out uncertainty in the dispersion is small, though noticeable, in this particular example.

We also evaluate 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) one at a time and recalculating the basic reproductive number $\mathcal{R}_0$ (Fig. 3). We find that incorporating uncertainties one at a time increases the width of the confidence intervals all but three cases. We estimate slightly narrower confidence intervals for Study 2 and Study 6 when we use our pooled estimate of the generation-interval dispersion μ_κ to recalculate $\mathcal{R}_0$ because they assume a narrow generation-interval distribution (compare **base** with **GI variation** in Fig. 3); 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 4 when we use our pooled estimate of the mean generation time μ_G to recalculate $\mathcal{R}_0$ (compare **base** with **GI mean** in Fig. 3) because the range of uncertainty in the mean generation time $\bar{G}$ they consider is much wider than the pooled range (Fig. 1).

Consistent with our previous observations (Fig. 2), we find that accounting for uncertainties in the estimate of r has the largest effect on the estimates of $\mathcal{R}_0$ (Fig. 3). For example, recalculating $\mathcal{R}_0$ for Study 6 by using our pooled estimate of r gives $\mathcal{R}_0 = 3.9$ (95% CI: 2.3–9.8), 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

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

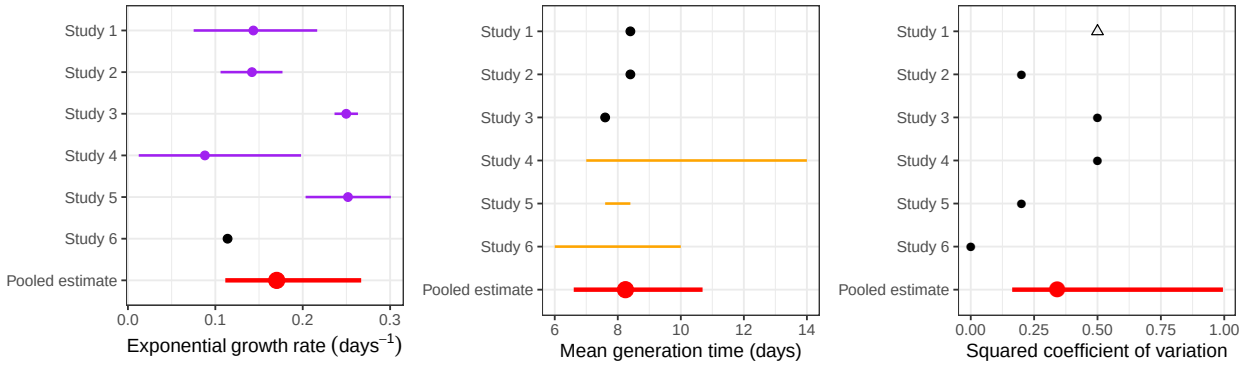

Figure 1: **Comparisons of the reported parameter values with our pooled estimates.** We inferred point estimates (black), uniform distributions (orange) or confidence intervals (purple) for each parameter from each study, and combined them into pooled estimates (red; see text). Open triangle: we assumed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

the mean generation time $\bar{G}$ have identical effects on $\mathcal{R}_0$ under the gamma approximation framework (Eq. 2 in Methods), r has a greater overall effect on $\mathcal{R}_0$ because it is associated with more uncertainty (Fig. 1). Second, assuming a fixed generation time ($\kappa = 0$) makes the estimate of $\mathcal{R}_0$ too sensitive to r and $\bar{G}$ as discussed previously.

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 3.1 (95% CI: 2.1–5.7). While the point estimate of $\mathcal{R}_0$ is similar to other reported values from this date range, the confidence intervals are wider than those of other studies. 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. Here, our focus is on certainty, not on the reason for these discrepancies.

3 Discussion

Estimating the basic reproductive number $\mathcal{R}_0$ is crucial for predicting the course of an outbreak and planning intervention strategies. Here, we used a simple framework (Park et al., 2019) 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 with the amount of dispersion in the generation intervals: although our pooled estimates are relatively insensitive to the estimated

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

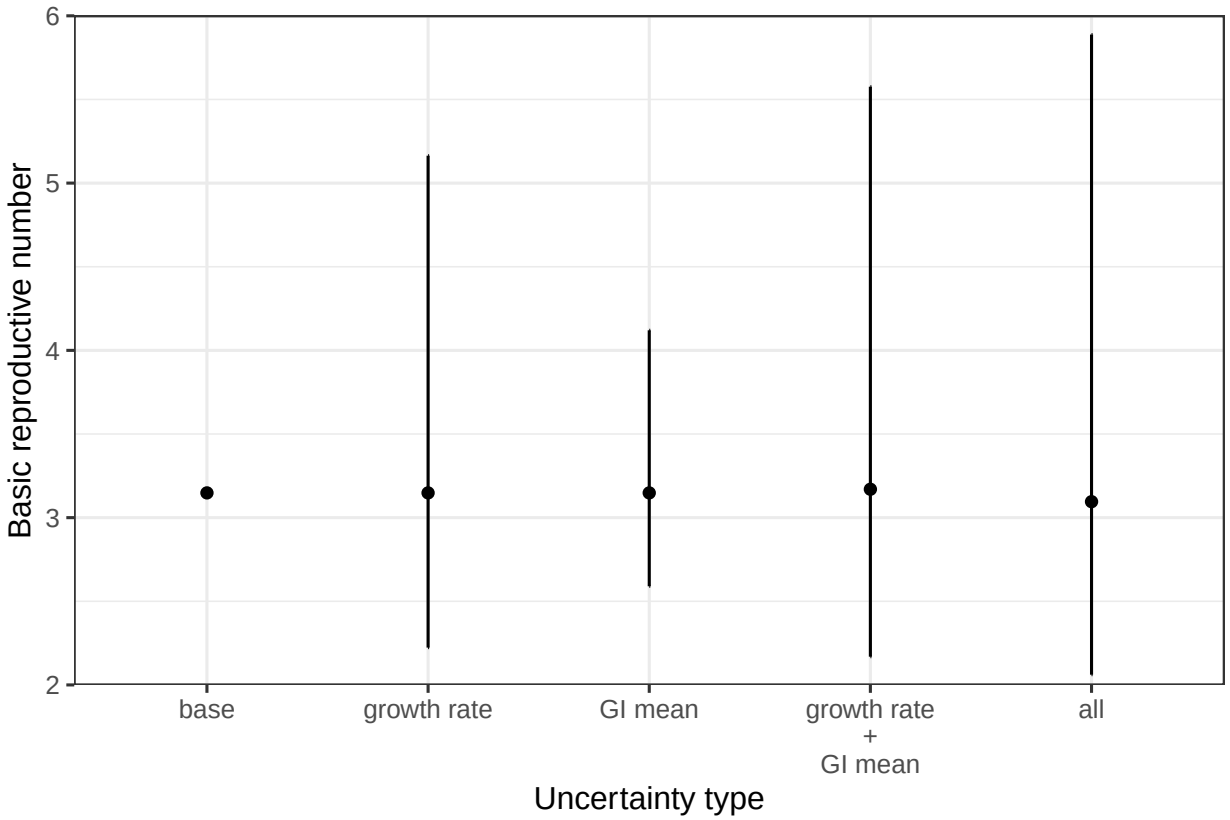

Figure 2: **Effects of r, $\bar{G}$, and κ on the estimates of $\mathcal{R}_0$.** We compare estimates of $\mathcal{R}_0$ under five scenarios that propagate different combinations of uncertainties. **base**: $\mathcal{R}_0$ estimates based on the median estimates of μ_r, μ_G, and μ_κ. **growth rate**: $\mathcal{R}_0$ 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.

uncertainty, our analysis of individual studies shows that assuming too narrow a generation-interval distribution can make the estimate of $\mathcal{R}_0$ too sensitive to the estimates of the exponential growth rate r.

In this study, we focused on propagating errors arising from implicit or explicit estimates of growth rate and generation intervals. Other key issues underlying early estimates of $\mathcal{R}_0$ include statistical independence and types of noise.

Of the six 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 result non-

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

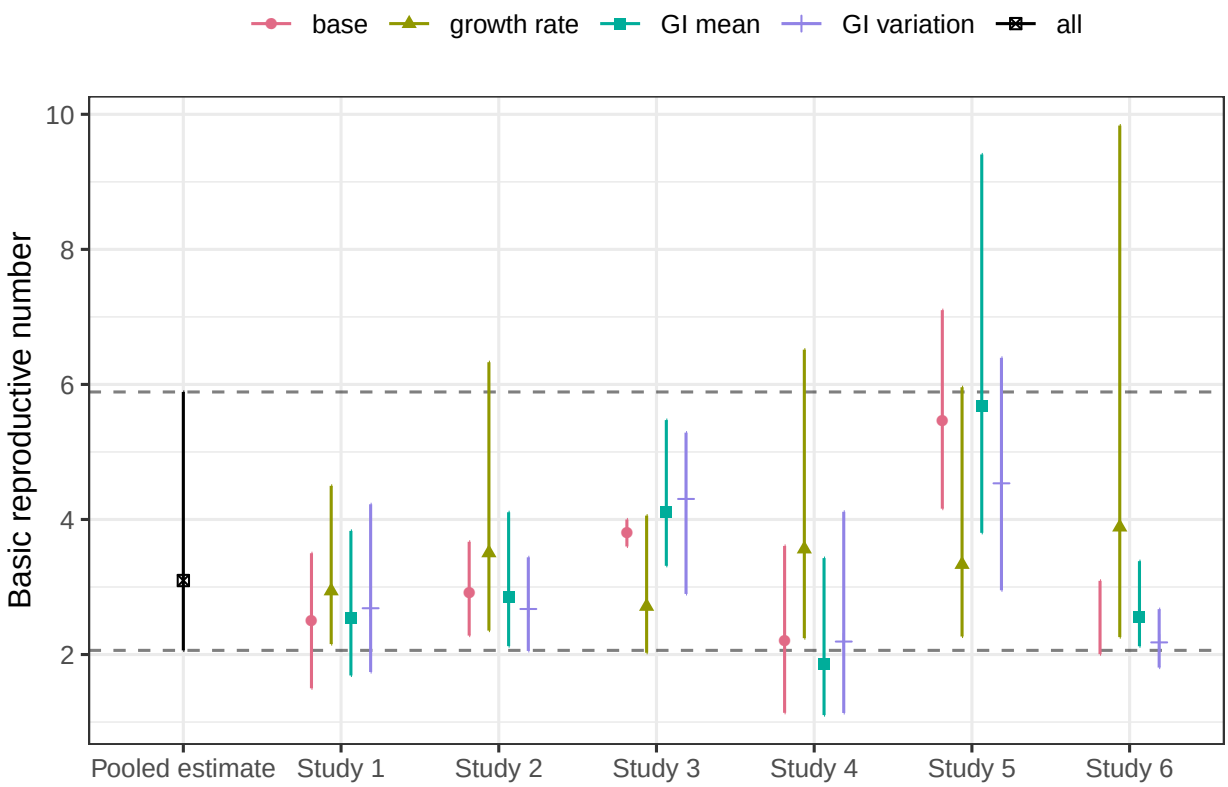

Figure 3: **Sensitivity of the reported $\mathcal{R}_0$ estimates with respect to our pooled estimates of the underlying parameters.** We replace the reported parameter values (growth rate r, GI mean $\bar{G}$, and GI variation κ) with our corresponding pooled estimates (μ_r, μ_G, and μ_κ) one at a time 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.

independent error structures are not taken into account (Ma et al., 2014; King et al., 2015). Naive fits to cumulative incidence data should be avoided.

There are many sources of noise in 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 (e.g., Ma et al. (2014)). Modelers should

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

be aware of the possibility that ignoring one kind of error can give overly narrow confidence intervals (e.g., King et al. (2015)).

Here, we focused on the estimates of $\mathcal{R}_0$ that were published within a very short frame of time (January 24th–26th). Although our analysis only reflects a snapshot of a fast-moving epidemic, our lessons will 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 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 framework for inferring generation-interval distributions from such data (Britton and Scalia Tomba, 2019). 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. Ideally, the code should not rely on closed-source programs.

We have provided a basis for evaluating and comparing exponential-growth based estimates of $\mathcal{R}_0$ in terms of three simple components: the exponential growth rate, mean generation interval, and generation interval dispersion. We are hopeful that this will provide a guide to understanding and reconciling different estimates early in an epidemic.

4 Methods

4.1 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 (Mills et al., 2004; Nishiura et al., 2009; Ma et al., 2014). Given an estimate of the exponential growth rate r and an *intrinsic* generation-interval distribution $g(\tau)$ (Champredon and Dushoff, 2015), the basic reproductive number can be estimated via the Euler-Lotka equation (Wallinga and Lipsitch, 2007):

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

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 (McBryde et al., 2009; Nishiura et al., 2009; Roberts and Nishiura, 2011; Park et al., 2019) to (1) characterize the amount of uncertainty present in the exponential growth rates and the shape of the generation-interval distribution and (2) 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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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.

4.2 Description of the studies

We reviewed 6 modeling studies of the novel coronavirus outbreak that were published online between January 24th, 2020 and January 26th, 2020 (Table 1). Five studies (Liu et al., 2020; Majumder and Mandl, 2020; Read et al., 2020; Riou and Althaus, 2020; Zhao et al., 2020) were uploaded to pre-print servers (bioRxiv, medRxiv, and SSRN), and one report was posted on the website of Imperial College London (Imai et al., 2020). There is a wide variation in their statistical methods and the amount of data they used to infer $\mathcal{R}_0$. Imai et al. (2020) and Riou and Althaus (2020) simulated branching process models and compared the predicted number of cases from their models with the estimated number of total cases by January 18th. Read et al. (2020) fitted a deterministic, metapopulation Susceptible-Exposed-Infected-Recovered (SEIR) model to incidence data between January 1st and January 21st from major cities in China and other countries. Zhao et al. (2020) and Liu et al. (2020) fitted exponential growth models to incidence data up to January 22nd and January 23rd, respectively, and inferred $\mathcal{R}_0$ via the Euler-Lotka equation (Eq. 1). Majumder and Mandl (2020) fitted the Incidence Decay and Exponential Adjustment (IDEA) model (Fisman et al., 2013) to incidence data up to January 26th, which is equivalent to fitting an exponential growth model and assuming a fixed generation-interval distribution.

4.3 Statistical framework

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 calculating 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:

$$\begin{aligned} r_i &\sim \text{Gamma}(\text{mean} = \mu_r, \text{shape} = \alpha_r), \\ \bar{G}_i &\sim \text{Gamma}(\text{mean} = \mu_G, \text{shape} = \alpha_G), \\ \kappa_i &\sim \text{Gamma}(\text{mean} = \mu_\kappa, \text{shape} = \alpha_\kappa). \end{aligned} \tag{4}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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; this approach is analogous to Bayesian methods for analyzing phylogenetic data, which often rely on drawing random samples of phylogenetic trees from a discrete set to account for phylogenetic uncertainty (Pagel et al., 2004; Bedford et al., 2014). Since the gamma distribution does not allow zeros, we use $\kappa = 0.02$ instead for Study 6. We note that this approach does not account for non-independence between the parameter estimates made by different modelers. As we add more models, we expect the pooled estimates to become sharper even when the models no longer add more information. Thus, the pooled estimator should be interpreted with care.

Weakly informative priors are assumed on the hyperparameters:

$$\begin{aligned} \mu_r &\sim \text{Gamma}(\text{mean} = 1\,\text{week}^{-1}, \text{shape} = 0.1) \\ \mu_G &\sim \text{Gamma}(\text{mean} = 1\,\text{week}, \text{shape} = 0.1) \\ \mu_\kappa &\sim \text{Gamma}(\text{mean} = 0.5, \text{shape} = 0.1) \\ (\alpha_r, \alpha_G, \alpha_\kappa) &\sim \text{Gamma}(\text{mean} = 1, \text{shape} = 0.1). \end{aligned} \tag{5}$$

We run 4 parallel Markov Chain Monte Carlo (MCMC) chains that consist of 200,000 burnin steps and 200,000 sampling steps. Posterior samples are thinned every 400 steps. Convergence is assessed by ensuring that the Gelman-Rubin statistic is below 1.01 for all hyperparameters (Gelman et al., 1992). 95% confidence intervals are calculated by taking 2.5% and 97.5% quantiles from the posterior distribution. `R` code is available in GitHub (https://github.com/parksw3/nCoV_framework).

Acknowledgements

We thank Ben Bolker for providing useful comments on the manuscript.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

References

Anderson, R. M. and R. M. May (1991). *Infectious diseases of humans: dynamics and control*. Oxford university press.

Bedford, T., M. A. Suchard, P. Lemey, G. Dudas, V. Gregory, A. J. Hay, J. W. McCauley, C. A. Russell, D. J. Smith, and A. Rambaut (2014). Integrating influenza antigenic dynamics with molecular evolution. *eLife 3*, e01914.

Britton, T. and G. Scalia Tomba (2019). Estimation in emerging epidemics: Biases and remedies. *Journal of the Royal Society Interface 16*(150), 20180670.

Centers for Disease Control and Prevention (2020). 2019 Novel Coronavirus (2019-nCoV), Wuhan, China. https://www.cdc.gov/coronavirus/2019-ncov/summary.html. Accessed 29, January, 2020.

Champredon, D. and J. Dushoff (2015). Intrinsic and realized generation intervals in infectious-disease transmission. *Proceedings of the Royal Society B: Biological Sciences 282*(1821), 20152026.

Diekmann, O., J. A. P. Heesterbeek, and J. A. Metz (1990). On the definition and the computation of the basic reproduction ratio $\mathcal{R}_0$ in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology 28*(4), 365–382.

Elderd, B. D., V. M. Dukic, and G. Dwyer (2006, October). Uncertainty in predictions of disease spread and public health responses to bioterrorism and emerging diseases. *Proceedings of the National Academy of Sciences 103*(42), 15693 –15697.

Fisman, D. N., T. S. Hauck, A. R. Tuite, and A. L. Greer (2013). An IDEA for short term outbreak projection: nearcasting using the basic reproduction number. *PloS one 8*(12).

Gelman, A., D. B. Rubin, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Huang, C., Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*.

Imai, N., A. Cori, I. Dorigatti, M. Baguelin, C. A. Donelly, S. Riley, and N. M. Ferguson (2020). Report 3: Transmissibility of 2019-nCoV. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-2019-nCoV-transmissibility.pdf. Accessed 26, January, 2020.

King, A. A., M. Domenech de Cellès, F. M. Magpantay, and P. Rohani (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences 282*(1806), 20150347.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Liu, T., J. Hu, M. Kang, L. Lin, H. Zhong, J. Xiao, G. He, T. Song, Q. Huang, Z. Rong, A. Deng, W. Zeng, X. Tan, S. Zeng, Z. Zhu, J. Li, D. Wan, J. Lu, H. Deng, J. He, and W. Ma (2020). Transmission dynamics of 2019 novel coronavirus (2019-nCoV). https://www.biorxiv.org/content/10.1101/2020.01.25.919787v1. Accessed 27, January, 2020.

Ma, J., J. Dushoff, B. M. Bolker, and D. J. Earn (2014). Estimating initial epidemic growth rates. *Bulletin of Mathematical Biology 76*(1), 245–260.

Ma, J. and D. J. Earn (2006). Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology 68*(3), 679–702.

Majumder, M. and K. D. Mandl (2020). Early transmissibility assessment of a novel coronavirus in Wuhan, China. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3524675. Accessed 27, January, 2020.

McBryde, E., I. Bergeri, C. van Gemert, J. Rotty, E. Headley, K. Simpson, R. Lester, M. Hellard, and J. E. Fielding (2009). Early transmission characteristics of influenza A (H1N1) v in Australia: Victorian state, 16 May–3 June 2009. *Eurosurveillance 14*(42), 19363.

Mills, C. E., J. M. Robins, and M. Lipsitch (2004). Transmissibility of 1918 pandemic influenza. *Nature 432*(7019), 904.

Nishiura, H., C. Castillo-Chavez, M. Safan, and G. Chowell (2009). Transmission potential of the new influenza A (H1N1) virus and its age-specificity in Japan. *Eurosurveillance 14*(22), 19227.

Pagel, M., A. Meade, and D. Barker (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology 53*(5), 673–684.

Park, S. W., D. Champredon, J. S. Weitz, and J. Dushoff (2019). A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics 27*, 12–18.

Read, J. M., J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell (2020). Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. https://www.medrxiv.org/content/10.1101/2020.01.23.20018549v1. Accessed 26, January, 2020.

Riou, J. and C. L. Althaus (2020). Pattern of early human-to-human transmission of wuhan 2019-nCoV. https://www.biorxiv.org/content/10.1101/2020.01.23.917351v1. Accessed 26, January, 2020.

Roberts, M. and J. Heesterbeek (2007). Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection. *Journal of mathematical biology 55*(5-6), 803.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.30.20019877; this version posted February 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Roberts, M. G. and H. Nishiura (2011). Early estimation of the reproduction number in the presence of imported cases: pandemic influenza H1N1-2009 in New Zealand. *PLoS One 6*(5).

Svensson, Å. (2007). A note on generation times in epidemic models. *Mathematical biosciences 208*(1), 300–311.

Wallinga, J. and M. Lipsitch (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences 274*(1609), 599–604.

Wearing, H. J., P. Rohani, and M. J. Keeling (2005). Appropriate models for the management of infectious diseases. *PLoS medicine 2*(7).

World Health Organization (2020a). Novel Coronavirus (2019-nCoV) Situation Report - 6. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200126-sitrep-6-2019–ncov.pdf?sfvrsn=beaeee0c_4. Accessed January 26, 2020.

World Health Organization (2020b). Novel Coronavirus (2019-nCoV) Situation Report - 8. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200128-sitrep-8-ncov-cleared.pdf?sfvrsn=8b671ce5_2. Accessed January 28, 2020.

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

World Health Organization (2020d). Statement on the second meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-ncov). https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov). Accessed January 30, 2020.

Zhao, S., J. Ran, S. S. Musa, G. Yang, Y. Lou, D. Gao, L. Yang, and D. He (2020). 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. https://www.biorxiv.org/content/10.1101/2020.01.23.916395v1. Accessed 26, January, 2020.