A practical generation interval-based approach to inferring the strength of epidemics from their speed

Sang Woo Park[1] David Champredon[2,3] Joshua S. Weitz[4,5] Jonathan Dushoff [2,*]

1 Department of Mathematics & Statistics, McMaster University, Hamilton, Ontario, Canada
2 Department of Biology, McMaster University, Hamilton, Ontario, Canada
3 Department of Mathematics & Statistics, Agent-Based Modelling Laboratory, York University, Toronto, Ontario, Canada
4 School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, United States of America
5 School of Physics, Georgia Institute of Technology, Atlanta, Georgia, United States of America

*dushoff@mcmaster.ca

Abstract

Infectious-disease outbreaks are often characterized by the reproductive number $\mathcal{R}$ and exponential rate of growth r. $\mathcal{R}$ provides information about outbreak control and predicted final size. Directly estimating $\mathcal{R}$ is difficult, while r can often be estimated from incidence data. These quantities are linked by the generation interval – the time between when an individual is infected by an infector, and when that infector was infected. It is often infeasible to obtain the exact shape of a generation-interval distribution, and to understand how this shape affects estimates of $\mathcal{R}$. We show that estimating generation interval mean and variance provides insight into the relationship between $\mathcal{R}$ and r. We use examples based on Ebola, rabies and measles to explore approximations based on gamma-distributed generation intervals, and find that use of these simple approximations are often sufficient to capture the r–$\mathcal{R}$ relationship and provide robust estimates of $\mathcal{R}$.

Keywords

Infectious disease modeling; Generation interval; basic reproductive number

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 reproductive number – $\mathcal{R}$. The reproductive number provides information about the disease's potential for spread and the difficulty of control. It is described in terms of an average [4] or an appropriate sort of weighted average [9].

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, and simpler mathematically than, the more commonly used doubling time (given by $T_2 = \ln(2)C \approx 0.69C$).

In disease outbreaks, the rate of spread, r, can be inferred from case-incidence reports, e.g., by fitting an exponential function to the incidence curve [27, 29, 24]. Estimates 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, $\mathcal{R}$ can be calculated from r and the generation-interval distribution using the generating function approach popularized by [44].

The *generation interval* is the amount of time between when an individual is infected by an infector, and the time that the infector was infected [39]. While r measures the speed of the disease at the population level, the generation interval measures speed at the individual level. Generation interval distributions are typically inferred from contact tracing, sometimes in combination with clinical data [7, 20, 16]. Generation interval distributions can be difficult to ascertain empirically [29, 8], and the generation-function approach depends on an entire distribution. Multiple studies have explored how general-interval distributions affect the r–$\mathcal{R}$ relationship (summarized in Table 1). However, the use of an entire distribution makes it difficult to determine which features of the distributions are essential to connect measurements of the rate of spread r, with the reproductive number, $\mathcal{R}$. Moreover, different assumptions about the shape of generation interval distributions

Method	Required information	Assumptions/Notes	References
Delta approximation	• Mean generation interval	• Assumes fixed generation interval • Provides an upper bound for $\mathcal{R}$ estimate	[45, 37, 26, 42]
Empirical (histogram-based)	• Generation interval samples	• Relies on binning generation intervals into discrete histograms	[45]
Empirical (sample-based)	• Generation interval samples	• Uses discretized Euler-Lotka equation	[14]
Euler-Lotka	• Entire generation interval distribution	• Uses moment generating function	[23]
Exponential approximation (SIR)	• Mean generation interval	• Assumes exponentially distributed generation interval • Predicts linear relationship between r and $\mathcal{R}$	[6, 33, 12, 45, 42]
Gamma approximation	• Mean/SD generation interval	• Assumes gamma distributed generation intervals	[29, 25, 42, 31]
Normal approximation	• Mean/SD generation interval	• Assumes normally distributed generation intervals • Predicts decreasing r–$\mathcal{R}$ relationship for high r	[45, 42]
SEIR (gamma)	• Mean/SD latent period	• Assumes gamma distributed latent an infectious periods • See [48] for summary of all special cases	[3, 46, 45, 37]
SEIR (generalized)	• Laplace transformation of latent period distribution • Laplace transformation of infectious period distribution	• Independent latent and infectious periods • Generalizes gamma distributed latent and infectious periods	[48]
Trapezoid approximation	• Period of no infection after exposure • Time until maxium infectivity • Duration of maximum infectivity • Time until recovery	• Assumes trapezoid shaped infection kernel	[37]

Table 1: Summary of previous studies that relate growth rate, r, with reproductive number, $\mathcal{R}$, using generation interval distributions.

have led to seemingly contradictory results [46, 37].

The primary goal of this paper is to further explore and explain the r–$\mathcal{R}$ relationship. We explore the qualitative relationship between generation time, initial rate of spread r, and initial reproductive number $\mathcal{R}$ using means, variance measures and gamma approximations. We show that summarizing generation-interval distributions using moments provides biological explanations that unify previous findings. We further discuss the generality of the gamma approximation and provide examples of its relevance in using realistic epidemiological parameters from previous outbreaks. By doing so, we shed light on the underpinnings of the relationship between r and $\mathcal{R}$, and on the factors underlying its robustness and its practical use when data on generation intervals is limited or hard to obtain.

2 Relating $\mathcal{R}$ and r

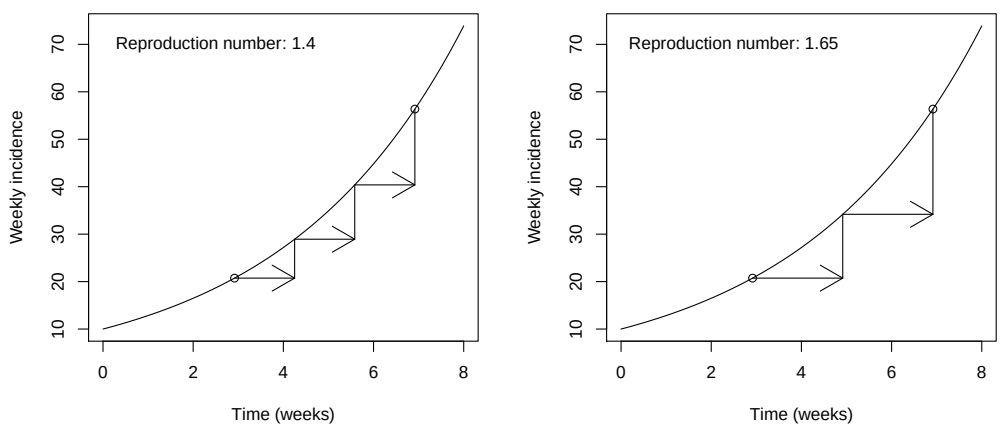

Figure 1: Two hypothetical epidemics with the same growth rate ($r = 0.25$/week) and fixed generation intervals. Assuming a short generation interval (fast transmission at the individual level) implies a smaller reproductive number $\mathcal{R}_0$ (panel A) when compared to a longer generation interval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

In this section, we introduce an analytical framewok, and recapitulate

previous work relating growth rate r and reproductive number $\mathcal{R}$. These two quantities are linked by the generation-interval distribution, which describes the interval between the time an individual becomes *infected* and the time that they *infect* another individual. In particular, if $\mathcal{R}$ is known, a shorter generation interval means a faster epidemic (larger r). Conversely (and perhaps counter-intuitively), if r is known, then shorter disease generations imply a *lower* value of $\mathcal{R}$, because more *individual* generations are required to realize the same *population* spread of disease [11, 32] (see Fig. 1).

The generation-interval distribution can be defined using a renewal-equation approach. A wide range of disease models can be described using the model [15, 10, 36, 1, 44, 37]:

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

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

The *basic* reproductive number, defined as the average number of secondary cases generated by a primary case in a fully susceptible population [4, 9], is

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

and the *intrinsic* generation-interval distribution is

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

The "intrinsic" interval can be distinguished from "realized" intervals, which can look "forward" or "backward" in time [8] (see also earlier work [39, 28]). In particular, it is important to correct for biases that shorten the intrinsic interval when generation intervals are observed through contact tracing during an outbreak.

In this model, disease growth is predicted to be approximately exponential in the early phase of an epidemic, 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 $\mathcal{R} = \mathcal{R}_0 S$ is the *effective* reproductive number. We can 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 the exact speed-strength relationship [10]:

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

This fundamental relationship dates back to the work of Euler and Lotka [23]. We will explore the shape of this relationship using parameters based on human infectious diseases, and investigate approximations based on gamma-distributed generation intervals.

3 Approximation framework

3.1 Approximation method, in theory

We do not expect to know the full distribution $g(\tau)$ – particularly while an epidemic is ongoing – so we are interested in approximations to $\mathcal{R}$ based on limited information. We follow the approach of [29] and approximate the generation interval with a gamma distribution. We prefer the gamma distribution to the standard normal approximation used in many applications for a number of reasons. First, it is more biologically realistic, since it is confined to non-negative values. Second, it has a convenient moment-generating function, and provides a corresponding simple form for the r–$\mathcal{R}$ relationship that can be parameterized with only two parameters that have biologically relevant meanings that can assist in explaining intuition behind the r–$\mathcal{R}$ relationship. Third, it generalizes the result obtained from simple Susceptible-Infectious-Recovered (SIR) models, and approximately matches Susceptible-Exposed-Infectious-Recovered (SEIR), in the case where the latent period and infectious period are similar (see Appendix). While the gamma approximation has been applied to infer $\mathcal{R}$ in previous outbreaks (Table 1), its theoretical properties and practical importance has not yet been explored in depth.

For biological interpretability, we describe the gamma distribution using the mean $\bar{G}$ and the squared coefficient of variation κ (thus $\kappa = 1/a$, and $\bar{G} = a\theta$, where a and θ are the shape and scale parameters under the standard parameterization of the gamma distribution). Substituting the gamma distribution into (5) then yields the gamma-approximated speed-strength relationship:

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

We write:

$$\mathcal{R}^{\text{gamma}} \equiv (1 + \kappa\rho)^{1/\kappa}, \tag{7}$$

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 C, the higher the estimate of $\mathcal{R}$ (see Fig. 1). We then explore the behaviour of the gamma-approximated speed-strength relationship $\mathcal{R}^{\text{gamma}}$ defined above (equivalent to the Tsallis "q-exponential", with $q = 1-\kappa$ [43]): its shape determines how the estimate of $\mathcal{R}$ changes with the estimate of normalized generation length ρ.

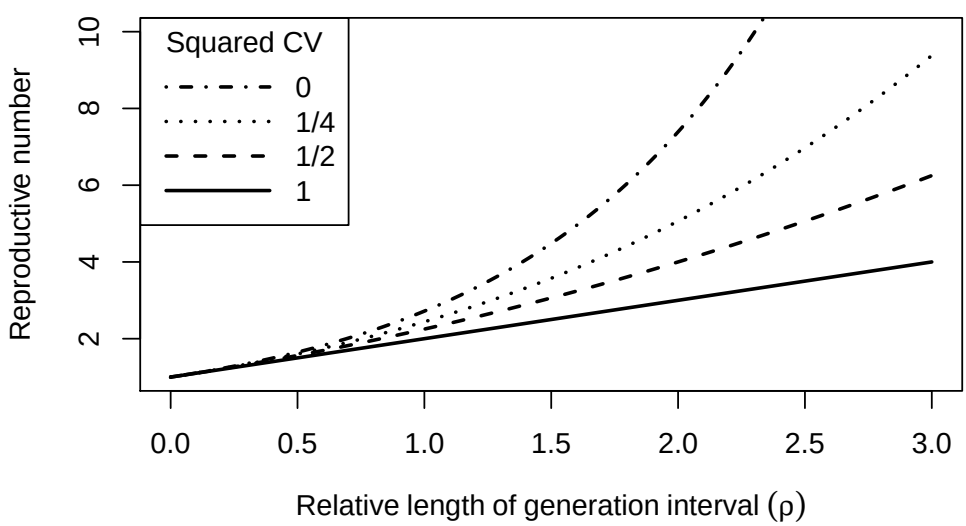

Figure 2: The approximate relationship (6) between mean generation time (relative to the characteristic time of exponential growth, $\rho = r\bar{G} = \bar{G}/C$) and reproductive number. The curves correspond to different amounts of variation in the generation-interval distribution.

For small ρ, $\mathcal{R}^{\text{gamma}}$ 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 $1/\kappa$, $\mathcal{R}^{\text{gamma}}$ converges toward $\exp(\rho)$.

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 $\kappa = 0$, the generation interval is fixed, so the average case must cause exactly $\mathcal{R} = \exp(\rho)$ 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 $\mathcal{R} < \exp(\rho)$ (as shown by [44]).

The straight-line relationship for $\kappa = 1$ also has a biological interpretation. This case corresponds to the classic SIR model, where the infectious period is exponentially distributed [6]. 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 the initial infectious individual.

Characterizing the r–$\mathcal{R}$ relationship with mean and coefficient of variation also helps explain results based on compartmental models [46, 37], because the mean and variance of the generation interval is linked to the mean and variance of latent and infectious periods. Shorter mean latent and infectious periods result in shorter generation intervals, and thus smaller $\mathcal{R}$ estimates. Less-variable latent periods result in less-variable generation intervals, and thus higher $\mathcal{R}$ estimates. Less-variable infectious periods have more complicated effects. While they reduce variation in generation interval (which would be expected to increase $\mathcal{R}$ estimates), they also reduce the mean generation interval [39] (which would lead to a decrease). This explains the apparent anomaly between earlier results: when mean generation interval is held constant, the effect of less-variable infectious periods is to reduce variation, leading to lower $\mathcal{R}$ for a given r [37]; when mean infectious period is held constant the effect of reducing generation interval will be stronger, leading to higher $\mathcal{R}$ [46].

There is a simple intuition for this latter result. In an exponentially growing epidemic, things that happen earlier have the largest effect. If we increase the variation in latent period while holding $\mathcal{R}$ constant, we have more early progression and more late progression to infectiousness. The former effect will be more important, and thus increasing variation should

increase r for a given value of $\mathcal{R}$ – or, equivalently, decrease $\mathcal{R}$ for a given value of r. Increasing variation in infectious period has the opposite effect: it increases the amount of both early and late recovery, and thus leads to smaller r for a given value of $\mathcal{R}$ (and conversely).

3.2 Approximation method, in practice

We test our approximation method by generating a pseudo-realistic generation-interval distributions using previously estimated/observed latent and infectious period distributions for different diseases (Table 2). For each pseudo-realistic distribution, we calculate the "true" relationship between r and $\mathcal{R}$ and compare it with a relationship inferred based on gamma distribution approximations. These approximations are first done with large amounts of data, allowing us to evaluate how well the approximations describe the r–$\mathcal{R}$ relationship under ideal conditions, and then tested with smaller amounts of data to assess how well methods perform when data is limited.

Estimating generation intervals is complex; our goal with pseudo-realistic distributions is not to precisely match real diseases, but to generate distributions that are likely to be roughly as challenging for our approximation methods as real distributions would be. We construct pseudo-realistic intervals from sampled latent and infectious periods by adding the sampled latent period to a an infection delay chosen uniformly from the sampled infectious period:

$$G_i = E_i + \mathrm{U}(0, I_i), \tag{8}$$

where G_i, E_i and I_i are the sampled intrinsic generation interval, latent period, and infectious period, respectively, and U represents a uniform random deviate. This implicitly assumes that infectiousness is constant across the infectious period [13]. We sample from latent and infectious periods obtained from observations (for empirical distributions), or by using a uniform set of quantiles (for parametric distributions). For the purpose of constructing pseudo-realistic distributions, we do *not* attempt to correct for the fact that observed intervals may be sampled in a context more relevant to backward than to intrinsic generation intervals (see [8]). We sample latent periods at random, and infectious periods by length-weighted resampling (since longer infectious period implies more opportunities to infect). For our examples, we used 10,000 quantiles for each parametric distribution and 10,000 sampled

generation intervals for each disease (see Appendix).

We then calculate "exact" relationships (for our pseudo-realistic distributions) by substituting sampled generation intervals into the exact speed-strength relationship (5). This relationship is then compared to the corresponding gamma-approximated relationship (6).

All calculations, numerical analyses and figures were made with the software platform R [34]. Code is freely available at https://github.com/dushoff/link_calculations.

4 Results

We investigate this approximation approach using three different examples. 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 work on the West African Ebola Outbreak [47], 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 Ebola virus disease (EVD), measles, and canine rabies.

4.1 Ebola

We generated a pseudo-realistic generation-interval distribution for Ebola virus disease (EVD) using information from [7] and a lognormal assumption for both the incubation and infectious periods. In contrast to gamma distributed incubation and infectious periods assumed by [7], we used a lognormal assumption for our components because it is straightforward and should provide a challenging test of our gamma approximation (see Appendix for results using gamma components). We used the reported standard deviation for the infectious period, and chose the standard deviation for the incubation period to match the reported coefficient of variation for the serial interval distribution, since this value is available and is expected to be similar to the generation interval distribution for EVD [7].

We then used our pseudo-realistic distribution to calculate both the exact (5) and gamma-approximated (6) speed-strength relationships (see Fig. 3).

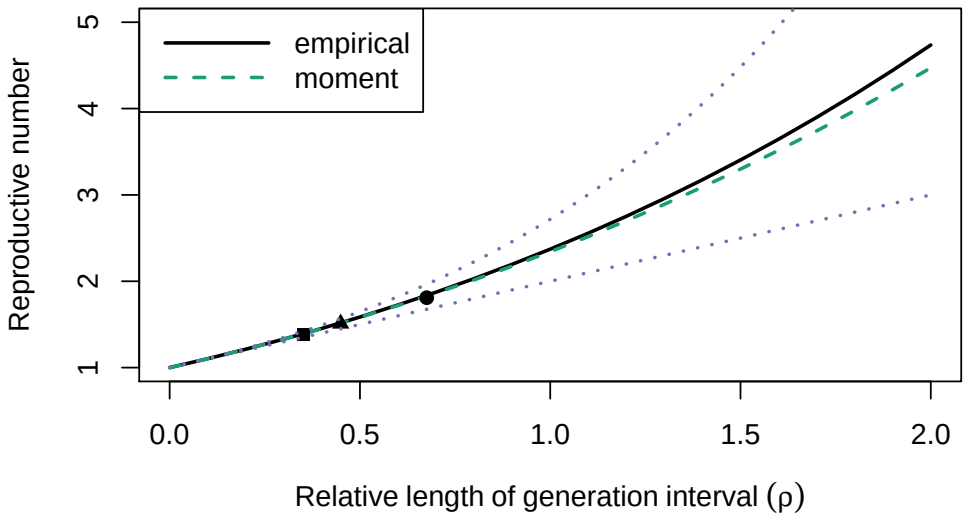

Figure 3: Estimating $\mathcal{R}$ for the West African Ebola Outbreak. (solid curve) The exact speed-strength relationship (5) using a pseudo-realistic generation-interval distribution based on [7]. (dashed curve) The Gamma approximated speed-strength relationship (6), using the mean and CV of a pseudo-realistic generation-interval distribution. (dotted curves) Naive 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 [7]: 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 [7] ($r = \ln(2)/T_2$).

The approximation is within 1% of the pseudo-realistic distribution it is approximating across the range of country estimates, and within 5% across the range shown. It is also within 2% of the World Health Organization (WHO) estimates derived from the observed time series assuming a poisson process with a known serial interval distribution.

4.2 Measles

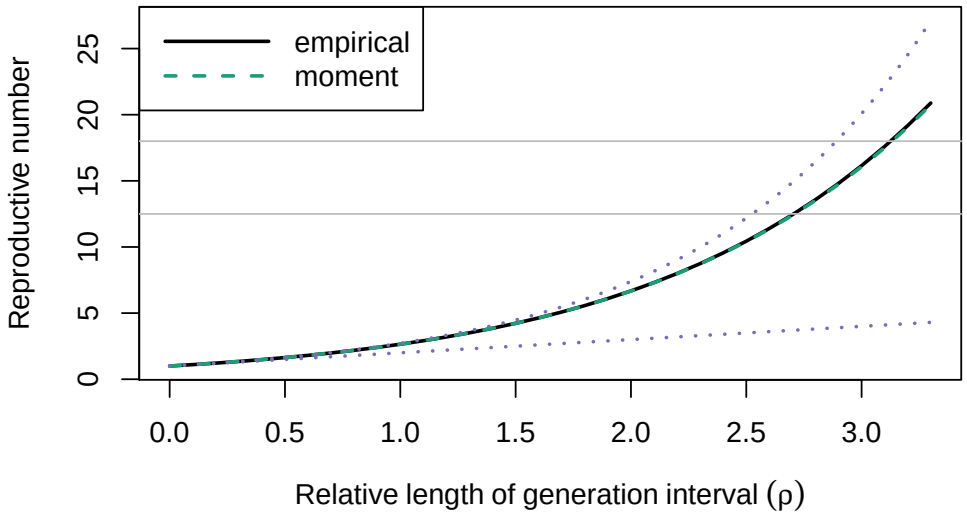

Figure 4: Estimating $\mathcal{R}$ for measles. (solid curve) The exact speed-strength relationship (5) using a pseudo-realistic generation-interval distribution. (dashed curve) The gamma-approximated relationship (6) using the mean and CV of a pseudo-realistic generation-interval distribution (this curve is almost invisible because it overlaps the solid curve) (dotted curves) Naive approximations based on exponential (lower) and fixed (upper) generation distributions. Gray horizontal lines represent $\mathcal{R}$ ranges estimated by [5]: 12.5-18.

We also applied the moment approximation to a pseudo-realistic generation-interval distribution based on information about measles from

[19], [22], and [5]. Incubation periods were assumed to follow a lognormal distribution [19]. Infectious periods were assumed to follow a gamma distribution with coefficient of variation of 0.2 [38, 22, 17]. Since variation in infectious period is relatively low [38, 17], and infectious period is short compared to incubation period, this choice is reasonable (and our results are not sensitive to the details).

Here, we found surprisingly close agreement between the exact and approximate relationships between r and $\mathcal{R}$ across a much wider range of interest (a difference of $< 1\%$ for $\mathcal{R}$ up to > 20) (see Fig. 4). On examination, this closer agreement is due to the smaller overall variation in generation times in measles: when overall variation is small, differences between distributions have less effect.

4.3 Rabies

We did a similar analysis for rabies by constructing a pseudo-realistic generation-interval distribution from observed incubation and infectious period distributions (see Fig. 5). Since estimates of $\mathcal{R}$ for rabies are near 1, there is small difference between the naive estimates and the gamma approximated speed-strength 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, a poorer job than if we use the approximation based on exponentially distributed generation times.

The reason for poor predictions of the moment approximation for higher $\mathcal{R}$ can be seen in the histogram shown in Fig. 5. 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 (in particular, the moment approximation misses the fact that the distribution has a density peak at a finite value). We expect short intervals to be particularly 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 pseudo-realistic generation intervals. This fit does a better job of matching the observed pattern of short generation intervals and of predicting the simulated relationship between r and $\mathcal{R}$ across a broad range (Fig. 5).

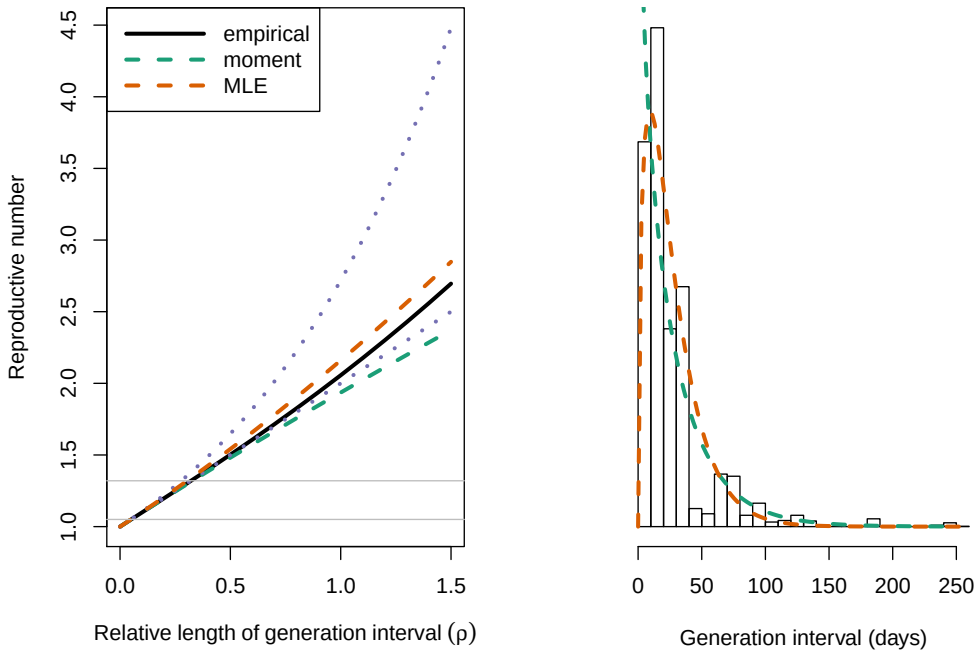

Figure 5: Estimating $\mathcal{R}$ for the Rabies. (Left) estimating $\mathcal{R}$ from rabies infectious case data. (solid curve) The exact speed-strength relationship (5) using a pseudo-realistic generation-interval distribution. (dashed curve) The Gamma approximated speed-strength relationship (6), using the mean and CV of a pseudo-realistic generation-interval distribution. "moment" approximation is based on the observed mean and CV of the distribution whereas "MLE" approximation uses the mean and CV calculated from a maximum-likelihood fit. (dotted curves) Naive approximations based on exponential (lower) and fixed (upper) generation distributions. Gray horizontal lines represent $\mathcal{R}$ ranges estimated by [13]: 1.05 - 1.32. (Right) histogram represents rabies generation interval distributions simulated from incubation and infectious periods observed by [13]. Dashed curves represent estimated distribution of generation intervals using method of moments and MLE (corresponding to approximate speed-strength relationships of the left figure).

Disease	Ebola	Measles	Rabies
Parameter		Values	
Reproduction number	1.38, 1.51, 1.81 [7]	12.5-18 [5]	1.05-1.32 [13]
Mean incubation period (days)	11.4 [7]	12.77 [19]	24.24 [13]
SD incubation period (days)	8.1 (see Sec. 4.1)	2.67 [19]	29.49 [13]
Mean infectious period (days)	5 [7]	6.5 [5]	3.57 [13]
SD infectious period (days)	4.7 [7]	2.9 (see Sec. 4.2)	2.26 [13]
Relative length of generation interval (ρ)	0.35, 0.45, 0.68 [7]	NA	NA
Mean generation interval (days)	16.2	15.0	26.6
CV generation interval	0.58	0.21	1.09

Table 2: Parameters that were used to obtain theoretical generation distributions for each disease. Reproduction numbers are represented as points in figure Fig. 3–5. Ebola parameters in triples represent Sierra Leone, Liberia, Guinea.

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 [44, 39, 28, 40]. 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 introduced and analyzed 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 used three disease examples to test the robustness of the framework. We also compared estimates based directly on estimated mean and variance of of the generation interval to estimates based on maximum-likelihood fits.

The gamma approximation for calculating $\mathcal{R}$ from r was introduced by [29], and provides estimates that are simpler, more robust, and more realistic than those from normal approximations (see Appendix). Here, we presented the gamma approximation in a form conducive to intuitive understanding of the relationship between speed, r, and strength, $\mathcal{R}$ (See Fig. 2). In doing so, we explained the general result that estimates of $\mathcal{R}$ increase with mean generation time, but decrease with *variation* in generation times [44, 46, 37]. We also provided mechanistic interpretations: when generation intervals are longer, 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 large, there is more early infection. As early infections contribute most to growth of an epidemic, faster exponential growth is expected for a given value of $\mathcal{R}$. Thus a higher value of $\mathcal{R}$ will be needed to match a given value of r.

We tested the gamma approximation framework by applying it to parameter regimes based on three diseases: Ebola, measles, and rabies. We found that approximations based on observed moments closely match true answers (based on known, pseudo-realistic distributions, see Sec. 4 for details) when the generation-interval distribution is not too broad (as is the case for Ebola and measles, but not for rabies), but that using maximum likelihood to estimate the moments provides better estimates for a broader range of parameters Sec. 4.3, and also when data are limited (see Appendix).

Many studies have linked r with $\mathcal{R}$ using generation interval distributions under various assumptions (Table 1). Simple methods, such as delta and exponential approximations, require minimal information (i.e., mean generation interval) but result in extreme estimates. More realistic methods, such as empirical and SEIR methods, may yield more realistic estimates, but are harder to interpret and require more data to implement.

Summarizing an entire generation interval distributions using two moments is a practically feasible approach that can give sensible and robust estimates of the relationship between r and $\mathcal{R}$ that lie between extreme estimates (see Appendix). More detailed methods will often be preferred when data are abundant but the gamma approximation can be easily used in preliminary analyses. In particular, 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 [2, 7, 30, 35, 18] but uncertainty in the generation-interval distribution was often neglected (but see [41]). During the outbreak, [47] 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 quantifying how various transmission routes and epidemic parameters

affect the moments of the generation interval distribution will help researchers better understand and predict the scope of future outbreaks.

Authors' Contributions

SWP led the literature review, wrote the first draft of the MS, assisted with analytic calculations, and finalized the simulations and data analysis; JD conceived the study and did the initial simulations and analytic calculations;all authors contributed to refining study design, literature review, and MS writing. All authors gave final approval for publication.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

The authors thank Steve Bellan for helpful comments, and Katie Hampson for making rabies data available.

Funding

This work was supported by the Canadian Institutes of Health Research [funding reference number 143486]. J.S.W. is supported by Army Research Office grant W911NF-14-1-0402.

References

[1] GK Aldis and MG Roberts. "An integral equation model for the control of a smallpox outbreak". In: *Mathematical biosciences* 195.1 (2005), pp. 1–22.

[2] C. L. Althaus. "Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa." In: *PLoS Curr* 6 (2014 Sep 2), 2014 Sep 2.

[3] Dorothy Anderson and Ray Watson. "On the spread of a disease with gamma distributed latent and infectious periods". In: *Biometrika* 67.1 (1980), pp. 191–198.

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

[5] Roy M Anderson and Robert M May. "Directly transmitted infections diseases: control by vaccination". In: *Science* 215.4536 (1982), pp. 1053–1060.

[6] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[7] 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.

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

[9] 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.

[10] Odo Diekmann and Johan Andre Peter Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Vol. 5. John Wiley & Sons, 2000.

[11] J. W. Eaton and T. B. Hallett. "Why the proportion of transmission during early-stage HIV infection does not predict the long-term impact of treatment on HIV incidence." In: *Proc Natl Acad Sci U S A* 111 (2014 Nov 11), pp. 16202–7.

[12] Neil M Ferguson et al. "Strategies for containing an emerging influenza pandemic in Southeast Asia". In: *Nature* 437.7056 (2005), p. 209.

[13] K. Hampson, J. Dushoff, S. Cleaveland, D. T. Haydon, M. Kaare, C. Packer, and A. Dobson. "Transmission dynamics and prospects for the elimination of canine rabies." In: *PLoS Biol* 7 (2009 Mar 10), e53.

[14] Katie Hampson, Jonathan Dushoff, Sarah Cleaveland, Daniel T Haydon, Magai Kaare, Craig Packer, and Andy Dobson. "Transmission dynamics and prospects for the elimination of canine rabies". In: *PLoS biology* 7.3 (2009), e1000053.

[15] JAP Heesterbeek and K Dietz. "The concept of $\mathcal{R}_0$ in epidemic theory". In: *Statistica Neerlandica* 50.1 (1996), pp. 89–110.

[16] J. H. Huber, G. L. Johnston, B. Greenhouse, D. L. Smith, and T. A. Perkins. "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.

[17] M. J. Keeling and B. T. Grenfell. "Disease extinction and community size: modeling the persistence of measles." In: *Science* 275 (1997 Jan 3), pp. 65–7.

[18] A. A. King, Domenech dCM, F. M. G. Magpantay, and P. Rohani. "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.

[19] J. Lessler, N. G. Reich, R. Brookmeyer, T. M. Perl, K. E. Nelson, and D. A. T. Cummings. "Incubation periods of acute respiratory viral infections: a systematic review." In: *Lancet Infect Dis* 9 (2009 May), pp. 291–300.

[20] 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 1), pp. 841–849.

[21] Marc Lipsitch et al. "Transmission dynamics and control of severe acute respiratory syndrome". In: *Science* 300.5627 (2003), pp. 1966–1970.

[22] 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.

[23] Alfred J Lotka. "Relation Between Birth Rates and Death Rates". In: *Science* 26.653 (July 1907), pp. 21–22.

[24] J. Ma, J. Dushoff, B. M. Bolker, and D. J. D. Earn. "Estimating initial epidemic growth rates." In: *Bull Math Biol* 76 (2014 Jan), pp. 245–60.

[25] ES McBryde et al. "Early transmission characteristics of influenza A (H1N1) v in Australia: Victorian state, 16 May–3 June 2009". In: *Eurosurveillance* 14.42 (2009), p. 19363.

[26] Hamish McCallum et al. "Transmission dynamics of Tasmanian devil facial tumor disease may lead to disease-induced extinction". In: *Ecology* 90.12 (2009), pp. 3379–3392.

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

[28] 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.

[29] H. Nishiura, C. Castillo-Chavez, M. Safan, and G. Chowell. "Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan." In: *Euro Surveill* 14 (2009 Jun 4), 2009 Jun 4.

[30] H. Nishiura and G. Chowell. "Theoretical perspectives on the infectiousness of Ebola virus disease." In: *Theor Biol Med Model* 12 (2015 Jan 6), p. 1.

[31] Hiroshi Nishiura and Gerardo Chowell. "Theoretical perspectives on the infectiousness of Ebola virus disease". In: *Theoretical Biology and Medical Modelling* 12.1 (2015), p. 1.

[32] K. A. Powers, M. E. Kretzschmar, W. C. Miller, and M. S. Cohen. "Impact of early-stage HIV transmission on treatment as prevention." In: *Proc Natl Acad Sci U S A* 111 (2014 Nov 11), pp. 15867–8.

[33] Oliver G Pybus, Michael A Charleston, Sunetra Gupta, Andrew Rambaut, Edward C Holmes, and Paul H Harvey. "The epidemic behavior of the hepatitis C virus". In: *Science* 292.5525 (2001), pp. 2323–2325.

[34] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[35] C. M. Rivers, E. T. Lofgren, M. Marathe, S. Eubank, and B. L. Lewis. "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.

[36] MG Roberts. "Modelling strategies for minimizing the impact of an imported exotic infection". In: *Proceedings of the Royal Society of London B: Biological Sciences* 271.1555 (2004), pp. 2411–2415.

[37] MG Roberts and JAP Heesterbeek. "Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection". In: *Journal of mathematical biology* 55.5-6 (2007), p. 803.

[38] RE Hope Simpson et al. "Infectiousness of communicable diseases in the household (measles, chickenpox, and mumps)." In: *Lancet* (1952), pp. 549–54.

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

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

[41] 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 7), pp. 145–54.

[42] J Trichereau et al. "Estimation of the reproductive number for A (H1N1) pdm09 influenza among the French armed forces, September 2009–March 2010". In: *Journal of Infection* 64.6 (2012), pp. 628–630.

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

[44] 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.

[45] Jacco Wallinga and Marc Lipsitch. "How generation intervals shape the relationship between growth rates and reproductive numbers". In: *Proceedings of the Royal Society of London B: Biological Sciences* 274.1609 (2007), pp. 599–604.

[46] Helen J Wearing, Pejman Rohani, and Matt J Keeling. "Appropriate models for the management of infectious diseases". In: *PLoS Med* 2.7 (2005), e174.

[47] 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 4), p. 8751.

[48] Ping Yan. "Separate roles of the latent and infectious periods in shaping the relation between the basic reproduction number and the intrinsic growth rate of infectious disease outbreaks". In: *Journal of theoretical biology* 251.2 (2008), pp. 238–252.

6 Appendix

6.1 Comparison with the SEIR model

The SEIR (susceptible-exposed-infectious-recovered) model is given by

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI/N \\ \frac{dE}{dt} &= \beta SI/N - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned} \tag{9}$$

It is assume that latent and infectious periods are exponentially distributed with mean $1/\sigma$ and $1/\gamma$, respectively.

For this model, generation interval distribution can be written as follows [39]:

$$g(\tau) = \frac{\sigma\gamma}{\sigma - \gamma}\left(\exp(-\gamma\tau) - \exp(-\sigma\tau)\right). \tag{10}$$

The mean generation interval is

$$\bar{G} = \frac{1}{\sigma} + \frac{1}{\gamma}. \tag{11}$$

and the square coefficient of variation is

$$\kappa = 1 - \frac{2}{\bar{G}^2\sigma\gamma} \tag{12}$$

[21, 37] derived the following r–$\mathcal{R}$ relationship under SEIR model:

$$\mathcal{R} = 1 + r\left(\frac{1}{\sigma} + \frac{1}{\gamma}\right) + \frac{r^2}{\sigma\gamma}. \tag{13}$$

Reparameterizing, we have

$$\begin{aligned}\mathcal{R} &= 1 + r\bar{G} + \frac{(1-\kappa)r^2\bar{G}^2}{2} \\ &= 1 + \rho + \frac{(1-\kappa)}{2}\rho^2\end{aligned} \tag{14}$$

Note that Taylor expansion of (7) yields the following:

$$\mathcal{R}^{\text{gamma}} \approx 1 + \rho + \frac{(1-\kappa)}{2}\rho^2 + O(\rho^3), \tag{15}$$

where $\mathcal{R}^{\text{gamma}}$ is an $\mathcal{R}$ estimate based on the gamma approximation. When $\sigma = \gamma$, $g(\tau)$ follows a gamma distributio and $O(\rho^3)$ vanishes. We expect the gamma approximation to work well when average duration of latent and infectious periods are similar.

6.2 Ebola example

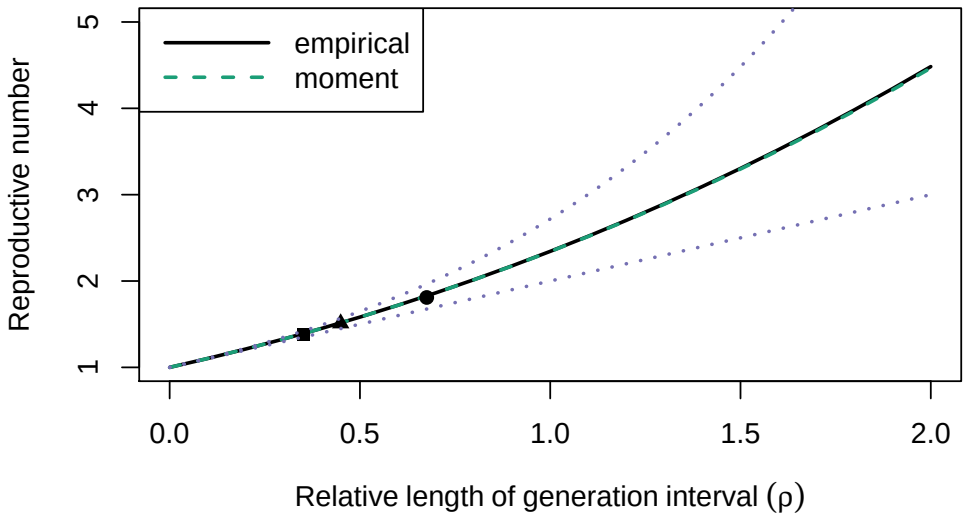

Figure S1: We perform the same analysis as we did in Sec. 4.1 assuming gamma distributed incubation and infectious periods. We find that the gamma approximated speed-strength relationship matches the true relationship almost perfectly in this case. Once again, we adjust the standard deviation of the incubation period to match the reported coefficient of variation in serial interval distributions. Rest of the parameters and points as in Fig. 3

6.3 Normal approximation

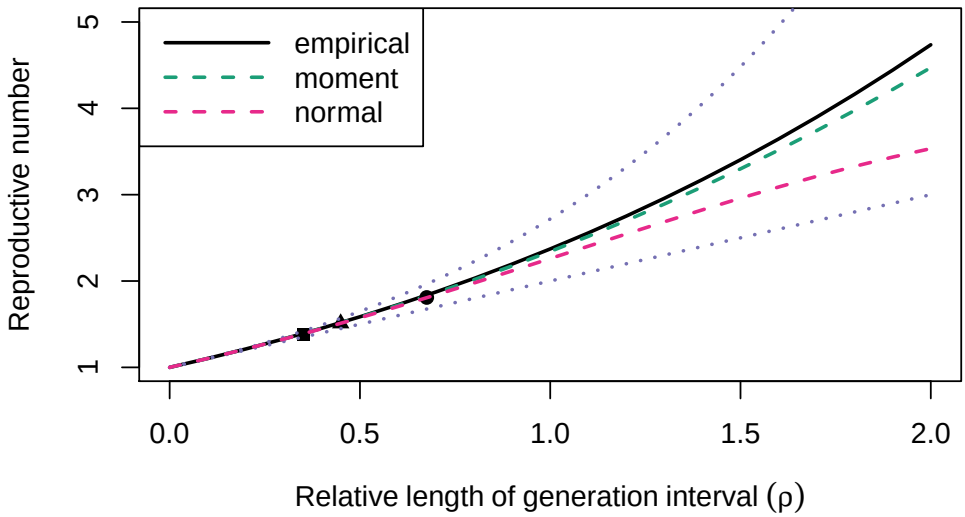

Figure S2: Approximating generation-interval distributions with a normal distribution has two problems. First, the distribution extends to negative values, which are biologically impossible. Second, as a consequence, the normal approximation predicts a saturating and eventually a decreasing $r-\mathcal{R}$ relationship for large r. Parameters and points as in Fig. 3.

6.4 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 given small sample sizes along with that of the more sophisticated maximum likelihood method. Fig. S3 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. As we increase sample size, our estimates become narrower. We also find that using the gamma approximated speed-strength relationship gives narrower estimates than the two naive estimates even when the sample size is extremely small ($n = 10$). It is important to note that Fig. S3 only conveys uncertainty in the estimate of coefficient of variation of generation interval distributions. Estimation of mean generation interval will introduce additional uncertainty into estimates of the reproductive number.

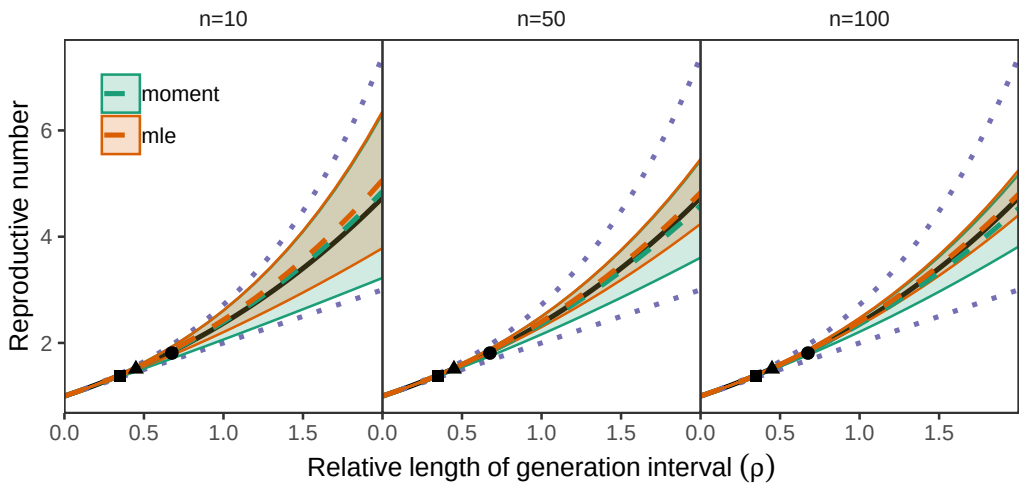

Figure S3: The effect of small sample size on approximated relationship between r and $\mathcal{R}$. (black solid curve) The relationship between growth rate and $\mathcal{R}$ using a known generation-interval distribution (see Fig. 3). (colored curves) Estimates based on finite samples from this distribution: dashed curves show the median and solid curves show 95% quantiles of 1000 sampling experiments. Note that the upper 95% quantile of the moment approximation and MLE approximation overlap. (dotted curves) Naive approximations based on exponential (lower) and fixed (upper) generation distributions.