Exploring how generation intervals link strength and speed of epidemics

Sang Woo Park David Champredon Joshua Weitz
Jonathan Dushoff

July 2017

Abstract

1 Introduction

Infectious disease research often focuses on estimating the reproductive number, i.e., the number of new infections caused on average by a single infection. This number is termed the basic reproductive number – $\mathcal{R}_0$ – in the event that a single infection emerges in an otherwise susceptible population. The basic reproductive number provides information about the disease's potential for spread and the difficulty of control. It is described in terms of an average [2] or an appropriate sort of weighted average [5].

The reproductive number has remained a focal point for research because it provides information about how a disease spreads in a population, on the scale of disease generations. As it is a unitless quantity, it does not, however, contain information about *time*. Hence, another important quantity is the population-level *rate of spread*, r. The initial rate of spread can often be measured robustly early in an epidemic, since the number of incident cases at time t is expected to follow $i(t) \approx i(0)\exp(rt)$. The rate of growth can also be described using the "characteristic time" of exponential growth $C = 1/r$. This is closely related to the doubling time (given by $T_2 = \ln(2)C \approx 0.69C$).

In disease outbreaks, the rate of spread, r, is often inferred from case-incidence reports, by fitting an exponential function to the incidence curve [13, 16, 12]. Estimate of the initial exponential rate of spread, r, can then

be combined with a mechanistic model that includes unobserved features of the disease to esimate the initial reproductive number, $\mathcal{R}$. In particular, $\mathcal{R}$ is often calculated from r and the generation-interval distribution using the generating function approach popularized by [22].

While the rate of spread measures the speed of the disease at the population level, the generation interval measures speed at the individual level. The *generation interval* denotes the time that elapses between when an individual is infected by an infector, and the time that the infector was infected [18]. Generation interval distributions are typically inferred from contact tracing, sometimes in combination with clinical data [3, 10, 7]. As a consequence, generation interval distributions are difficult to ascertain empirically. Further, the generation function approach depends on entire distribution – which makes it difficult to ascertain which features of the distributions are essential to connect measurements of the rate of spread r, with the reproductive number, $\mathcal{R}$.

Here, we re-interpret the work of [22] using means, variance measures and approximations. By doing so, we hope to shed light on the underpinnings of the relationship between r and $\mathcal{R}$, and to shed light on the factors underlying its robustness and its practical use even when data on generation intervals is limited or hard to obtain.

2 Relating $\mathcal{R}$ and r

We are interested in the relationship between r, $\mathcal{R}$ and the generation-interval distribution, which describes the interval between the time an individual becomes *infected* and the time that they *infect* another individual. This distribution links r and $\mathcal{R}$. In particular, if $\mathcal{R}$ is known, a shorter generation interval means a faster epidemic (larger r). Conversely, and somewhat counter-intuitively, if r is known, then faster disease generations imply a *lower* value of $\mathcal{R}$, because more *individual* generations are required to realize the same *population* spread of disease (see Fig. 1).

We define the generation-interval distribution using a renewal-equation approach. A wide range of disease models can be described using this model:

$$i(t) = S(t) \int K(s) i(t-s)\, ds, \tag{1}$$

where t is time, $i(t)$ is the incidence of new infections, $S(t)$ is the *proportion* of the population susceptible, and $K(s)$ is the intrinsic infectiousness of

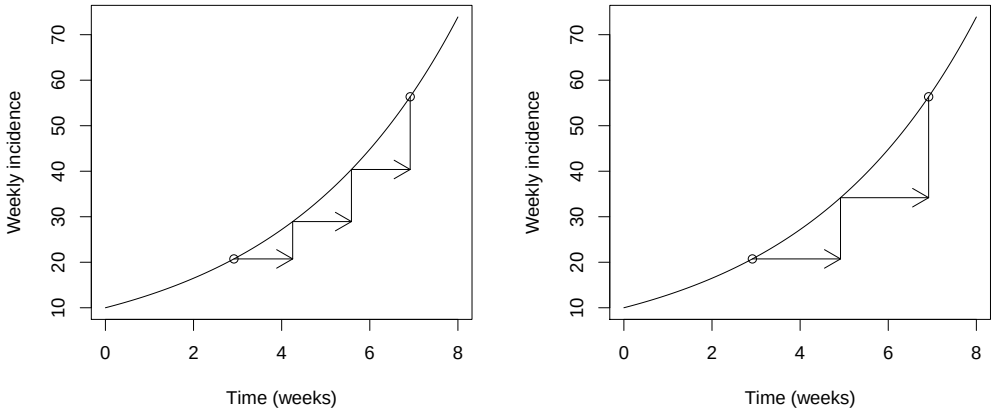

Figure 1: Two hypothetical epidemics with the same growth rate ($r = 0.25\text{weeks}^{-1}$) and fixed generation intervals. Assuming a short generation interval (fast transmission at the individual level) implies a smaller value for the reproduction number $\mathcal{R}_0$ (panel A) when compared to a longer generation interval (slow transmission at the individual level, panel B).

individuals who have been infected for a length of time s.

We then have the basic reproductive number:

$$\mathcal{R}_0 = \int K(s)ds, \tag{2}$$

and the *intrinsic* generation-interval distribution:

$$g(s) = \frac{K(s)}{\mathcal{R}_0}. \tag{3}$$

The "intrinsic" interval can be distinguished from "realized" intervals, which can look "forward" or "backward" in time [4] (see also earlier work [18, 14]).

Disease growth is predicted to be approximately exponential in the early phase, because the depletion in the effective number of susceptibles is relatively small. Thus, for the exponential phase, we write:

$$i(t) = \mathcal{R} \int g(\tau) i(t-\tau)\, d\tau, \tag{4}$$

where $\mathcal{R} = \mathcal{R}_0 S$.

We then solve for the characteristic time C by assuming that the population is growing exponentially: i.e., substitute $i(t) = i(0)\exp(t/C)$ to obtain

$$1/\mathcal{R} = \int g(\tau)\exp(-\tau/C)\, d\tau. \tag{5}$$

3 Approximation framework

3.1 Approximation method, theory

Here, we propose an approximation approach to estimate $\mathcal{R}$ from (5). The approximation approach was inpired by the work of [22] who used a normal approximation to construct such a moment approximation. Instead, we follow [16] and approximate the generation interval with a gamma distribution. This provides a more robust framework, based on more realistic assumptions, since the gamma distribution is confined to non-negative values. For biological interpretability, we describe the distribution using the mean $\bar{G}$ and the squared coefficient of variation κ (thus $\kappa = 1/a$, and $\bar{G} = \theta/\kappa$, where a and θ are the shape and scale parameters under the standard parameterization of

the gamma distribution). Substituting the gamma distribution, $g_\gamma(\tau)$, into (5) then yields:

$$\mathcal{R} \approx (1 + \kappa r \bar{G})^{1/\kappa}. \tag{6}$$

We try to interpret this equation further by writing:

$$\mathcal{R} \approx (1 + \kappa\rho)^{1/\kappa} \equiv X(\rho; 1/\kappa), \tag{7}$$

where $\rho = \bar{G}/C = r\bar{G}$ measures how fast the epidemic is growing (on the time scale of the mean generation interval) – or equivalently, the length of the mean generation interval (in units of the characteristic time of exponential growth). The longer the generation interval is compared to T_c, the higher the estimate of $\mathcal{R}$ (see Fig. 1). We define the generalized exponential function X above – it is equivalent to the Tsallis "q-exponential", with $q = 1 - \kappa$ [21] – its shape determines how the estimate of $\mathcal{R}$ changes with the estimate of normalized generation length ρ.

For small ρ, X always looks like $1 + \rho$, regardless of the shape parameter $1/\kappa$, which determines the curvature: if $1/\kappa = 1$, we get a straight line, for $1/\kappa = 2$ the curve is quadratic, and so on (see Fig. 2). For large values of $1/\kappa$, X looks like the "compound-interest approximation" to the exponential; and when $\kappa \to 0$, $X(\rho)$ converges to $\exp(\rho)$.

The limit as $\kappa \to 0$ is reasonably easy to interpret. The incidence is increasing by a factor of $\exp(\rho)$ in the time it takes for an average disease generation. If the generation interval is fixed, then this means the average case must cause $\mathcal{R} = \exp(\rho)$ new cases. If variation in the generation time (i.e., κ) increases, then some new cases will be produced before, and some after, the mean generation time. Since we assume the disease is increasing exponentially, infections that occur early on represent a larger proportion of the population, and thus will have a disproportionate effect: individuals don't have to produce as many lifetime infections to sustain the growth rate, and thus we expect $\mathcal{R} < \exp(\rho)$.

The straight-line relationship for $\kappa = 1$ also has a biological interpretation. In our approximation, this corresponds to a generation distribution that is approximated by an exponential distribution. In this case, recovery rate and infection rate are constant for each individual. The rate of exponential growth per generation is then given directly by the net per capita increase in infections: $\mathcal{R} - 1$, where one represents the recovery of an infectious individual.

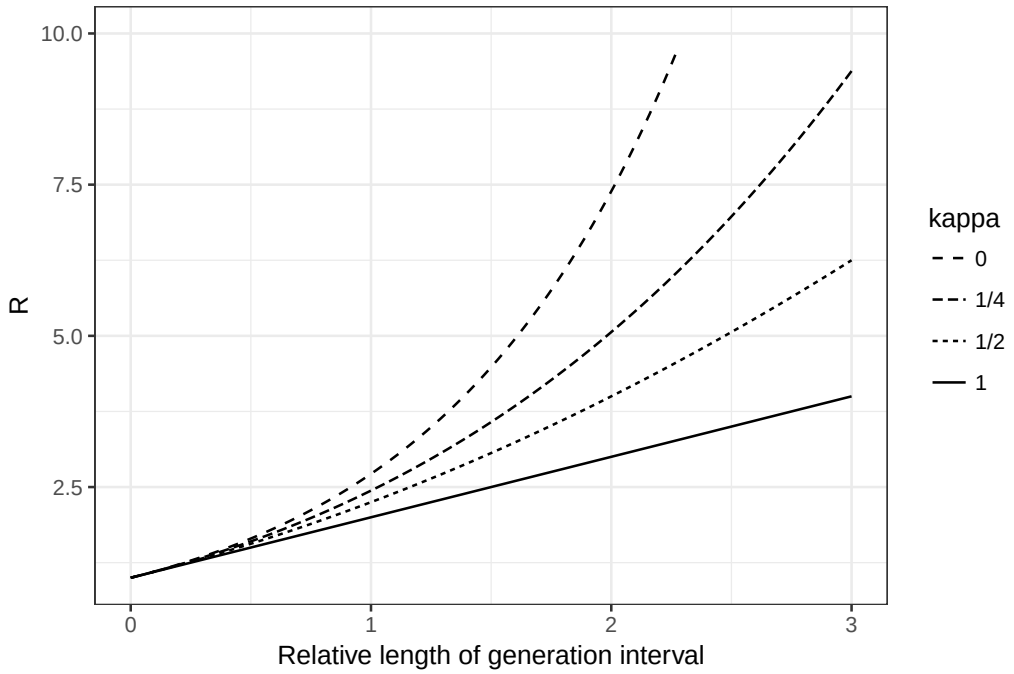

Figure 2: The approximate relationship ((6)) between mean generation time (relative to the characteristic time of exponential growth) and reproductive number, as for different amounts of variation in the generation-interval distribution. Recall that the relative length of generation interval ρ is the ratio of mean generation time $\hat{G}$ to characteristic time $C = 1/r$.

3.2 Approximation method, in practice

To apply our approximation method, we need an estimated distribution of generation intervals. This may be based directly on empirical observations of generation intervals, or inferred from observations of (or estimated parameters from) latent and infectious periods. Given the estimated distribution of generation intervals, we can obtain the gamma approximation by estimating moments from data and using (6); these moments can be calculated directly, or estimated using a maximum-likelihood fit.

To generate an estimated distribution of generation intervals from information about latent and infectious periods, we use a simulation method [6]: briefly, we generate deviates from each distribution independently (by bootstrap sampling for empirical observations, or using quantiles for estimated distributions). We then sample an infection delay uniformly from the infectious period (i.e., assuming constant infectiousness) and estimate generation interval as the sum of latent period and infection delay. For our examples, we drew 10000 values from each distribution.

The estimated relationship between exponential growth rate and reproductive number is obtained by substituting the simulated generation interval sample into (5). As (5) is a weighted mean of $\exp(-\tau/C)$ over a continuous distribution, equivalent expression given discrete samples is then:

$$1/\mathcal{R} = \frac{1}{N}\sum_{i=1}^{N}\exp(-x_i/C), \tag{8}$$

where x_i represents each sample generation interval and N is the total number of samples. This is then compared to the approximate relationship based on (6)

4 Examples

We investigate this approximation approach using three different examples. In doing so, we demonstrate the practicality of using our gamma approximation to estimate the basic reproductive number. These examples also serve to demonstrate that robust estimates could be made with less data and potentially earlier in an outbreak – a point we revisit in the Discussion. Our initial investigation of this question was motivated by the West African Ebola Outbreak [23], so we start with that example. To probe the approximation

more thoroughly, we also chose one disease with high variation in generation interval (canine rabies), and one with a high reproductive number (measles). For simplicity, we assumed that latent and infectious periods are equivalent to incubation and symptomatic periods for EVD and canine rabies.

4.1 Ebola

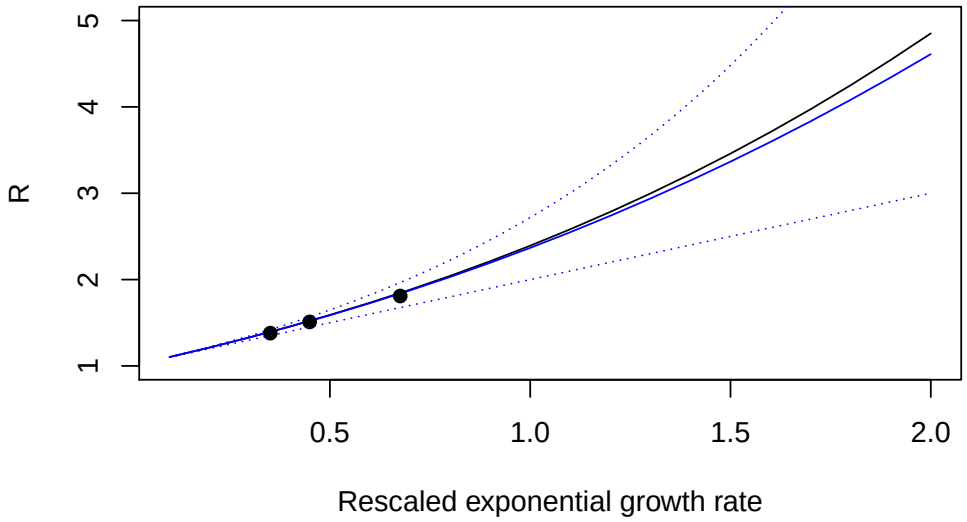

Figure 3: Estimating $\mathcal{R}$ from Ebola infectious case data. (black curve) the relationship between growth rate and $\mathcal{R}$ using a realistic generation-interval distribution based on [3]. (blue curve) the same relationship, approximated using the observed mean and CV. The blue dotted curves show the approximations based on exponential (lower) and fixed (upper) generation distributions. Points indicate estimates for the three focal countries of the West African Ebola Outbreak calculated by [3]: Sierra Leone (square, $\mathcal{R} = 1.38$), Liberia (triangle, $\mathcal{R} = 1.51$), and Guinea (circle, $\mathcal{R} = 1.81$). Initial growth rate for each outbreak was inferred from doubling periods reported by [3] ($r = \ln(2)/T_2$).

We first simulated a generation-interval distribution for Ebola virus disease (EVD) using information from [3] and a lognormal assumption for both the incubation and infectious periods. We used lognormal (rather than gamma, the other simple alternative) for our components because we thought this would provide a more challenging test of our gamma approximation. We used the reported standard deviation for the infectious period, and chose the standard deviation for the incubation period to match the reported standard deviation for the serial interval distribution, since this value is available and is expected to be similar to the generation interval distribution for EVD [3]. We then simulated the relationship between r and $\mathcal{R}$ implied by our simulated distribution (8), and the approximate relationship (7) based only on the mean and CV (see Fig. 3). The approximation appears to work well over relevant parameter ranges, implying that it may be sufficient to understand the mean and CV of the generation-interval distribution when investigating this relationship.

4.2 Measles

To test whether gamma moment matching works for high $\mathcal{R}$ values, we applied the moment approximation to a simulated generation intervals distribution using information from [9] and [11]. Here, we found that the approximation matches the theoretical distribution almost perfectly (no visible differences in the curves) across our range of interest ($\mathcal{R}$ up to > 20). Although infectious period distributions was inferred from a modelling study, this result is not sensitive to our estimate of the variation in infectious period: because the length of infectious period is much shorter than that of latent period, changing the variation in the infectious period distribution has little effect on $\mathcal{R}$ (results not shown).

4.3 Rabies

We did a similar analysis for rabies, and found that approximation is generally harder for this high-variance case (see Fig. 5). Since rabies estimates point to a value of $\mathcal{R}$ near 1, results are not very sensitive to any tested assumption about the relationship. But, looking at the relationship more broadly, we see that the moment-based approximation would do a poor job of predicting the relationship for intermediate or large values of $\mathcal{R}$ – in fact,

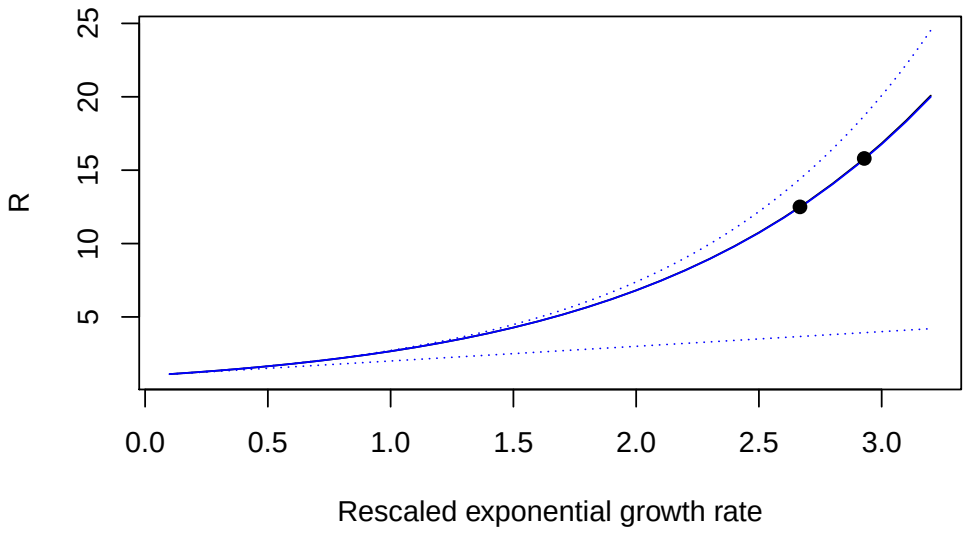

Figure 4: Estimating $\mathcal{R}$ from measles data. (solid curve) the relationship between growth rate and $\mathcal{R}$ using a realistic generation-interval distribution. (dashed curve) the same relationship approximated using the estimated mean and CV (this curve is almost invisible because it overlaps the solid curve) The blue dotted curves show the approximations based on exponential (lower) and fixed (upper) generation distributions.

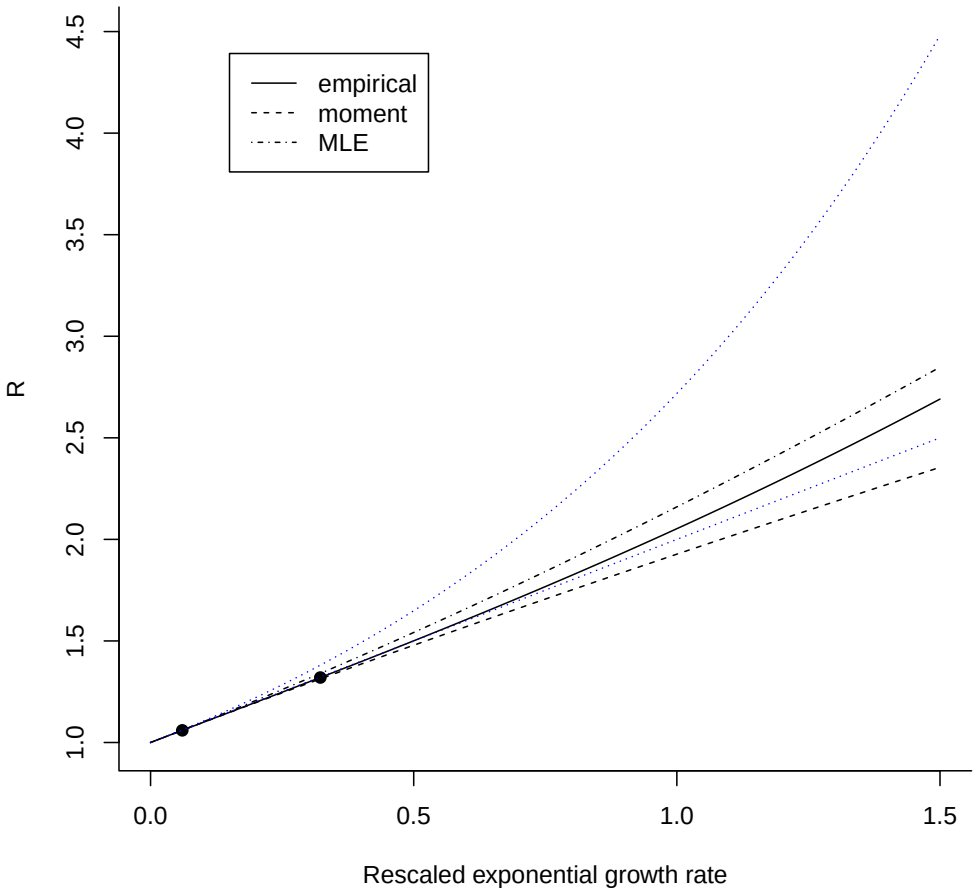

Figure 5: Estimating $\mathcal{R}$ from rabies infectious case data. (solid curve) the relationship between growth rate and $\mathcal{R}$ using a realistic generation-interval distribution. (dashed curve) the same relationship approximated using the observed mean and CV. (dash-dotted curve) the same relationship approximated using the mean and CV calculated from a maximum-likelihood fit. The blue dotted curves show the approximations based on exponential (lower) and fixed (upper) generation distributions. Points are estimates from [6]: Serengeti (circle, $\mathcal{R} = 1.06$) and Ngorongoro (triangle, $\mathcal{R} = 1.32$).

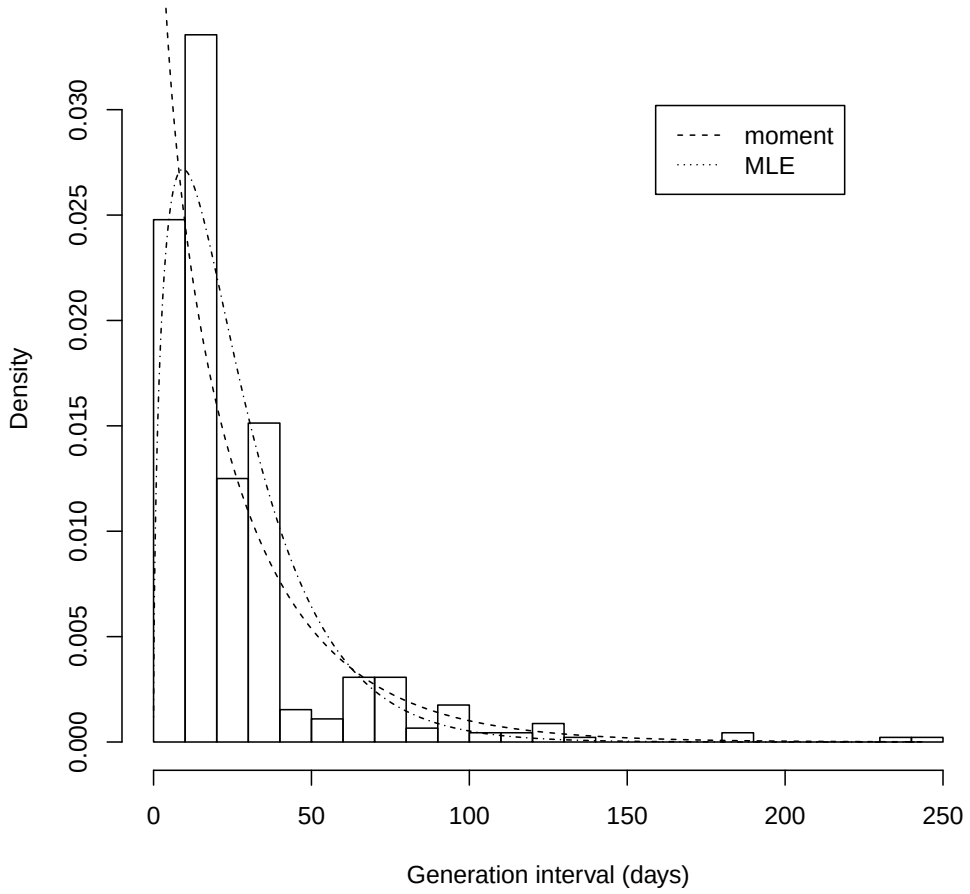

Figure 6: Fitting gamma distributions to generation intervals. Rabies generation distributions simulated from incubation and infectious periods observed by [6], with gamma approximations based on moments (dashed curve) and a maximum-likelihood fit (dotted curve). Initial growth rate for each outbreak was inferred by solving (5) numerically using the simulated generation interval distributions.

a poorer job than if we use the approximation based on exponentially distributed generation times.

The reason for this can be seen in Fig. 6. The moment approximation is strongly influenced by rare, very long generation intervals, and does a poor job of matching the observed pattern of short generation intervals. Short intervals will be much more important in driving the speed of the epidemic, and therefore in determining the relationship between r and $\mathcal{R}$. We can address this problem by estimating gamma parameters formally using a maximum-likelihood fit to the generation intervals we simulated based on observations. This fit does a better job of matching the observed pattern of short generation intervals (Fig. 6) and of predicting the simulated relationship between r and $\mathcal{R}$across a broad range (Fig. 5).

Disease	Ebola	Measles	Rabies
Parameter	Values		
Reproduction number	1.38, 1.51, 1.81 [3]	NA	1.06, 1.32 [6]
Mean incubation period (days)	11.4 [3]	12.77 [9]	24.24 [6]
SD incubation period (days)	7.5 (see Sec. 4.1)	2.67 [9]	29.49 [6]
Mean infectious period (days)	5 [3]	3.65 [11]	3.57 [6]
SD infectious period (days)	4.7 [3]	1.63 [11]	2.26 [6]

Table 1: Parameters that were used to obtain theoretical generation distributions for each disease. Reproduction numbers are represented as points in figure Fig. 3–5.

5 Discussion

Estimating the reproductive number $\mathcal{R}$ is a key part of characterizing and controlling infectious disease spread. The initial value of $\mathcal{R}$ for an outbreak is often estimated by estimating the initial exponential rate of growth, and then using a generation-interval distribution to relate the two quantities. Much work has been done in exploring such links [22, 18, 14, 19]. However, detailed estimates of the full generation interval are difficult to obtain, and the link between uncertainty in the generation interval and uncertainty in estimates of $\mathcal{R}$ are often unclear. Here we discuss a simple framework for *estimating* the relationship between $\mathcal{R}$ and r, using only the estimated mean and CV of the generation interval. The framework is based on the gamma distribution.

We use examples to test the robustness of the framework. We also compare estimates based directly on estimated mean and variance of of the generation interval to estimates based on maximum-likelihood fits.

The gamma approximation was developed by [16], and provides estimates that are simpler, more robust and more realistic than those from the normal approximations developed by [22] (see Appendix). Here, we attempted to present the gamma approximation in a form conducive to intuitive understanding of the relationship between speed, r, and strength, $\mathcal{R}$ (See Fig. 2). We discuss the general result that estimates of $\mathcal{R}$ increase with mean generation, but decrease with *variation* in generation times [22]. We also provide mechanistic interpretations: when generation intervals are slower, more infection is needed per generation (larger $\mathcal{R}$) in order to produce a given rate of increase r. Similarly, when variance in generation time is low, there is less early infection, and thus slower exponential growth, also meaning that a larger $\mathcal{R}$ is needed.

We tested the gamma approximation framework by applying it to parameter regimes based on three diseases: Ebola, measles, and rabies. We found that approximation based on observed moments gives good estimates when the generation-interval distribution is not too broad (as is the case for Ebola and measles, but not for rabies, see Table 1), but that using maximum likelihood to estimate the moments provides better estimates for a broader range of parameters, and also when data are limited (see Appendix).

Our key finding is that summarizing an entire generation interval distribution with its moments can give sensible and robust estimates. This framework has potential advantages for understanding the likely effects of parameter changes, and also for parameter estimation with uncertainty: since $\mathcal{R}$ can be estimated from three simple quantities ($\bar{G}$, κ and r), it should be straightforward to propagate uncertainty from estimates of these quantities to estimates of $\mathcal{R}$.

For example, during the Ebola outbreak in West Africa, many researchers tried to estimate $\mathcal{R}$ from r [1, 3, 15, 17, 8] but uncertainty in the generation-interval distribution was often neglected (but see [20]). During the outbreak, [23] used a generation-interval argument to show that neglecting the effects of post-burial transmission would be expected to lead to underestimates of $\mathcal{R}$. Our generation interval framework provides a clear interpretation of this result: as long as post-burial transmission tends to increase generation intervals, it should result in higher estimates of $\mathcal{R}$ for a given estimate of r. Knowing the exact shape of the generation interval distribution is difficult,

but thinking about how various transmission routes and epidemic parameters affect the distribution will help researchers better understand future outbreaks.

Acknowledgments

References

[1] C. L. Althaus. "Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa." In: *PLoS Curr* 6 (2014 Sep 02), 2014 Sep 02.

[2] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press, 1991.

[3] B. Aylward et al. "Ebola virus disease in West Africa–the first 9 months of the epidemic and forward projections." In: *N Engl J Med* 371 (2014 Oct 16), pp. 1481–95.

[4] D. Champredon and J. Dushoff. "Intrinsic and realized generation intervals in infectious-disease transmission." In: *Proc Biol Sci* 282 (2015 Dec 22), p. 20152026.

[5] O. Diekmann, J. A. Heesterbeek, and J. A. Metz. "On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations." In: *J Math Biol* 28 (1990), pp. 365–82.

[6] K. Hampson et al. "Transmission dynamics and prospects for the elimination of canine rabies." In: *PLoS Biol* 7 (2009 Mar 10), e53.

[7] J. H. Huber et al. "Quantitative, model-based estimates of variability in the generation and serial intervals of Plasmodium falciparum malaria." In: *Malar J* 15 (2016 Sep 22), p. 490.

[8] A. A. King et al. "Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola". In: *Proceedings of the Royal Society B: Biological Sciences* 282 (2015), p. 2015.

[9] J. Lessler et al. "Incubation periods of acute respiratory viral infections: a systematic review." In: *Lancet Infect Dis* 9 (2009 May), pp. 291–300.

[10] J. Lessler et al. “Times to key events in Zika virus infection and implications for blood donation: a systematic review.” In: *Bull World Health Organ* 94 (2016 Nov 01), pp. 841–849.

[11] A. L. Lloyd. “Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics.” In: *Theor Popul Biol* 60 (2001 Aug), pp. 59–71.

[12] J. Ma et al. “Estimating initial epidemic growth rates.” In: *Bull Math Biol* 76 (2014 Jan), pp. 245–60.

[13] C. E. Mills, J. M. Robins, and M. Lipsitch. “Transmissibility of 1918 pandemic influenza.” In: *Nature* 432 (2004 Dec 16), pp. 904–6.

[14] H. Nishiura. “Time variations in the generation time of an infectious disease: Implications for sampling to appropriately quantify transmission potential”. In: *Mathematical Biosciences and Engineering* 7 (2010), p. 2010.

[15] H. Nishiura and G. Chowell. “Theoretical perspectives on the infectiousness of Ebola virus disease.” In: *Theor Biol Med Model* 12 (2015 Jan 06), p. 1.

[16] H. Nishiura et al. “Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan.” In: *Euro Surveill* 14 (2009 Jun 04), 2009 Jun 04.

[17] C. M. Rivers et al. “Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia.” In: *PLoS Curr* 6 (2014 Oct 16), 2014 Oct 16.

[18] A. Svensson. “A note on generation times in epidemic models.” In: *Math Biosci* 208 (2007 Jul), pp. 300–11.

[19] A. Svensson. “The influence of assumptions on generation time distributions in epidemic models.” In: *Math Biosci* 270 (2015 Dec), pp. 81–9.

[20] B. P. Taylor, J. Dushoff, and J. S. Weitz. “Stochasticity and the limits to confidence when estimating R0 of Ebola and other emerging infectious diseases.” In: *J Theor Biol* 408 (2016 Nov 07), pp. 145–54.

[21] Constantino Tsallis. “What are the numbers that experiments provide”. In: *Quimica Nova* 17.6 (1994), pp. 468–471.

[22] J. Wallinga and M. Lipsitch. "How generation intervals shape the relationship between growth rates and reproductive numbers." In: *Proc Biol Sci* 274 (2007 Feb 22), pp. 599–604.

[23] J. S. Weitz and J. Dushoff. "Modeling post-death transmission of Ebola: challenges for inference and opportunities for control." In: *Sci Rep* 5 (2015 Mar 04), p. 8751.

6 Appendix

6.1 Normal approximation

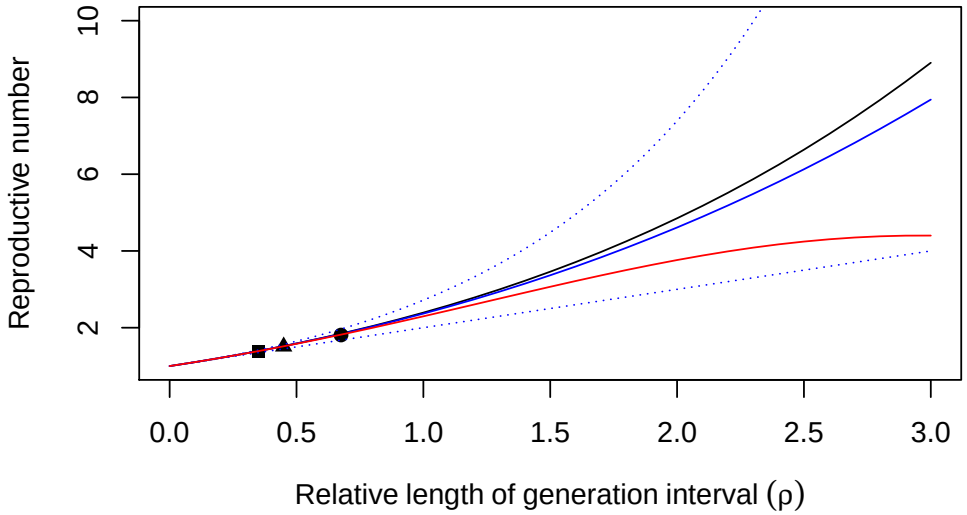

Figure 7: Approximating a generation intervals distribution with a normal distribution can be problematic as it supports negative values, which are biologically impossible. As a consequence, the normal approximation (red curve) eventually predicts a decreasing $r - \mathcal{R}$ relationship for large r.

6.2 Robustness of the gamma approximation

The moment-matching method (approximating $\mathcal{R}$ based on estimated mean and variance of the generation interval) has an appealing simplicity, and works well for all of the actual disease parameters we tested (the breakdown for rabies distributions occurs for values of $\mathcal{R}$ well above observed values). We therefore wanted to compare its robustness in statistical estimation to that of the more sophisticated maximum likelihood method. Fig. 8 shows results of this experiment. When sample size is limited, estimates using MLE

tend to be substantially close to the known true values in these experiments. However, it is worth noting that using the observed moment gives narrower estimates than the two naive estimates.

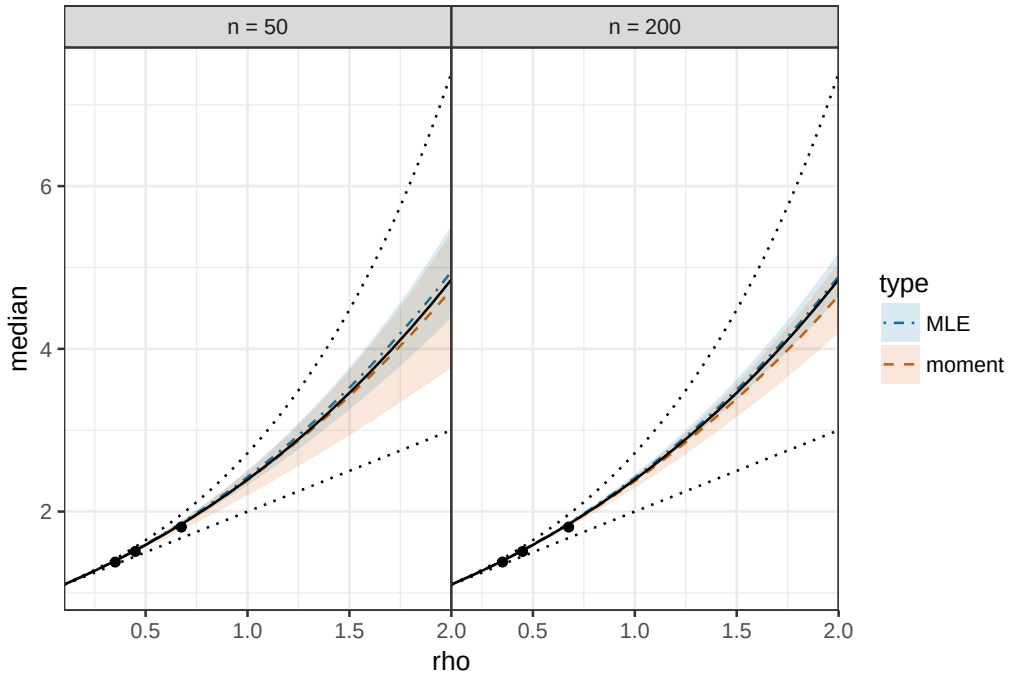

Figure 8: The effect of sample size on estimates of $\mathcal{R}$. (black) the relationship between growth rate and $\mathcal{R}$ using a known generation-interval distribution (see Fig. 3). (colors) estimates based on finite samples from this distribution: solid curves show the median of 1000 sampling experiments, and shading shows the range where 95% of the results fall. Blue shows estimates based on estimated mean and CV. Red shows estimates based on maximum likelihood fits.