The nature of individual heterogeneity shape outbreak dynamics

Baptiste Elie*, Christian Selinger, Samuel Alizon

MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France

* author for correspondance: baptiste.elie@ird.fr

Abstract

It is now common-place that pathogen transmission during an outbreak can be more heterogeneous than what is commonly assumed, and that it can have major consequences on their dynamics. However, previous studies did not explore the impact of the different biological sources of heterogeneity while controlling for the resulting heterogeneity in the number of secondary cases. In this study, we explore the role of individual variation in infection duration and transmission rate on parasite emergence and spread in a population. We simulate outbreaks using a custom stochastic SIR model, with and without evolution of the parasite. We show that for a given mean, the variance in the number of secondary cases is the main driver of the outbreak probability, with or without evolution, while it does not play a role on the outbreak dynamic once it emerged. On the opposite, a smaller and more realistic variance in the infection duration causes a faster outbreak. It is therefore useful to take into consideration more realistic distributions when modelling infectious diseases outbreaks.

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 $\mathscr{R}_0$. Only parasites with $\mathscr{R}_0 > 1$ are able to cause major outbreaks [1]. Super-spreading events have been shown to occur in many epidemics [3, 4, 5, 6, 7], *i.e.* where individuals cause an unusually large number of secondary cases. This tends to increase the variance in the number of secondary cases, which is also informative: a higher variance strongly decreases the probability of outbreak emergence, but increases the speed at which they grow [4].

There are many biological processes underlying the heterogeneity in the number of secondary cases (for a review, see [8]). However, less attention has been paid in modelling those different mechanisms while considering the heterogeneity in the number of secondary cases. There are some exceptions and models show that this is potentially meaningful: heterogeneity in susceptibility or in mixing have a lesser impact than heterogeneity in transmission rate in the probability of emergence [9].

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 in recovery rates have different effects on epidemic spread. Theoretical models show that assuming more homogeneous infectious period duration distributions leads to more variable dynamics in the long term, leading to fewer disease persistence [10], but also a faster epidemic growth, and higher epidemic peak size [11]. Even if initially maladapted (*i.e.* $\mathscr{R}_0 < 1$), a parasite can evolve before it fades out, a mechanism called evolutionary emergence [12, 13]. In parallel, when $\mathscr{R}_0 < 1$, the final epidemic size is larger if heterogeneity in transmission rate is increased [14]. Therefore, increasing heterogeneity in transmission rate can also increase the risk of evolutionary emergence [9, 15]. However, host heterogeneity implemented by vaccination is expected to decrease the potential of evolutionary emergence [13].

A limitation of existing studies is that they do not maintain a constant distribution of secondary cases. Therefore, they cannot disentangle the specific effect of the studied source of heterogeneity from the more general effect of the variation in $\mathscr{Z}$. In this work, we want to study, for a given $\mathscr{Z}$, whether the origin of this distribution heterogeneity matters, *i.e.* whether the coefficient of variation (CV) of the infectious contact rate (CV_B), or the CV of infection duration (CV_Γ) matter.

To control for the heterogeneity in the number of secondary infections, we assume that its distribution $\mathscr{Z}$ follows follows a Negative-Binomial distribution, with mean $\mathscr{R}_0$ and dispersion parameter k. This is a common assumption which allows to easily characterize the outbreak heterogeneity [4, 14, 16]. 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 SARS outbreak in Singapore in 2003, the dispersion was estimated to be $k = 0.16$ [4]. We model individual infectiousness and infection duration using gamma distributions, denoted respectively B and Γ. We therefore have three distributions of interest, and we want to study the effect of each, while keeping $\mathscr{Z}$ constant.

We simulate outbreaks, without and with evolution, and measure several key 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 frequency of emergence, the variance in the infectious period duration also affects both the risk of emergence and the entire dynamic of an outbreak. Furthermore, the nature of heterogeneity matters in terms of evolutionary emergence.

Emergence risk

In our model, as described in the Methods, the number of secondary cases dispersion k is entirely determined by the coefficients of variation in transmission rate (CV_B) and in infection duration (CV_Γ). Therefore, fixing two of the tree parameters determines the value of the third one as shown in Figure 1A. Distributions of number of secondary infections and infections duration are easier to estimate than distributions of transmission rates, which means k and CV_Γ are easier to estimate than CV_B. We illustrate this by indicating the parameter value for several well studied outbreaks on Figure 1A. This also shows that the parameters studied are within biologically realistic ranges. It is to note however that we can not always correctly estimate those two parameters because we could only find joint distributions of secondary cases and serial interval in a measles outbreak, while we used independent distributions for the other outbreaks. However, beyond the fact that it enlarges the confidence interval, we do not expect our results to be really biased from this assumption because the correlation is expected to be quite low. In fact, for measles, the correlation coefficient between serial interval and secondary cases generated by the infector is 0.2.

Figure 1B shows the proportion of outbreaks that emerge, assuming that $\mathscr{R}_0 = 1.5$. 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 reproduction 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 $\mathscr{R}_0$ values (Fig. S3).

In the next section, we analyse 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). The early growth during this stochastic phase is estimated here through the time until the prevalence reaches 70 infected individuals, which is an appropriate outbreak threshold according to existing models [17]. We find that decreasing k leads to faster epidemic growth (as in [4]). For a given k, we further find that increasing the heterogeneity in infection duration (*i.e.* decreasing that in infectiousness) also leads to faster growth (Fig. 2A).

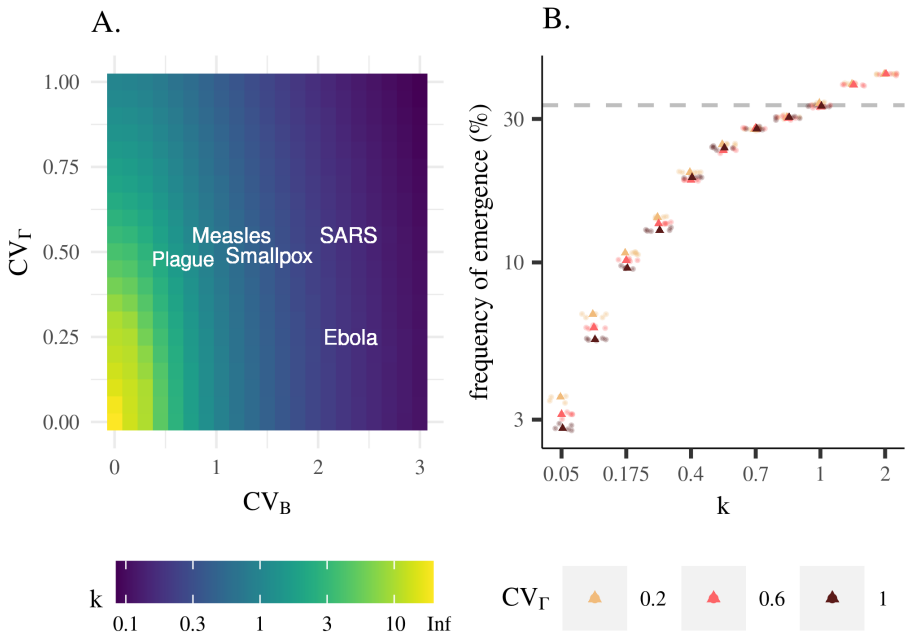

Figure 1 – **Relationship between k, CV_B, CV_Γ and the emergence of an outbreak.** A) Numerical estimation of the secondary cases heterogeneity (k) as a function of CV_β and CV_γ. Names in white show the range of values estimated using maximum likelihood methods from outbreak data. If k remains constant, increasing CV_Γ 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. Each condition is repeated 5 times, and the bold triangle represents the median value. Note that when $k > 1$ it is mathematically impossible to have $CV_\Gamma = 1$.

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 [17] (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 very minimal. When assuming an exponential distribution for the infectious period duration ($CV_\Gamma = 1$), CV_B or k have little effect and the doubling time remains around 2 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 infectious period duration, *e.g.* by implicitly assuming an exponential distribution, can lead to over-estimating $\mathscr{R}_0$ by up to 33%.

Epidemic peak and size

The prevalence peak value is highly affected by the heterogeneity in the infectious period 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

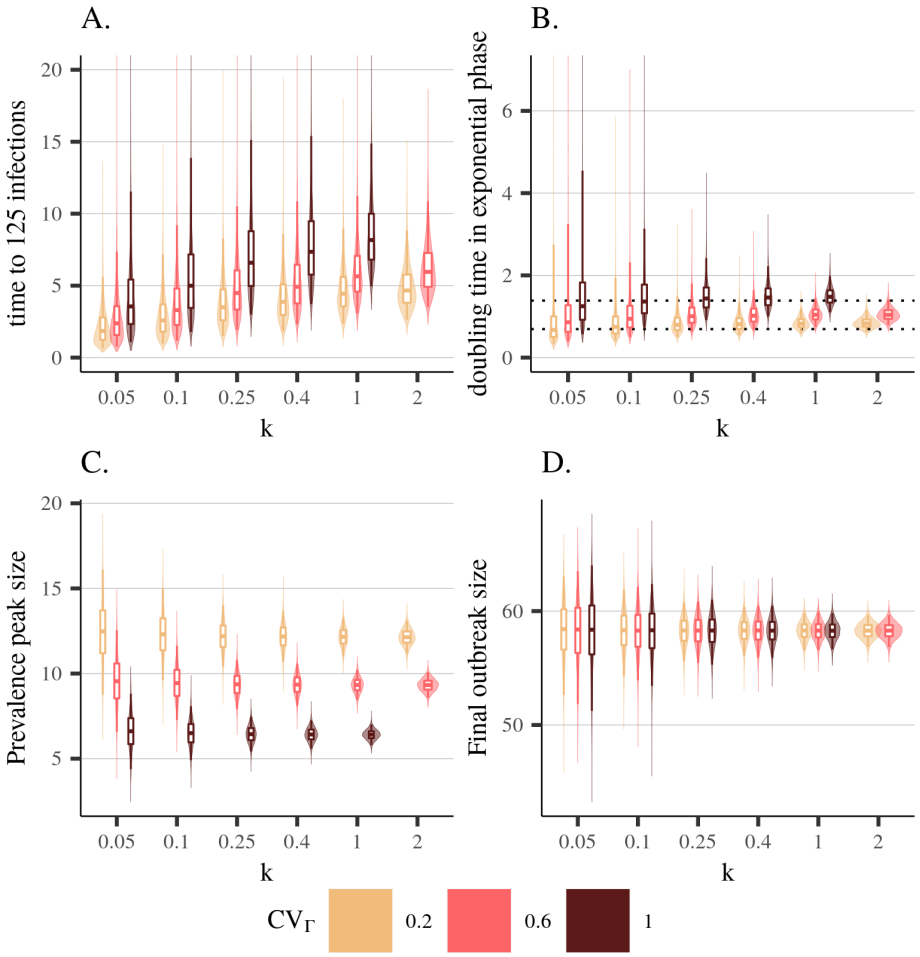

Figure 2 – **Properties of emerging outbreaks.** A) Time until prevalence reaches 70 infected hosts. B) Doubling time during the exponential phase (*i.e.* going from a prevalence of 250 to 500 infections). Top and bottom dotted lines indicate the expected doubling time for a memoryless model with $R_0 = 1.5$ 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.

peak size and that of $\mathscr{Z}$.

Finally, none of the heterogeneity metrics we used seem to affect the median estimate for the final epidemic size does not seem to be affected by individual heterogeneity, which is always close to 58% of the population (Fig. 2 D), corresponding to the expected value for a $\mathscr{R}_0 = 1.5$ [18]. As for the other metrics, its variance decreases with k.

Evolutionary emergence

We considered up to now the case where $\mathscr{R}_0 > 1$. However, even if $\mathscr{R}_0 < 1$, parasites can cause major outbreaks in a population if they acquire adaptive mutations. We assume that the introduced 'resident'

strain has a $R_0^r \approx 1$ and, therefore, will go extinct unless it evolves during the chains of transmission into a phenotypically different 'mutant' strain with $R_0^m > 1$. We also assume that the mutant arises either by replacing the resident in a host, or during a transmission event.

We first consider the frequency of having the mutation appearing at least once, because it does not depend on its origin (replacement in the host or via transmission event). In figure 3A, we show that it follows the same pattern as the frequency of (non-evolutionary) emergence shown earlier (increasing CV_Γ decreases the probability of emergence, Fig. 1B).

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

Discussion

Heterogeneity during an infectious disease outbreak is an important parameter to take into consideration, but only few studies take into consideration its different origins. In this study, we focus on two major biological sources of disease outbreak heterogeneity: the infectious period duration, and the individual transmission rate. Our model enables us to disentangle the specific impact of those two sources of heterogeneity during disease outbreaks from the more general role of secondary cases heterogeneity on outbreak dynamics. Secondary cases heterogeneity remains the main driver of disease emergence probability, which confirms previous observations [4, 14]. However, when considering the dynamic of emerging outbreaks, we observe significant differences whether the source of heterogeneity comes from the infectious period duration, or the transmission rate.

Classical ordinary differential equations (ODE) 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 [19, 20]. We first observe that with a high, but biologically realistic [10, 7] heterogeneity in the number of secondary cases, increased heterogeneity in the infectious period duration is associated with decreased frequencies of outbreak emergence. When

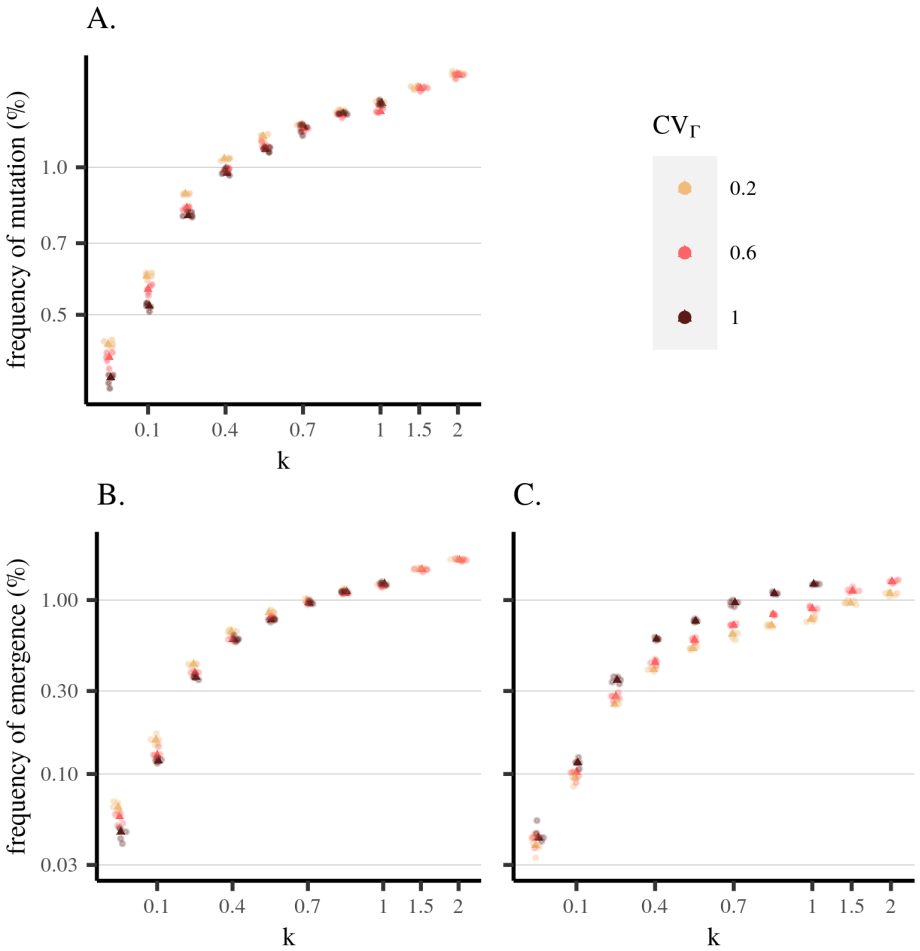

Figure 3 – **Individual heterogeneity and evolutionary emergence.** A) Frequency of mutation (the origin of the mutation can be ignored), B) Frequency of emergence with scenario when mutations occur during transmission, and C) when mutations occur within the host.

the heterogeneity in the number of secondary cases is high, the stochastic processes are amplified, and emergence is owing to rare superspreading events. Decreasing the infection duration heterogeneity also decreases the stochastic phase duration (Fig.2A), which would explain why we observe a higher frequency of emergence.

In branching process models, increasing the heterogeneity in transmission rate leads to a faster increase in cases per generation among the outbreaks that do emerge [4]. We show that this effect does not translate in an increase in the growth rate once the epidemic dynamic becomes deterministic. We detect this effect because, contrarily to earlier studies [9, 4], our method enables us to simulate the entire course of an epidemic. Note that recent developments of branching process theory in epidemics can incorporate the depletion of susceptible hosts [21]. Therefore, we can compare, among the epidemics that emerge, key characteristics such as the final cumulative incidence, and the peak incidence size. We find 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 study already reported a similar effect on both metrics [22, 23], 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). It is easy to explain these observations. When the infectious period is homogeneous, the mean generation interval (that is the time from infection of the 'infector' to the infection 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, transmission is then biased in favour 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 beyond secondary cases heterogeneity, infectious period duration heterogeneity can affect evolutionary emergence depending on the process that generates the mutant infection [13]. When the mutation appears during transmission, we observe a similar trend as when the parasite emerges without evolution: in case of important secondary cases heterogeneity, a more homogeneous infection duration leads to higher frequency of emergence. Unsurprisingly, the mutation frequency follows the same pattern: when the mutation appears during transmission, its probability of appearance is dependent on the outbreak size. On the opposite, when the mutation appears within a host during the infection, infectious period 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. Furthermore, this mechanism completely counterbalances the initial effect of mutation frequency of appearance. Interestingly, there is no difference between the two mutation models if the infections duration are assumed to be exponentially-distributed ($CV_\Gamma = 1$). This further underlines the importance of questioning this biologically-unrealistic assumption [11, 10, 23].

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 and from variations in susceptibility. Address this requires 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 $\mathscr{R}_0$ estimates but in a deterministic model that did not take into account superspreading events [22]. More generally, our results suggest that investigating other sources that may contribute to the heterogeneity of the number of secondary infections can help understand biases of classical models.

We assumed that the population does not have any spatial structure, which is more realistic for directly-transmitted diseases, such as SARS or measles, that tend not to be transmitted on a strict contact network contrarily to to sexually transmitted infections [14]. Moreover, 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 $\mathscr{R}_0$ and the control effort [24]). However, it is known that heterogeneity in host susceptibility is known to decrease the final epidemic size, *i.e.* the total proportion of the population infected throughout the epidemic [25, 26]. 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 mathematical modelling [27] and a statistical inference [28] point of view. However, it limits our analysis to the area of punctual parameters that we selected as being biologically relevant.

These theoretical results have implications for outbreak monitoring. First, we show that $\mathscr{R}_0$ can be over-estimated when estimated by combining and exponential growth rate and a serial interval. Therefore, accounting for realistic distributions of infectious period duration matters. The epidemic prevalence peak size also strongly depends on the heterogeneity in infection period duration. Given the risk of saturation of healthcare systems, being able to anticipate acute values is a major issue.

Methods

Model without evolution

We implement a non-Markovian version of the classical Susceptible-Infected-Recovered (SIR) epidemiological model [18, 29], which means that not all rates are held constant over the course of 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 characterised 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 infec-

tivity by drawing the *per capita* transmission rate β_i for each individual i from a Gamma distribution, denoted B. For mathematical convenience, and without further qualitative impact, we set the mean of 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 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 $\mathscr{R}_0$ in our model and we vary its coefficient of variation between 0.2 and 1, which determines its shape $\kappa = 1/CV_\Gamma^2$.

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

Heterogeneities in transmission rate and infection duration duration determine the distribution of the number of secondary infections. 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 $\mathscr{R}_0$ 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 B distribution.
- Let γ_i be the infectious period 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 infectious period 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 $\mathscr{R}_0$, 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

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 $\mathscr{R}_0$, 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 $\mathscr{R}_0^m$

Numerical simulations

We simulate epidemics using Gillespie's next reaction method [30] 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 [31]. 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 [32, 33], pneumonic plague [34], Smallpox [35, 36] and SARS outbreaks [4, 37] (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 [38, 39, 33, 40]. We consider that this biological property does not change between different outbreak contexts.

Assuming constant infectiousness during the infectious period and absence of infection during the latent period, we can compute the mean infection duration as:

$$\bar{\Gamma} = 2.(\bar{s} - \bar{e})$$

where $\bar{s}$ is the mean serial interval and $\bar{e}$ is the mean latent period. We also use independent estimates of $\mathscr{R}_0$ [37, 31, 33]. Therefore, we restrict our estimation to two free parameters: 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 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-o

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

[9] 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).

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

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

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

[13] 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).

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

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

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

[18] 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).

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

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

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

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

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

[24] 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).

[25] 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).

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

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

[28] 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).

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

[30] 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).

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

[32] 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).

[33] 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).

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

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

[36] 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).

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

[38] 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).

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

[40] 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).