The source of individual heterogeneity shapes infectious disease outbreaks

Baptiste Elie*, Christian Selinger, Samuel Alizon

MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France
* author for correspondance: baptiste.elie@ird.fr

Abstract

Infectious disease transmission patterns in some outbreaks can be more heterogeneous than in others, with striking effects on the way epidemics unfold. Some studies show that the biological sources of heterogeneity may matter, but they tend to do so without controlling for the overall heterogeneity in the number of secondary cases caused by an infection. Here, we control for this important bias to explore the role of individual variation in infection duration and transmission rate on parasite emergence and spread. We simulate outbreaks using a stochastic SIR model, with and without parasite evolution. Consistently with existing studies, we show that the variance in the number of secondary infections has the strongest effect on outbreak emergence probability but has little effect on the epidemic dynamic once emergence is certain. The origin of heterogeneity also affects the probability of emergence, but its more striking effects are about properties of epidemics that do emerge. In particular, assuming more realistic variances in infection duration distributions lead to faster outbreaks and a higher peak of incidence. When the parasite requires evolutionary changes to be able to spread, the impact of heterogeneity depends on the underlying evolutionary model. If the parasite evolves within the host, decreasing the infection duration variance decreases the probability of emergence. These results show that using realistic distributions for infection duration is necessary to accurately capture the effect of individual heterogeneity on epidemiological dynamics, which has implications for the monitoring and control of infectious diseases, as well as data collection.

Introduction

The expected number of secondary cases produced by an infected individual in a naive population is a key concept in epidemiology [1, 2], classically referred to as the basic reproduction number (denoted R_0). Only infections with $R_0 > 1$ can cause major outbreaks. However, beyond this mean value, super-spreading events, where an individual causes an unusually large number of secondary cases, have been shown to occur in many epidemics [3, 4, 5, 6, 7, 8]. The more frequent these events are, the higher the variance in the number of secondary cases, and the higher this variance, the lower the probability of outbreak emergence and the higher the epidemic growth (for outbreaks that do emerge) [4].

Many biological processes can explain the heterogeneity in the number of secondary cases (for a review, see [9]). However, models that investigate the effect of such processes tend to do so without controlling for the (overall) heterogeneity in the number of secondary cases, and often explore one source of heterogeneity at a time. One of

the few exceptions suggests that this is meaningful by showing that heterogeneity in host susceptibility has a lesser impact than heterogeneity in transmission rate in the probability of emergence [10].

Here, for a given distribution of the number of secondary cases, we use a stochastic mechanistic model to explore whether heterogeneity in transmission rates or recovery rates have different effects on an epidemic spread. Theoretically, assuming a more homogeneous distribution of infectious period duration increases the variability of population dynamics in the long term, therefore increasing the probability of outbreak extinction [11], but also increasing epidemic growth as well as epidemic peak size [12].

Even if initially maladapted (*i.e.* $R_0 < 1$), before fading out a parasite can evolve into a well-adapted strain that can cause a major outbreak, a phenomenon called evolutionary emergence [13, 14]. Since, when $R_0 < 1$, the final epidemic size increases with increasing heterogeneity in transmission rate [15], this can affect the risk of evolutionary emergence.

To study the effect of the source of individual heterogeneity, it is crucial to maintain a constant distribution of secondary cases ($\mathscr{Z}$) otherwise the specific effect of the studied source of heterogeneity cannot be distinguished from the more general effect of the variation in $\mathscr{Z}$, which is known to be strong [4].

Following earlier studies [4, 15, 16], we assume that the distribution of the number of secondary infections $\mathscr{Z}$ is Negative-Binomial, with mean R_0 and dispersion parameter k. The smaller k, the more dispersed $\mathscr{Z}$. While in a classical SIR model, $k = 1$, superspreading events tend to increase the heterogeneity of this distribution. For example, in the case of the 2003 SARS outbreak in Singapore, the dispersion was estimated to be $k = 0.16$ [4], and most recent data for Covid-19 estimate k to be around 0.3 [17].

We model individual transmission rate and infection duration using gamma distributions, denoted respectively $\mathscr{B}$ and Γ. For a given distribution $\mathscr{Z}$, we vary the coefficient of variation (CV) of the transmission rate (CV_B) and the CV of infection duration (CV_Γ) to compare heterogeneity originating from differences in infectious period duration and from transmission rate. Transmission rate can be seen biologically as the product of individual contact rate, and infectiousness.

We simulate outbreaks, without and with evolution, and measure key summary statistics to analyze the impact of different sources of heterogeneity on emerging outbreaks properties. We show that although the shape of the distribution of the number of secondary infections ($\mathscr{Z}$) is the main driver of the frequency of emergence, the source of this heterogeneity affects the risk of emergence, the properties of emerging epidemics, and even some aspects of evolutionary emergence.

Emergence risk

In our model the dispersion of the number of secondary cases k is entirely determined by the coefficients of variation in transmission rate (CV_B) and in infection duration (CV_Γ). This is shown in Figure 1A and further explained in the Methods. Data to infer distributions of number of secondary infections and of infection durations are easier to collect than that to inter distributions of transmission rates. This means that practically it is more straightforward to estimate the values of k and CV_Γ, which then impose a value for CV_B. We illustrate the feasibility of such inferences by indicating the parameter value for several well-studied outbreaks in Figure 1A. This also shows that the parameters ranges studied here are biologically realistic.

Figure 1B shows the proportion of outbreaks that emerge when $R_0 = 1.5$. Emergence is assumed to occur if the outbreak final size is greater than the classical herd immunity threshold, *i. e.* $1 - \frac{1}{R_0}$ [1]. The probability of emergence appears to be mainly governed by the individual repro-

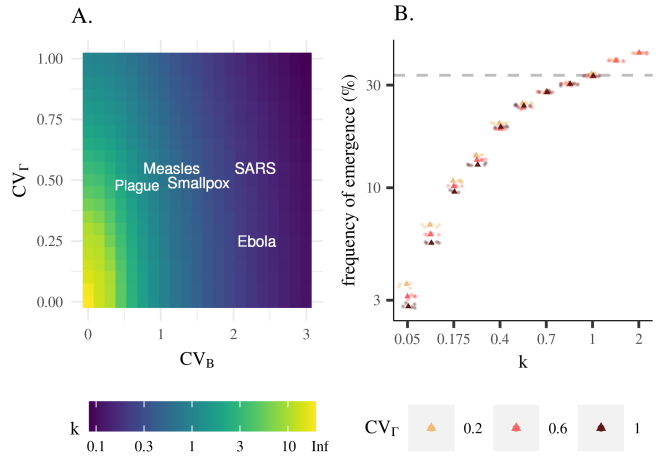

Figure 1 – **Relationship between k, CV_B, CV_Γ and the risk of outbreak emergence.** A) Numerical estimation of the secondary cases heterogeneity (k) as a function of CV_β and CV_γ. Names in white show the range of values estimated using maximum likelihood methods from outbreak data. If k remains constant, increasing CV_Γ always decreases CV_B. B) Frequency of emergence of an outbreak starting from one infection as a function of model heterogeneity. The value of CV_B is determined by that of CV_Γ and k. Each dot represents the frequency of emergence out of 20,000 runs. To illustrate the variability which can be expected from the computation, each frequency estimation is repeated 5 times, and the bold triangle represents the median value. Note that when $k > 1$ it is mathematically impossible to have $CV_\Gamma = 1$.

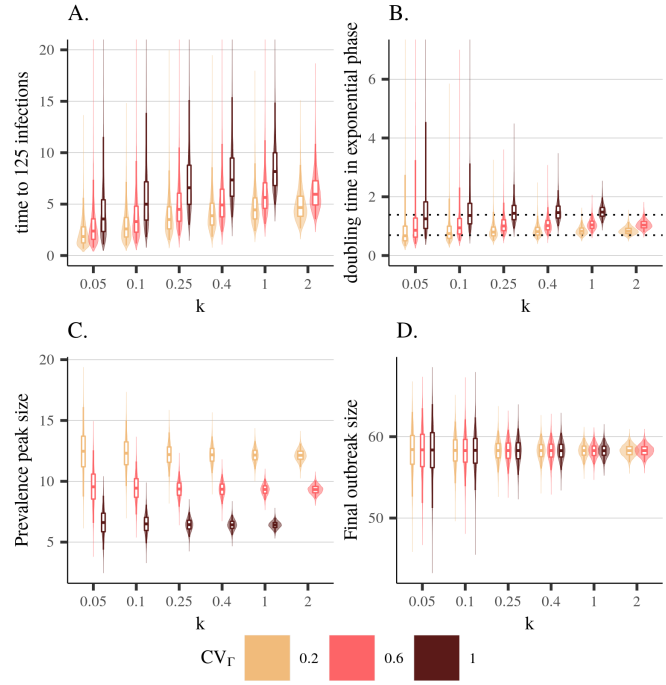

Figure 2 – **Properties of emerging outbreaks.** A) Time until prevalence reaches 70 infected hosts. B) Doubling time during the exponential phase (*i.e.* going from a prevalence of 250 to 500 infections). Top and bottom dotted lines indicate the expected doubling time for a memoryless model with $R_0 = 1.5$ or $R_0 = 2$, respectively. C) Prevalence peak size. D) Cumulative number of infections. Values are expressed as % of the total population. Each dot in the background represents one simulation. Boxplots and whiskers indicate the 97.5, 75, 50, 25 and 2.5% quantiles.

duction number dispersion k. However, for epidemics with many superspreading events (low k values), the frequency of emergence also increases with more homogeneous infection duration distributions (higher CV_Γ values). This effect is amplified for lower R_0 values (Fig. S3).

In the next section, we analyze the properties of the outbreaks that do emerge.

Dynamic of emerging outbreaks

Growth rates

We first study the initial phase of emerging outbreaks, when the number of events is low. Since the law of large numbers does not yet apply, variations in prevalence are strongly affected by stochasticity (Fig. S2). We quantify the early growth during this stochastic phase by measuring the time until the prevalence reaches 70 infected individuals, which is an appropriate outbreak threshold according to existing models [18]. As expected [4], we find that decreasing k leads to faster epidemic growth. For a given k, we further find that increasing the heterogeneity in infection duration (*i.e.* decreasing that in transmission rate) also increases epidemic growth (Fig. 2A).

We then study the deterministic exponential growth phase, which follows the stochastic phase and ends when the depletion of susceptible host population cannot be ignored anymore [18] (Fig. S2). Fig. 2B shows that, contrarily to the beginning of the outbreak, the growth during this exponential phase is mostly impacted by CV_Γ and the

effect of k is extremely limited. When assuming an exponential distribution for the infection duration (*i.e.* $CV_\Gamma = 1$), CV_B and, therefore, k have little effect: the doubling time remains in the order of 1.4 days, which is expected with $R_0 = 1.5$. Decreasing CV_Γ leads to a strong decrease in doubling time. Therefore, ignoring the distribution of the infection duration, *e.g.* by implicitly assuming an exponential distribution, can lead to over-estimating R_0 by up to 33%.

Epidemic peak size and final size

The prevalence peak value is highly affected by the heterogeneity in the infection duration: its median increases from 6% to 12% of the population when the CV_Γ decreases from 1 to 0.2 (Fig. 2C). k has little effect on the mean epidemic peak size, but we do find a correlation between the variance in peak size and that of $\mathscr{Z}$.

Finally, none of our heterogeneity metrics seem to affect the median final epidemic size, which is always close to 58% of the population (Fig. 2 D), corresponding to the expected value for $R_0 = 1.5$ according to classical theory [19]. As for the other metrics, the variance in the total epidemic size decreases with k.

Evolutionary emergence

Until now, we assumed that $R_0 > 1$. However, even if $R_0 < 1$, parasites can cause major outbreaks in a population if they acquire adaptive mutations. We now assume that the introduced 'resident' strain has a $R_0^r < 1$ and, therefore, will go extinct unless it evolves into a phenotypically different 'mutant' strain with $R_0^m > 1$. We also assume that the mutant strain arises either by taking over a host infected by the resident strain or during a transmission event.

We first consider the frequency of having at least one mutation in a simulation because this does not depend on the origin of the mutant (replacement

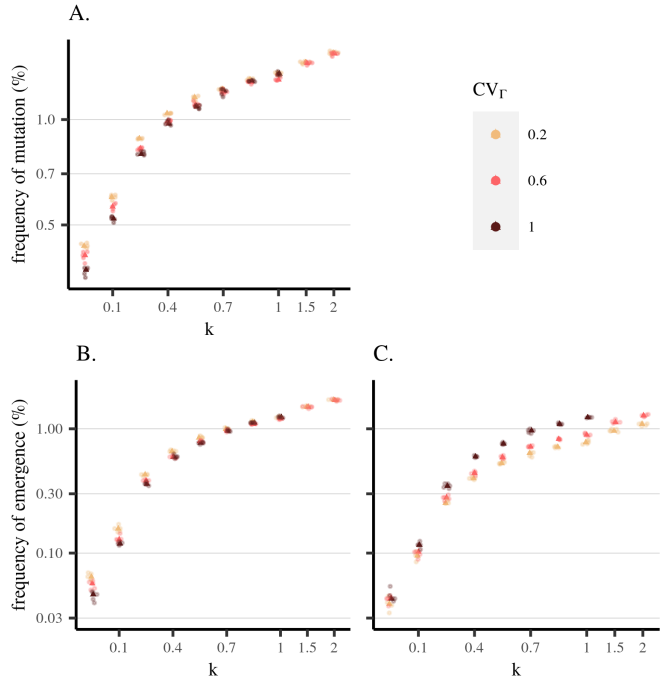

Figure 3 – **Individual heterogeneity and evolutionary emergence.** A) Frequency of mutation (the origin of the mutation can be ignored), B) Frequency of emergence with scenario when mutations occur during transmission, and C) when mutations occur within the host.

in the host or via transmission event). In figure 3A, we show that this frequency follows the same pattern as the frequency of (non-evolutionary) emergence (increasing CV_Γ decreases the probability of emergence, Fig. 1B). This makes sense since mutants cannot emerge if outbreaks go rapidly extinct.

However, the way the mutant emerges does affect the role of heterogeneity on the probability of outbreak (evolutionary) emergence. The frequency of emergence is similar in the two scenarios when $CV_\Gamma = 1$, but it diverges when the heterogeneity in infectious duration increases. When the mutation occurs during a transmission event (Fig. 3B), the probability of emergence is directly correlated to the frequency of mutation. Conversely, when the mutation occurs within the host, decreasing the heterogeneity of the infection leads to a decrease in the frequency of emergence, regardless of the individual secondary cases heterogeneity (Fig. 3C).

Discussion

Individual heterogeneity is known to shape infectious disease outbreaks [4], but few modelling studies take into consideration its origins and, among those that do, they do not control for the (overall) heterogeneity of $\mathscr{Z}$, the distribution of the number of secondary cases caused by an infected individual. In this study, we investigate the specific impact of heterogeneity from two biological sources: the infection duration, and the individual transmission rate.

Ordinary differential equations (ODE) that are classically used to model epidemics make the mathematically convenient assumption that the probability of occurrence of an event does not depend on the history of individuals *i.e.* that they are memoryless. While this may be true for transmission events, assuming a constant infection rate, it is biologically unrealistic for recovery because the daily probability to recover depends on the number of days since infection [20, 21]. When assuming a high, but biologically realistic [11, 7] heterogeneity in the number of secondary cases (low k), we find that increased heterogeneity in the infection duration is associated with decreased frequencies of outbreak emergence. Our interpretation is that in this scenario, the effect of stochastic processes is increased, and emergence is driven by rare superspreading events. Since decreasing the infection duration heterogeneity also decreases the stochastic phase duration (Fig.2A), this can increase the probability of emergence.

In branching process models, increasing the heterogeneity in transmission rates leads to a faster increase in cases per generation among the outbreaks that do emerge [4]. We show that this effect does not translate into an increased growth rate after the epidemic evades the stochastic phase and its dynamic becomes deterministic. We are able to detect this effect because, contrarily to earlier studies, we use a framework that allows us to simulate the entire course of an epidemic. Note that recent developments of branching process theory in epidemiology can incorporate the depletion of susceptible hosts [22]. Being able to compare, among the epidemics that emerge, key characteristics such as the final cumulative incidence, and the peak incidence size allows us to show, for instance, that even if the early growth rate is high, once the exponential growth phase is reached, superspreading events play a limited role in the epidemic growth.

Contrarily to superspreading events (parameter k), we show that the heterogeneity in infectious period duration plays an important role in the deterministic phase of the disease, by increasing the growth rate and, more strikingly, the prevalence peak size. While previous studies already reported a similar effect on both metrics [23, 24], our study further shows that this phenomenon is intrinsically related to the infectious period duration, and not to the secondary cases heterogeneity (which was not kept constant in these previous studies). These observations can be explained in the following way. When the infectious period is homogeneous, the mean generation interval (that is the time between the infection of the 'infector' and that of the 'infectee') is half that of the mean infectious period. The more heterogeneous the infectious period duration, the more some individuals have a short infectious period and never transmit. As a consequence, the transmission is biased in favor of individuals with longer infectious periods, which increases the mean of the generation interval. This also increases the doubling time and flattens the epidemic curve compared to an epidemic with a more homogeneous infectious period duration.

Finally, we show that infectious period duration heterogeneity can affect evolutionary emergence depending on the process that generates the mutant infection [14]. The impact of the mutational pathway and evolutionary scenario have been already pointed out by several studies [25, 10]. First, logically, we find no difference between the two

mutation scenarios if the process is memoryless - that is the duration of the infections is assumed to be exponentially-distributed ($CV_\Gamma = 1$). This further underlines the importance of questioning this biologically-unrealistic assumption [12, 11, 24]. However, when using more realistic infection duration distributions, we find that: i) When the mutation appears during transmission, we observe a similar trend as when the parasite emerges without evolution, which makes sense since mutant spread directly relates to the number of transmission events. ii) Conversely, when the mutation appears within a host during the infection, infection duration heterogeneity increases the frequency of emergence. We explain this by the fact that when the mutation appears within the host, if the infectious period is longer, it will have more chances to transmit new infections, which increases the probability that the mutation ignites the epidemic. This process overcomes the trends observed when considering only the emergence of the mutation (and not its spread).

Our effort to maintain a simple and tractable model of outbreak emergence naturally leads to several limitations. In particular, there is an identifiability issue regarding the biological origin of the transmission rate heterogeneity, which could come from variations in transmission rate or in host susceptibility. Addressing this issue would require to introduce multiple host types in the model, but also to make additional assumptions regarding the correlation between infectivity and susceptibility. It could also be interesting to enrich the model by considering a latent period during which exposed hosts are not yet infectious. This has been shown to affect R_0 estimates but in a deterministic model that did not take into account superspreading events [23]. More generally, our results suggest that investigating other sources that may contribute to the heterogeneity of the number of secondary infections can uncover potential biases.

We assumed that the population does not have any spatial structure, which is more realistic for directly-transmitted diseases, such as SARS or measles, than for sexually-transmitted infections for which contact networks impose strong constraints [15]. Furthermore, in the context of emergence, at the beginning of an epidemic, it has been shown that the spatial structure does not affect the outbreak metrics (especially R_0 and the control effort [26]). However, it is known that heterogeneity in host susceptibility decreases the final epidemic size, *i.e.* the total proportion of the population infected throughout the epidemic [27, 28]. We also do not include host demography and limit our analysis to a single epidemic wave.

Finally, this analysis relies on numerical results. This enables us to explore the role of stochasticity, which is particularly important to consider in the context of outbreak emergence from a mathematical modeling [29] and a statistical inference [30] point of view. However, it limits our analysis to the area of punctual parameters that we selected as being biologically relevant.

A limitation of the estimations of CV_Γ and CV_B from real epidemics is that we could only find joint distributions of secondary cases and serial interval in a measles outbreak. For the others, we used independent distributions for the other outbreaks. However, although this assumption does increase the confidence interval, we do not expect it to bias our results because any potential correlation is expected to be low. Indeed, we do not expect this correlation to be really high. For instance, where it is measurable, in the measles outbreak, the Spearman correlation coefficient between serial interval and secondary cases generated by the infector is 0.20.

These theoretical results have implications for outbreak monitoring. In particular, we show that making simplifying but biologically-unrealistic assumptions about the distributios of infection duration can lead to underestimate the risk of emergence, the epidemic doubling time, and the prevalence peak size. Given the risk of saturation of healthcare systems, accurately anticipating these

values is a major issue. This stresses the importance of collecting detailed data biological data to better inform epidemiological models.

Methods

Model without evolution

We implement a non-Markovian version of the classical Susceptible-Infected-Recovered (SIR) epidemiological model [19, 31], which means that not all rates are held constant throughout an infection. We assume that the host population is of fixed size N and that epidemics are initiated by a single infectious individual. At time t, each individual is characterized by its current state (susceptible, infectious, or removed), and, if infected, the time at which it will recover as well as its infectious rate.

There are two sources of heterogeneity in the model:

i) Transmission rate. We take into account individual heterogeneity either in contact rate or in infectivity by drawing the *per capita* transmission rate β_i for each individual i from a Gamma distribution, denoted $\mathscr{B}$. For mathematical convenience, and without further qualitative impact, we set the mean of $\mathscr{B}$ such that $\mathbb{E}[B]S_0 \approx \mathbb{E}[B]N = 1$, with S_0 the initial number of susceptible individuals. The standard deviation of $\mathscr{B}$ is imposed by the choice of the coefficient of variation (CV_B) which is further detailed below.

ii) Infection duration. We assume that individuals remain in the I compartment for a time drawn randomly from a Gamma distribution, noted Γ. By construction, the expectation of Γ is R_0 in our model and we vary its coefficient of variation between 0.2 and 1, which determines its shape, which equals to $1/CV_\Gamma^2$.

Coefficients of variation and $\mathscr{Z}$ dispersion

Heterogeneities in transmission rate and infection duration determine the distribution of the number of secondary infections ($\mathscr{Z}$). Little is known about the biological range of CV_B, whereas contact-tracing studies can provide some insights into the ranges of values of k and CV_Γ. Therefore, we use fixed values of k and CV_Γ and compute the required value of CV_B.

More specifically, we look for the value of CV_B that results in the smallest Kolmogorov-Smirnov distance between the Negative Binomial distribution with mean R_0 and dispersion parameter k and the computed distribution of the secondary cases determined by CV_B and CV_Γ, simulated by repeating 10^6 times the following procedure:

- We draw the transmission rate β_i following the $\mathscr{B}$ distribution.
- Let γ_i be the infection duration of this individual. Since transmission events are assumed to be memoryless and independent, the number of secondary cases follows an homogeneous pure-birth/Poisson process, the rate of which is proportional to the infection duration γ_i, *i.e.* $\zeta_i \sim \text{Poisson}(\beta_i N \gamma_i)$. Since γ_i is drawn from a Gamma distribution with shape κ and mean R_0, the overall process follows a Negative Binomial distribution with mean $\beta_i R_0$ and size κ. We can draw the number of secondary cases of this individual following this distribution.

Evolutionary emergence model

We introduce an additional class of individuals by distinguishing between I_r and I_m, which refer to individuals infected by the initial resident (resp. new mutant) parasite strain. The reproduction numbers of these strains are denoted R_0^r and R_0^m. We start with one infected individual of class I_r. Parasites can mutate during a transmission event, or within a host, at a given rate μ. The mutation increases their mean transmission rate, without altering CV_B (therefore also increasing the standard deviation of $\mathscr{B}$). We assume that the infectious period is not impacted by the mutation. Parasites within a host are assumed to be clonal: in the case of within-host mutation, we assume that the mutant instantaneously takes over the host. The mutant transmission rate is multiplied by $\frac{R_0^m}{R_0^r}$.

Frequency of emergence

We use the total epidemic size to determine if an outbreak has emerged or not. Emergence is assumed to occur when the total epidemic size is greater than the herd immunity threshold, *i.e.* $1 - \frac{1}{R_0}$ [1]. This implies any cross-reactive strain with the same R_0, would not be able to create an outbreak in the population. In the case of evolutionary emergence, we only consider the emergence of the mutant strain, with R_0^m

Numerical simulations

We simulate epidemics using Gillespie's next reaction method [32] to generate non-Markovian distributions. In this simple algorithm, two events can occur: infection, or recovery. We chose them as follows:

1. *Initialize (i.e. set S, $I = 1$, $t = 0$)*
2. *For each new infection, draw β_i and γ_i.*
3. *Compute the force of infection (i.e. $S.\sum_{i=1}^{I} \beta_i$), and draw the time to the next infection assuming an exponential distribution.*
4. *Look for the event with the closest time of occurrence, and update the compartments (S, I).*
5. *Update the time t to the time of the new event.*
6. *Go to step 2.*

In the evolutionary emergence scenario with mutation during transmission, the model is similar but there are two forces of infection, one for each class of infected host (I_r, I_m), and two additional events: infection by the mutant strain (assuming an exponential distribution with rate $\sum_{i=1}^{I_r} \beta_i \mu + \sum_{j=1}^{I_m} \beta_j$), and recovery of an individual in the I_m class.

In the scenario with mutation within the host, we distinguish the event of infection by the mutant strain (assuming an exponential distribution with rate $\sum_{j=1}^{I_m} \beta_j$) from the within host mutation of a resident strain into a mutant strain (assuming an exponential distribution with rate $I_r * \mu$).

The model was developed for this study and implemented in Java 11.0.7. We use parallel computation to decrease computing time. The code will be made available upon publication.

We monitor the following metrics:

- The time until prevalence reaches 30, 50, 100, 250, 500, and 1,000 infected hosts. This enables us to estimate the growth rate during the exponential phase.
- The prevalence peak size and outbreak final size. In the case of evolutionary emergence, we distinguish the resident final size from the mutant one.
- When the pathogen can mutate into a more adapted form, we register the time when the mutation occurs, the number of infected hosts by the resident at this time, and the cumulative number of infected hosts by resident at this time.

Simulation outputs were analyzed with R 3.6.3.

Parameters estimation for known outbreaks

In order to estimate CV_B and CV_Γ from observed outbreaks, we looked for epidemiological data in the literature where both serial interval and secondary cases are available from the same outbreak. Ideally, since those two distributions are not independent, we looked for joint distributions. However this was only possible for Measles [33]. Most of the time, the inline data gathering those two information together are not available, and we needed to assume independence of the two distributions for Ebola [34, 35], pneumonic plague [36], Smallpox [37, 38] and SARS outbreaks [4, 39] (see the Table S1 for further details about the data and parameters sources).

In order to get biologically relevant parameters, we infer parameters from a model with a latent period. Its distribution is retrieved from independent sources in the literature [40, 41, 35, 42]. We consider that this biological property does not change between different outbreak contexts. We also use independent estimates of R_0 [39, 33, 35]. Therefore, we restrict our estimation to three free parameters: the mean infection period, CV_B and CV_Γ.

Assuming the pool of susceptible hosts remained constant during the outbreak, we simulate 5.10^6 secondary cases and serial interval for a given mean Γ, CV_B and CV_Γ, in order to obtain their distribution. From that, we can compute the likelihood of the observed data, jointly when possible, or independently otherwise. The maximum likelihood parameters were then obtained using a grid search optimization.

Code and data availability

The different scripts are available on a github repository:
https://gitlab.in2p3.fr/ete/heterogeneity-outbreak.

The simulation results (necessary to run the code) are available on demand.

Acknowledgments

The authors thank the CNRS, the IRD, and acknowledge the itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this study (https://bioinfo.ird.fr/).

References

[1] RM Anderson, RM May, *Infectious Diseases of Humans: Dynamics and Control*. (Oxford Univ. Press, United Kingdom), Original edition, (1992).

[2] MJ Keeling, P Rohani, *Modeling Infectious Diseases in Humans and Animals*. (Princeton University Press), (2008).

[3] MEJ Woolhouse, et al., Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences* **94**, 338–342 (1997).

[4] JO Lloyd-Smith, SJ Schreiber, PE Kopp, WM Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).

[5] A Marm Kilpatrick, P Daszak, MJ Jones, PP Marra, LD Kramer, Host heterogeneity dominates West Nile virus transmission. *Proceedings of the Royal Society B: Biological Sciences* **273**, 2327–2333 (2006).

[6] MGM Gomes, et al., How host heterogeneity governs tuberculosis reinfection? *Proceedings of the Royal Society B: Biological Sciences* **279**, 2473–2478 (2012).

[7] A Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, S Abbott, AJ Kucharski, S Funk, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research* **5**, 67 (2020).

[8] JE Lemieux, et al., Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, eabe3261 (2021).

[9] KL VanderWaal, VO Ezenwa, Heterogeneity in pathogen transmission: Mechanisms and methodology. *Functional Ecology* **30**, 1606–1622 (2016).

[10] A Yates, R Antia, RR Regoes, How do pathogen evolution and host heterogeneity interact in disease emergence? *Proceedings of the Royal Society B: Biological Sciences* **273**, 3075–3083 (2006).

[11] AL Lloyd, Realistic Distributions of Infectious Periods in Epidemic Models: Changing Patterns of Persistence and Dynamics. *Theoretical Population Biology* **60**, 59–71 (2001).

[12] MP Malice, RJ Kryscio, On the Role of Variable Incubation Periods in Simple Epidemic Models. *Mathematical Medicine and Biology* **6**, 233–242 (1989).

[13] R Antia, RR Regoes, JC Koella, CT Bergstrom, The role of evolution in the emergence of infectious diseases. *Nature* **426**, 658–661 (2003).

[14] S Gandon, ME Hochberg, RD Holt, T Day, What limits the evolutionary emergence of pathogens? *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120086 (2013).

[15] T Garske, C Rhodes, The effect of superspreading on epidemic outbreak size distributions. *Journal of Theoretical Biology* **253**, 228–237 (2008).

[16] J Hellewell, et al., Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health* **8**, e488–e496 (2020).

[17] K Sun, et al., Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science* **371** (2021).

[18] M Hartfield, S Alizon, Epidemiological Feedbacks Affect Evolutionary Emergence of Pathogens. *The American Naturalist* **183**, E105–E117 (2014).

[19] WO Kermack, AG McKendrick, A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A* **115**, 700–721 (1927).

[20] J Lessler, et al., Incubation periods of acute respiratory viral infections: A systematic review. *The Lancet Infectious Diseases* **9**, 291–300 (2009).

[21] M Chan, MA Johansson, The Incubation Periods of Dengue Viruses. *PLoS ONE* **7**, e50972 (2012).

[22] A Barbour, G Reinert, Approximating the epidemic curve. *Electronic Journal of Probability* **18** (2013).

[23] HJ Wearing, P Rohani, MJ Keeling, Appropriate Models for the Management of Infectious Diseases. *PLoS Medicine* **2**, e174 (2005).

[24] T Britton, D Lindenstrand, Epidemic modelling: Aspects where stochasticity matters. *Mathematical Biosciences* **222**, 109–116 (2009).

[25] HK Alexander, T Day, Risk factors for the evolutionary emergence of pathogens. *Journal of The Royal Society Interface* **7**, 1455–1474 (2010).

[26] P Trapman, et al., Inferring R_0 in emerging epidemics—the effect of common population structure is small. *Journal of The Royal Society Interface* **13**, 20160288 (2016).

[27] N Becker, I Marschner, The effect of heterogeneity on the spread of disease in *Stochastic Processes in Epidemic Theory*, eds. JP Gabriel, C Lefèvre, P Picard. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 90–103 (1990).

[28] E Volz, SIR dynamics in random networks with heterogeneous connectivity. *Journal of Mathematical Biology* **56**, 293–310 (2008).

[29] T Britton, E Pardoux, Stochastic epidemics in a homogeneous community. *arXiv:1808.05350 [math]* **2255** (2019).

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

[31] RM Anderson, RM May, Population biology of infectious diseases: Part I. *Nature* **280**, 361–367 (1979).

[32] MA Gibson, J Bruck, Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *The Journal of Physical Chemistry A* **104**, 1876–1889 (2000).

[33] T Jombart, S Frost, P Nouvellet, F Campbell, B Sudre, Outbreaks: A Collection of Disease Outbreak Data (2020).

[34] O Faye, et al., Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: An observational study. *The Lancet Infectious Diseases* **15**, 320–326 (2015).

[35] WHO Ebola Response Team, Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections. *New England Journal of Medicine* **371**, 1481–1495 (2014).

[36] R Gani, S Leach, Epidemiologic Determinants for Modeling Pneumonic Plague Outbreaks. *Emerging Infectious Diseases* **10**, 608–614 (2004).

[37] F Fenner, et al., *Smallpox and Its Eradication*. (World Health Organization), (1988).

[38] H Nishiura, M Eichner, Infectiousness of smallpox relative to disease age: Estimates based on transmission network and incubation period. *Epidemiology and Infection* **135**, 1145–1150 (2007).

[39] M Lipsitch, et al., Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science* **300**, 6 (2003).

[40] AYC Kuk, S Ma, The estimation of SARS incubation distribution from serial interval data using a convolution likelihood. *Statistics in Medicine* **24**, 2525–2537 (2005).

[41] NTJ Bailey, On Estimating the Latent and Infectious Periods of Measles: I. Families with Two Susceptibles Only. *Biometrika* **43**, 15 (1956).

[42] H Nishiura, Determination of the appropriate quarantine period following smallpox exposure: An objective approach using the incubation period distribution. *International Journal of Hygiene and Environmental Health* **212**, 97–104 (2009).