medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

The stochastic dynamics of early epidemics: probability of establishment, initial growth rate, and infection cluster size at first detection

Peter Czuppon[1,2,3], Emmanuel Schertzer[4], François Blanquart[2,5,*], Florence Débarre[1,*]

[1] Institute of Ecology and Environmental Sciences of Paris (iEES-Paris, UMR 7618), Sorbonne Université, CNRS, UPEC, IRD, INRAE, 75252 Paris, France

[2] Center for Interdisciplinary Research in Biology, CNRS, Collège de France, PSL Research University, 75005 Paris, France

[3] Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany

[4] Faculty of Mathematics, University of Vienna, 1090 Wien, Austria

[5] Université de Paris, INSERM, IAME, 75018 Paris, France

[*] equal contributions

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Abstract. Emerging epidemics and local infection clusters are initially prone to stochastic effects that can substantially impact the epidemic trajectory. While numerous studies are devoted to the deterministic regime of an established epidemic, mathematical descriptions of the initial phase of epidemic growth are comparatively rarer. Here, we review existing mathematical results on the epidemic size over time, and derive new results to elucidate the early dynamics of an infection cluster started by a single infected individual. We show that the initial growth of epidemics that eventually take off is accelerated by stochasticity. These results are critical to improve early cluster detection and control. As an application, we compute the distribution of the first detection time of an infected individual in an infection cluster depending on the testing effort, and estimate that the SARS-CoV-2 variant of concern Alpha detected in September 2020 first appeared in the United Kingdom early August 2020. We also compute a minimal testing frequency to detect clusters before they exceed a given threshold size. These results improve our theoretical understanding of early epidemics and will be useful for the study and control of local infectious disease clusters.

Keywords: early epidemic dynamics, establishment probability, renewal equation, mass testing, timing of detection, testing frequency

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

1 Introduction

The emergence and spread of infectious diseases pose an increasing threat in an ever more interconnected world. A quantitative understanding of epidemic dynamics is necessary to improve control measures. Deterministic models are a suitable tool to describe the epidemiological dynamics once a large number of individuals has been infected. During the early phase of an epidemic or a local infection cluster however, stochastic effects cannot be neglected. These stochastic effects are due to the initially low number of infected individuals, and to the inherent stochasticity of the transmission process. Understanding and quantifying these stochastic effects will help, for example, assess the risk of new infection clusters emerging or estimate the size of a cluster associated with a new variant when such a variant is detected.

The infectiousness of an individual may vary over the course of their infection because of within-host viral dynamics if the transmission rate is correlated with the viral load. We consider a generic stochastic model in which infectiousness is an arbitrary function of time since infection. This stochastic model is called a Crump-Mode-Jagers process [1–3]. When the number of infected individuals gets large, this stochastic model can be approximated by a deterministic partial differential equation describing the distribution of the time since infection of the host population. This equation is known as the McKendrick-von Foerster partial differential equation [4–6].

Transmission timings are particularly influential during the early stages of the growth of an infection cluster, which is the focus of our work. It is therefore important to use biologically realistic distributions of transmission times [7], rather than assuming mathematically convenient but biologically unrealistic exponential distributions. A constant infectiousness over the duration of an individual's infection leads to the predominantly used framework of ordinary differential equations, while non-constant infectiousness can be captured by a partial differential equation. In addition to the added biological realism, a time-varying infectiousness of infected individuals can also properly capture the dynamical consequences of abrupt changes in transmission rate [6, 8]. This is not possible with an ordinary differential equation framework [9].

Here, we provide key results about the epidemic dynamics as described by the McKendrick-von Foerster equation. Stochasticity in transmission does not merely add noise to the dynamics, but also causes a systematic deviation from the deterministic description, which underestimates the initial growth of an establishing epidemic [10, 11]. This is in contradiction to a common misconception that stochasticity generally slows down the initial epidemic growth rate. We quantify the deviation between the deterministic and observed stochastic growth rate by conditioning the individual-based process on survival. After initial stochastic effects, the process converges to exponential growth with an asymptotic growth rate, denoted r, derived from the reproduction number R and the transmission rate. The distribution of time since infection in the stationary regime is exponential with parameter r, the asymptotic growth rate.

The reviewed and newly derived results can provide answers to public health related questions: How many importations will eventually result in a local infection cluster? How large is a local cluster once a first case is detected? When did a new variant – like Alpha, first detected in the United Kingdom (UK) – arise? How large is the detection rate of infectious individuals by a single mass testing effort? How many daily tests need to be conducted to detect local clusters before they exceed a certain size? We show how our theoretical results provide quantitative answers to these questions.

2 Expected epidemic size

We study the epidemic size of a cluster initiated by a single infected individual. We refer to a 'cluster' as the entire tree of infections that was initiated by a single infected individual. In particular, we do not

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

spatially restrict a cluster, nor do we constrain the time period in which transmissions need to occur.

Because some of our developments will also need them, we first recall results on *deterministic* epidemiological dynamics, then develop new analytical results on the expected early growth and the expected number of infected individuals once a stationary regime has been reached. We illustrate with simulations the variability across stochastic trajectories (Fig. 1). As observed before [10, 11], the expected growth rate during the early phase of cluster growth is greater than the long-term deterministic expectation, because clusters that do not die out are typically those that initially grow faster. We show how to account for this phenomenon in the mathematical description of the early phase and of the stationary regime.

In our stochastic simulations, we assume that the epidemic starts with a single infected individual at time $t = 0$. Each infected individual i is assigned a time since infection a_i. The time since infection determines the infectiousness of an individual through time. The term 'time since infection' is also referred to as 'age of infection' in the mathematical literature. We decouple the transmission rate $\tau(a)$ into a mean number of secondary infections R and a transmission probability density over time $\mu(a)$. We then have

$$\tau(a) = R \times \mu(a)\,. \tag{1}$$

This equation holds because $\int_0^\infty \mu(a)\,da = 1$, so that indeed the average number of secondary infections is given by R. This decoupling allows us, in a relatively simple way, to study different offspring distributions for R, while leaving the transmission density $\mu(a)$ unchanged.

For illustration, we assume that the distribution of transmission times follows a gamma distribution, but any distribution would be possible. In particular, a constant transmission rate (uniform distribution) would result in an exponential distribution of the transmission times (i.e., the memory-less distribution), which would reduce this general model into an ordinary differential equation (ODE).

2.1 Previous results on deterministic dynamics: renewal equation, growth rate and time-since-infection distribution

Throughout our analysis, we assume that the fraction of susceptible individuals is sufficiently large compared to the number of individuals infected in the early epidemic that it remains approximately constant. The overall rate at which new infections occur at time t, denoted by $i(t)$, in the deterministic regime is described by the following renewal equation [12]:

$$i(t) \;=\; \tau(t) \;+\; \int_0^t \tau(a)\,i(t-a)\,da\,, \tag{2}$$

where $\tau(a)$ is the transmission rate of an individual with time since infection a. The first term $\tau(t)$ reflects the new infections by the first infected individual at time t. The integral in Eq. (2) is the continuous version of the sum over the number of new infections caused by individuals with time since infection a (term $i(t-a)da$), which happens at rate $\tau(a)$. Intuitively, one can think about $i(t)dt$ being the incidence at time t, i.e., the number of newly infected individuals in the small time interval $[t, t+dt]$.

The cumulative number of infected individuals, i.e., the total epidemic size, which we denote by $I(t)$, is then given by

$$I(t) = 1 + \int_0^t i(s)\,ds = 1 + \int_0^t I(t-a)\tau(a)\,da\,, \tag{3}$$

with $I(0) = 1$ (mathematical details are given in the Supplementary Information (SI), Section S4). For simplicity, we do not consider recovery of infected individuals. However, individuals will of course stop transmitting when the time since infection is such that the transmission rate $\tau(a)$ becomes very small.

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

The epidemic size $I(t)$ will, for large times t, grow exponentially if $R > 1$. Formally, the asymptotic exponential growth rate r is obtained by solving the classical Euler-Lotka equation [12, 13]:

$$1 = \int_0^\infty e^{-rt}\tau(t)dt \quad \Leftrightarrow \quad \frac{1}{R} = \int_0^\infty e^{-rt}\mu(t)dt, \tag{4}$$

where r is also called the Malthusian parameter of the supercritical branching process [14]. In the case where $\mu(t)$ is given as the density of a gamma distribution with shape parameter α and scale parameter β, the exponential growth rate r is

$$r = \frac{R^{1/\alpha} - 1}{\beta}. \tag{5}$$

Convergence speed from the initial condition towards the asymptotic growth rate r is determined by the average number of secondary infections R and the transmission probability density μ. Intuitively, the faster a large number of infected individuals is reached (high R and/or small average transmission time), the faster is convergence towards the stationary growth regime.

Furthermore, it is possible to derive an explicit expression for the number of infected individuals over time, once asymptotic growth is reached. It follows from results of supercritical general branching processes and renewal theory [14], that the expected cumulative epidemic size is, for asymptotically large times t, given by

$$I(t) = I(0)\frac{e^{rt}}{rR\int_0^\infty e^{-rs}s\mu(s)ds}. \tag{6}$$

The integral in the denominator is the mean generation time of the Malthusian process [13, 15]. This is the time between the infection of the infecting individual and the time of infection of a randomly chosen secondary infection event. If the transmission density $\mu(s)$ were constant, the integral would be $1/(rR)$ and the epidemic size would be the solution of a constant infection process without depletion of susceptibles: $I(t) = I(0)e^{rt}$.

For an uncontrolled COVID-19-epidemic (we set $R = 2.9$, estimated for the French epidemic in Spring 2020 [16]), we obtain $r \approx 0.18$ per day, which corresponds to a doubling time of about 4 days. When interventions are in place (e.g., $R = 1.3$), then the Malthusian parameter is $r \approx 0.048$ per day, which corresponds to a doubling time of 14 days.

Under exponential growth, the distribution of the ages of infection in the population is given by an exponential distribution with parameter r, the exponential growth rate [14]. Intuitively, in an exponentially growing population, the number of individuals who were infected a days ago is e^r times greater than the number of individuals who were infected $a+1$ days ago. The exponential distribution also implies that for a large growth rate r, a large proportion of the cumulative number of infections will be very recent. For example, with $R = 2.9$, 30% among the total cumulative number of infections occurred within the last two days.

We now turn to the stochastic simulations and show how systematic deviations from the deterministic regime can be understood and mathematically described. We first give a stochastic correction for the asymptotic growth rate and then apply a similar idea to the general epidemic size process over time.

2.2 Asymptotic growth rate and epidemic size in the stochastic epidemic model

For large enough times after the initially infected individual started the local cluster, the epidemic grows exponentially at the rate predicted by the Euler-Lotka equation (Eq. (4)). However, the expected cumulative epidemic size derived for the deterministic case (Eq. (6)) includes epidemics that eventually

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

die out. Since we are only interested in epidemic clusters that eventually result in a large epidemic outbreak, we rescale the initial epidemic size by dividing by the survival probability p_{surv}:

$$I_{\text{surv}}(t) = \frac{I(t)}{p_{\text{surv}}} = \frac{I(0)}{p_{\text{surv}}} \frac{e^{rt}}{rR\int_0^\infty e^{-rs} s\mu(s)ds} . \tag{7}$$

This rescaling reflects conditioning of the epidemic process on survival (Fig. 1). The survival probability is $p_{\text{surv}} = 1 - p_{\text{ext}}$, where the probability of extinction p_{ext} is numerically computed as the fixed point of the probability generating function of the distribution of secondary infections. In words, the probability of extinction is equal to the probability that the initial infected individual does not produce any secondary infection, plus the probability that it produces one secondary infection which goes extinct (p_{ext}), plus the probability that it produces two secondary infections which both go extinct (p_{ext}^2), and so on; this intuition is outlined in SI, Section S1. Formally, the correction of the asymptotic limit in Eq. (7) is derived from a convergence result of a general branching process (SI, Section S3).

2.3 Initial stochastic growth of an epidemic

The initial growth rate of an epidemic that does not become extinct is initially steeper than its final asymptotic growth rate [10, 11] (compare the initial slope of the mean of stochastic simulations with the asymptotic growth for large times; gray dots vs. blue solid line in Fig. 1). This is due to the inherent stochasticity of the transmission process, which strongly affects the dynamics when there are only a small number of infected individuals. Clusters that escape extinctions are typically those that by chance benefited from a larger initial growth than the long-term expectation. This also means that deterministic models tend to underestimate epidemic sizes early on, or, if parameters are inferred from data, overestimate epidemic parameters such as the true basic reproduction number R_0, as for example observed in [17].

To account for this initial stochastic phase, one can alter the individual-based dynamics by conditioning the stochastic process on the survival of the epidemic. A similar procedure has been employed in [11]. This conditioning results in an adjustment of the transmission rate τ, which we denote by $\tilde{\tau}$. Formally, this adjustment is only justified for the stochastic process by Doob's h-transform [18] (details in SI, Section S5). In the large population size limit, we then approximate the adjusted transmission rate by the continuous analog of the adjusted transmission rate of the stochastic process. This approximation, while mathematically not fully justified, is a natural analogy of the conditioning of the asymptotic epidemic size in Eq. (7). The mean epidemic size of the adjusted process is then computed by

$$\tilde{I}(t) = 1 + \int_0^t \tilde{i}(s)ds, \tag{8}$$

where $\tilde{i}(s)ds$ is the incidence in the time interval $[s, s+ds)$ under the adjusted process. The rate of new infections $\tilde{i}(t)$ in the conditioned process now depends non-linearly on the history of the epidemic and therefore does not satisfy a renewal equation as in Eq. (2), but a delay differential equation:

$$\tilde{i}(t) = F(\tilde{i}(s); s \in [0, t]) . \tag{9}$$

The function F is explicitly computed in SI, Section S5 (Eq. (S37)). In short, the conditioning on survival of the epidemic results in an adjustment of the transmission rate τ by a factor that varies over time. This adjustment factor reflects the survival probability of the epidemic at a certain time and depends on the size and the age structure of the epidemic over time. The adjustment factor is largest at time $t = 0$, where it equals $(1 + p_{\text{ext}})$. Over time, the adjustment factor decreases and asymptotically approaches 1 for a large epidemic size, where the probability of extinction becomes negligible, i.e., for large times $\tilde{\tau} = \tau$.

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

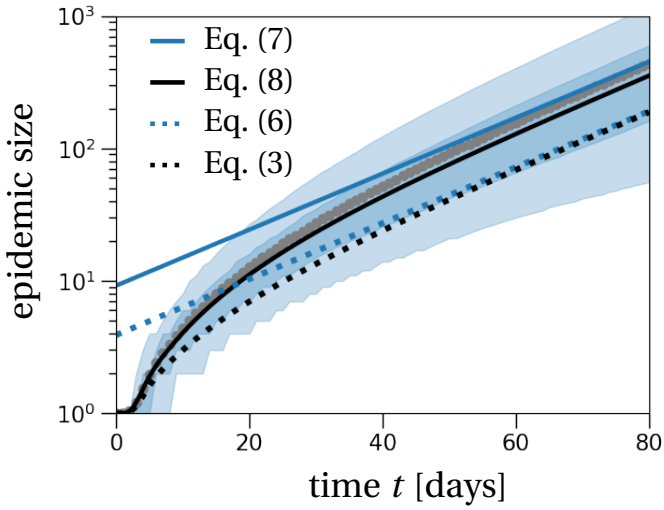


Figure 1: **The cumulative number of infected individuals over time.** The light and dark shaded regions show the 90% and 50% inter-quantile ranges obtained from 10,000 stochastic simulations that resulted in cluster establishment. Gray dots show the average of these simulations over time. The theoretical prediction (black solid line) is calculated from Eq. (8) with the adjusted transmission rate as computed in SI, Section S5. The black dotted line shows the prediction obtained from Eq. (3) without the conditioning for the epidemic to establish. The solid blue line is the epidemic size predicted by the asymptotic growth rate as stated in Eq. (7). The blue dotted line is the corresponding quantity without the stochastic adjustment (Eq. (6)). The effective reproduction number is set to $R = 1.3$, the number of secondary transmission events is Poisson-distributed, and the transmission density $\mu(t)$ is a Gamma distribution with the parameters given in Table 1.

In Fig. 1, we plot both the adjusted and non-adjusted versions of the mean epidemic size (Eqs. (3) and (8)). As mentioned above, the non-adjusted formula (black dotted line) underestimates the mean epidemic sizes as obtained from 10,000 stochastic simulations (gray dots). In contrast, conditioning the transmission density on survival (black solid line) predicts the mean epidemic size over time reasonably well, and also equilibrates approximately at the correct level. Overall, there is large variation in the epidemic sizes between different trajectories, as shown by the broad light shaded region corresponding to the 90% inter-quantile range of the simulated trajectories. To model the number of secondary infections, we have used the Poisson distribution in the figure because the adjustment of the transmission rate does not result in explicit expressions for the negative binomial case. Cumulative epidemic sizes in case the number of secondary infections is distributed according to a negative binomial or geometric distribution show more variation due to the larger variance in the number of secondary infections (Fig. S2 in SI, Section S6).

3 Applications

We now apply the theoretical results obtained above. First, we use the approximation of the epidemic size (Eq. (8)) to estimate the probability distribution of the emergence time of the Alpha variant, first detected in the UK in September 2020. The distribution of the emergence time also provides insight into the probability distribution of the size of the cluster when the variant was first sampled. As a

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Interpretation	Distribution	Parameters	Reference
mean number of secondary infections	Poisson	$R = 1.5$	[20]
time of secondary infection	Gamma (density: $\mu(t)$)	shape: 6.6, scale: 0.833 (mean: 5.5 days)	[24]
time from infection to sampling	Gamma (density: $f_{\text{sampling}}(t)$)	shape: 12, scale: 7/12 (mean: 7 days)	–
sequencing probability	Bernoulli	$p_{\text{sequencing}} = 0.042$	[22]
sampling probability	Bernoulli	$p_{\text{sampling}} = 0.25 \times p_{\text{sequencing}}$	[23]

Table 1: **Probability distributions and parameter values used in the case study of the Alpha variant.**

second application, we estimate the minimal testing frequency necessary to detect new emerging clusters before they exceed a certain size (on average). This prediction is especially relevant when the number of infected individuals is rare.

3.1 Distribution of the first detection time and cluster size at detection, and application to the origin of the Alpha variant

The Alpha variant initially consisted only of the B.1.1.7 lineage. This lineage was first detected in the UK from a sample that was collected on September 20th 2020 [19] and has rapidly become a major variant of concern due to its increased transmissibility [20] and pathogenicity [21]. Here, we develop a method to estimate the first infection of an individual with the Alpha variant and the distribution of the size of the Alpha-cluster on the day when the sample was taken in September, based on the dynamics of the epidemic size of a local cluster.

Our analysis requires the effective reproduction number, estimated to be $R = 1.5$ for the Alpha variant in November 2020 in the UK [20], and the probability for a sample taken in the UK to be sequenced, which was around 4.2% in October 2020 [22]. We will use this value in our analysis, keeping in mind that this might be an underestimate because the number of cases has been lower in September so the percentage of samples that could have been sequenced is potentially higher. Since only reported cases can be sampled, we additionally account for underreporting of cases. We assume that around 25% of all infections are detected [23]. Lastly, we need to define a distribution for the time that passes between infection and sampling of an infectious individual. We assume that the time from infection to sampling is a gamma distributed random variable (but any distribution would work) with a mean of seven days and a standard deviation of two days. The parameter values (Table 1) are chosen such that they give a probability of sampling an infected individual up until 3 days of their infection that is less than 1%, and a probability of sampling an infected individual after ten days of their infection that is less than 10%. All parameters are summarised in Table 1.

Distribution of the first detection time

To estimate the time of the first detection of an individual infected by the Alpha variant, we combine the sampling probability distribution f_{sampling} with the expected epidemic size at time t, given by the

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

adjusted version of the epidemic size in Eq. (8), and the number of infections until the first infected in the cluster is sampled and sequenced, which happens with probability p_{sampling} per infected individual. For readability, we refer to this first infected individual that is sampled and sequenced by case X and only write sampling when in fact we mean sampling and sequencing. The number of infection events till case X is infected, including case X, is denoted N_{inf}. It is a geometrically distributed number with probability p_{sampling}. Note that if we were interested in the jth sampling event, the number of infected individuals until the jth sampling event would be distributed according to a negative binomial distribution with 'success' probability p_{sampling} and dispersion $\kappa = j$.

We combine the distribution of N_{inf} with the *deterministic* time needed for the infected population to reach N_{inf} individuals (conditioned on non-extinction of this epidemic cluster as computed in Eq. (8)). We also refer to this time as *hitting time* and denote it by $t^{\text{det}}_{N_{\text{inf}}}$. To this, we add the time from infection of case X to their sampling. Denoting by T_{sampling} the random variable corresponding to the time of first detection and sampling, its probability density is given by:

$$\begin{aligned} h_{\text{sampling}}(t) &:= \lim_{\text{dt}\to 0} \mathbf{P}\left(T_{\text{sampling}} \in (t-\text{dt}, t+\text{dt})\right) \\ &\approx \sum_{i=1}^{\infty} \mathbf{P}\left(N_{\text{inf}} = i\right)\ f_{\text{sampling}}\left(t - t_i^{\text{det}}\right) \\ &= \sum_{i=1}^{\infty} p_{\text{sampling}}\left(1-p_{\text{sampling}}\right)^{i-1} f_{\text{sampling}}\left(t - t_i^{\text{det}}\right), \end{aligned} \tag{10}$$

where $f_{\text{sampling}}(s)$ denotes the probability density of the time from infection to sampling evaluated at time s (Table 1). We emphasize that the density of the first sampling time $h_{\text{sampling}}(t)$ is an approximation, because it is based on the mean epidemic size and not the whole distribution of the epidemic size. The mean epidemic size directly provides the deterministic hitting time t_i^{det}, neglecting the whole distribution of the epidemic size.

With our COVID-19-specific parameter set given in Table 1, we find that the mean time between the first infection of an individual with the Alpha variant and sampling of case X is around 46 days, indicating that the strain was present in the UK the 4th of August 2020 – yet, the variance is quite large for this distribution: the standard deviation is 19.5 days. The emergence date of the Alpha variant strongly depends on the sampling probability: smaller sampling probabilities result in earlier possible emergence dates than larger probabilities (Fig. 2). The distribution of secondary cases also impacts the timing: if the number of secondary infections is distributed as a negative binomial distribution, the date of emergence shifts closer to the date of sampling of case X. This effect is secondary though, compared to the impact of the sampling probability (Fig. 2).

In general, we find that the theoretical prediction of the probability distribution of the first sampling time captures the shape of the empirical distribution from the stochastic simulation results (Fig. 3a). Note that this implies that most of the variability in time does not come from stochasticity in epidemic size, but from the variability emerging from the random sampling of infected individuals (p_{sampling}) and the variability in the time from infection to sampling of infected individuals (f_{sampling}). Biologically, the variability in the time from infection to sampling arises from inter-individual variability in viral dynamics, symptom development, test seeking behaviour, etc. We find the largest discrepancy between theory and simulations at large first sampling times, i.e., we underestimate the right tail of the first sampling time distribution. This difference arises because our theoretical approximation does not take into account variability in the epidemic size process. Fig. 1 shows a large variation in the number of infected individuals over time between different stochastic trajectories. Most notably, there are several trajectories that remain at low cumulative epidemic sizes for a relatively long time. These trajectories are responsible for the long right tail of the sampling time distribution in Fig. 3a.

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

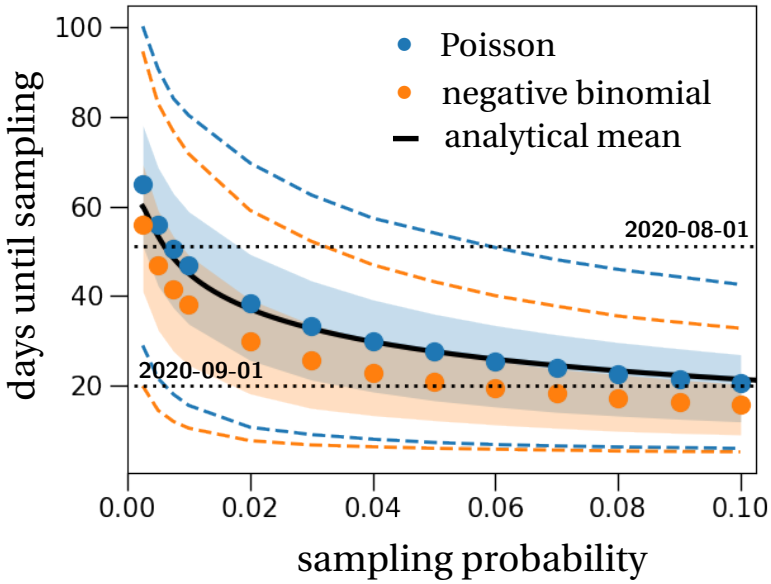


Figure 2: **The date of emergence of the first infected with the Alpha variant in the UK when varying the sampling probability.** The shaded regions and dashed lines show the 50% and 90% inter-quantile ranges obtained from 10,000 stochastic simulations that resulted in cluster establishment; blue for the secondary infections being Poisson distributed, orange for a negative binomial distribution. Dots represent the means of these simulations when varying the sampling probability. The effective reproduction number is set to $R = 1.5$, the dispersion parameter is $\kappa = 0.57$ [16], and the transmission density $\mu(t)$ and the waiting time between infection and sampling (f_{sampling}) are Gamma distributions with parameters as stated in Table 1. The theoretical mean (black solid line) of the first sampling time is calculated from Eq. (10), which only applies to the Poisson case.

Cluster size at the first detection time

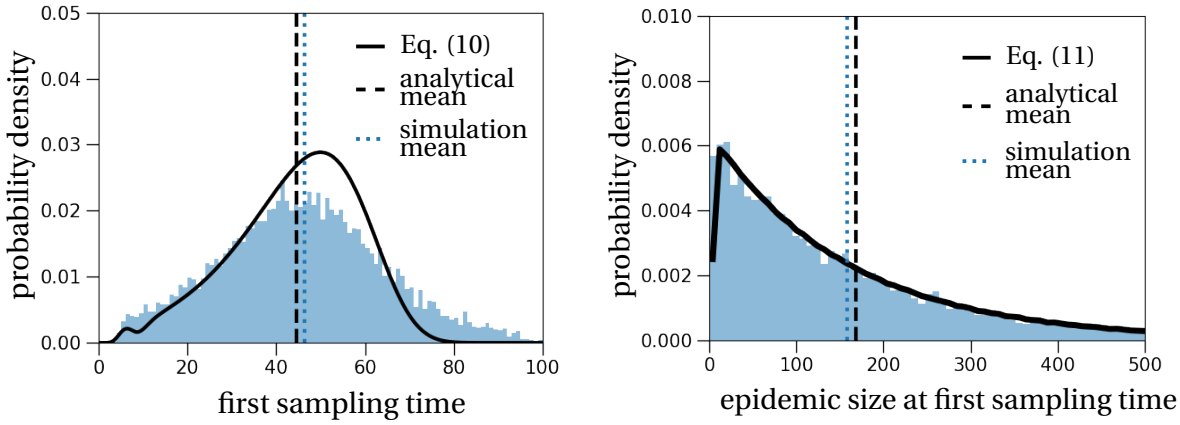
Next, we use this distribution of the first sampling time to infer the size of the epidemic cluster at that time. Therefore, we combine the adjusted epidemic size in Eq. (8) with Eq. (10) and obtain the following probability mass function for the size of the cluster at the sampling time of case X:

$$\mathbf{P}\left(I(T_{\text{sampling}}) = k\right) = \int_0^\infty h_{\text{sampling}}(t)\mathbb{1}_{\{\tilde{I}(t)\in[k-1/2,k+1/2)\}}\,dt\,, \tag{11}$$

where h_{sampling} is the probability that the first sampling time lies in the interval $[t, t+dt)$, given in Eq. (10).

This estimate of the epidemic size distribution approximates the simulated data reasonably well (Fig. 3b). The only notable difference occurs for very low epidemic sizes, where the epidemic size at the first sampling time ranges from 0-8 (bin size is set to 8 – the smallest bin size that produces a continuous theoretical prediction), as can be seen in the histogram in Fig. 3b. The mean size of the cluster with the Alpha variant at the first sampling time (obtained from stochastic simulations) consists of 159 individuals, yet again with a large standard deviation of 158 individuals. For example, the 95-percentile of the simulations predicts a cluster size of 476 infected individuals with the Alpha variant by the time of the first sampling of the variant.

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



(a) first sampling time

(b) cluster size at first sampling

Figure 3: **Distribution of the first sampling time and the cluster size at that time, parameterized to the case of the Alpha variant.** The histograms are obtained from 10,000 stochastic simulations and represent (a) the first sampling time of an infected individual with the Alpha variant, measured since the first infection of an individual with the Alpha variant (in days), and (b) the cluster size at this first sampling time. The theoretical predictions (black solid lines) are computed by Eqs. (10) and (11). The parameters and distributions used in the stochastic simulations are given in Table 1.

3.2 Minimal testing frequency to detect clusters of a given size

A single mass testing effort only results in a detection rate of between 25-48% of potentially infectious individuals, depending on the utilized test (rapid test or polymerase chain reaction) and the exponential growth rate r corresponding to reproduction numbers R between 1.3 and 3 (details in SI, Section S7). Therefore, we now ask whether repeated random testing in the population is a more feasible strategy to contain an infection cluster. Specifically, how often should we randomly test the population to detect a cluster before it exceeds a certain size? As a numerical example we will use a threshold cluster size of 30 infected individuals. We assume that testing is applied population-wide at random, independently of the infection state of an individual. The probability to test positive depends on the time since infection of an individual [25–27]. We denote the probability to test positive by a rapid test if the infected individual has been infected a days ago by $Q(a)$ (Fig. S3 in SI, Section S7).

If a fraction f of the population is tested every day, the detection probability of an infected individual is approximately given by

$$p_{\text{detect}} = 1 - \prod_{a=1}^{\infty} (1 - fQ(a)) \approx f \sum_{a=1}^{\infty} Q(a) \,. \tag{12}$$

The term $(1 - fQ(a))$ is the probability that an infected individual is not detected at their time since infection a. Hence, the product is the probability that an individual is never detected over the course of the individual's infection. The probability of detection is one minus this product. The approximation is valid when it is very unlikely that the same individual is tested more than once during the period when there is a high chance to detect their infection.

To determine the testing frequency above which the expected cluster size is smaller than 30 infected individuals, we repeat the steps from the previous sections: first, we determine the first detection time and then translate this result to the average cluster size at detection. Since our analytical result tends

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

to overestimate the cluster size at detection (Fig. S4 in SI, Section S8), this analytical procedure will provide an upper bound for the true testing frequency required to detect clusters of a certain size. In our numerical example with $R = 1.1$, this procedure results in a testing frequency of 0.013 for a threshold cluster size of 30 infected individuals.

Importantly, increasing the testing frequency when it is still low offers large benefits in terms of cluster size at detection because the epidemic size at detection reflects the exponential growth of the epidemic: it decreases exponentially with increasing testing frequency (Fig. S4 in SI, Section S8).

4 Discussion

We have collected key equations and derived novel results to account for stochasticity during the early phase of epidemic trajectories. Explicitly taking into account stochastic effects during the early phase of an epidemic allowed us to compute a good description of the mean epidemic size for all times (Eq. (8)). Importantly, our result captures the increased initial growth rate of surviving epidemics when compared to the asymptotic growth rate (Fig. 1). This is a known effect [10, 11], yet cannot be captured by deterministic epidemiological models. One important consequence of this theoretical underestimation of classically used models is that parameter inference during the early phase of an epidemic of, for example, the basic reproduction number, will result in an overestimation of the true value [17]. We provide a new mathematical description of the expected epidemic size over time that could be used in statistical inference during the early phase of emerging epidemics.

As a first application, we analytically derived the probability distribution of the first detection time of an epidemic cluster. While in principle applicable to any type of detection event, as for instance the first death or the first hospitalization event, we have focused on dating the emergence of the Alpha variant that was first sampled in the UK the 20th of September 2020. Our analysis is appropriate for clusters that descend from a single infected individual, and as long as population immunity is low enough for the supply of susceptible individuals to be unlimited. The Alpha variant was first detected in England in September 2020 and likely emerged there once, so our analysis can be applied to it. It would not apply, for example, to the Delta variant in the UK, unless the cluster linked to the first importation of the variant could be identified – and so the date of importation could be estimated. On average, we find that the Alpha cluster was started 46 days before its detection, which means that the variant was likely present in the UK on the 4th of August 2020. Usually, phylogenetic methods are used to date the evolutionary history of mutations [28]. In this particular case, a phylogenetic approach is difficult because of the large divergence between Alpha and non-Alpha variants sampled at a similar time [19]. Indeed, we did not find a published estimate of the date of emergence of the Alpha variant based on a phylogenetic analysis. In an attempt to date the origin of SARS-CoV-2, a combination of phylogenetic and epidemiological methods has been used to obtain a more complete picture of the very early dynamics of the COVID-19 epidemic [29]. Our new description for early epidemic growth provides a formal non-spatial description of the individual-based simulations that were used in [29] to date the very first COVID-19 case.

We additionally derived an analytical approximation for the probability distribution of the epidemic size at the first detection event. In contrast to a previous numerical estimate of the cluster size at the first disease-caused death that relies solely on the waiting time distribution until detection, e.g. the distribution from infection to death [30], we consider the whole epidemic trajectory of the cluster, i.e., from the first infected individual to the day of detection. The previously proposed method inevitably results in an overestimate of the actual epidemic size. Previous research has also shown that if the probability of detection since infection were constant over time, which is not the case in our setting, the cluster size at detection would be geometrically distributed [31, 32]. Whether the distribution of cluster sizes at detection is a geometric distribution if the detection process is not constant in time, is

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

an open question. In our specific data set, this seems to be the case (Fig. 3b).

We also applied our results to the evaluation of testing strategies. Currently (May 2021), aside from vaccination campaigns, frequent testing is seen as a possible solution to relax COVID-19-related restrictions in the short term. Our modelling approach gives an estimation of the minimal testing frequency per day to detect epidemic clusters of a certain size, for example small enough for manual contact tracing to be feasible. The minimal testing frequency depends on the test that is employed. In our numerical example, we have used the detection probability estimated for rapid tests, which were collected during the early phase of the epidemic in the UK in 2020 [27]. Since then, tests have improved so that our estimation of the minimal testing frequency is very likely an overestimate. We find that for a cluster size to be below 30 infected individuals (on average), each day around 0.13% of a total population would need to be randomly selected for testing, i.e., independently of the individual's infection status. Pooled sample testing strategies could be a solution to reduce the number of testing kits needed, and is a particularly reasonable option when the prevalence of infected individuals in a community is close to zero [33].

Additionally, we estimated the fraction of cases that can be detected during a single mass testing effort, as has been for example conducted in Slovakia in fall 2020 [34]. We find that with either a rapid test or a polymerase chain reaction test and with a reproduction number between $R = 1.3$ and $R = 2.9$, the detection rate of infectious individuals is between 26-48% (SI, Section S7). During the mass testing effort, a certain fraction of undetected individuals is still in the latent phase (0-3 days post-infection) and will become infectious after the mass testing event. Similar observations have also been made by employing a deterministic SEIR-model [35]. This indicates that only isolating positively detected individuals would be insufficient to contain the epidemic and that mass testing would need to be repeated to efficiently control the epidemic.

In conclusion, we have summarised existing theoretical results describing the early, stochastic dynamics of an epidemic, and developed new results on the mean epidemic size trajectory. We combined the establishment probability with the deterministic McKendrick-von Foerster equation to obtain a precise description of the expected epidemic size of an establishing epidemic over time. As an application, we approximated the probability distribution for the timing of a first infected individual in an epidemic cluster. This distribution can be used to estimate, for example, the emergence of new variants of a pathogen, like the Alpha variant. In addition, we derived the minimal testing frequency to detect clusters below a certain size. These applications are relevant from a public health perspective and could be used to guide the policy to contain and fight any infectious disease.

Data availability

The C++ codes, data files and Python scripts used to generate the figures are available at `https://gitlab.com/pczuppon/early-epidemic-dynamics`.

Funding

PC has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement PolyPath 844369. FD is funded by an Agence Nationale de la Recherche JCJC grant TheoGeneDrive ANR-19-CE45-0009-01. FB is funded by a Momentum grant from the CNRS.

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

References

[1] Crump K. S. and Mode C. J. A general age-dependent branching process. I. *Journal of Mathematical Analysis and Applications*, 24(3):494–508, 1968. doi: 10.1016/0022-247x(68)90005-x.

[2] Crump K. S. and Mode C. J. A general age-dependent branching process. II. *Journal of Mathematical Analysis and Applications*, 25(1):8–17, 1969. doi: 10.1016/0022-247x(69)90210-8.

[3] Jagers P. A general stochastic model for population development. *Scandinavian Actuarial Journal*, 1969(1-2):84–103, 1969. doi: 10.1080/03461238.1969.10405220.

[4] McKendrick A. G. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1925. doi: 10.1017/s0013091500034428.

[5] Diekmann O. and Heesterbeek J. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Wiley series in mathematical and computational biology. John Wiley and Sons, United States, 2000. ISBN 9780471986829.

[6] Foutel-Rodier F., Blanquart F., Courau P., Czuppon P., Duchamps J.-J., Gamblin J., Kerdoncuff É., Kulathinal R., Régnier L., Vuduc L., Lambert A., and Schertzer E. From individual-based epidemic models to McKendrick-von Foerster PDEs: A guide to modeling and inferring COVID-19 dynamics. *arXiv preprint*, 2020.

[7] Linton N., Kobayashi T., Yang Y., Hayashi K., Akhmetzhanov A., Jung S., Yuan B., Kinoshita R., and Nishiura H. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 9(2):538, 2020. doi: 10.3390/jcm9020538.

[8] Forien R., Pang G., and Pardoux É. Estimating the state of the COVID-19 epidemic in France using a model with memory. *Royal Society Open Science*, 8(3), 2021. doi: 10.1098/rsos.202327.

[9] Gatto M., Bertuzzo E., Mari L., Miccoli S., Carraro L., Casagrandi R., and Rinaldo A. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19):10484–10491, 2020. doi: 10.1073/pnas.2004978117.

[10] Mercer G. N., Glass K., and Becker N. G. Effective reproduction numbers are commonly overestimated early in a disease outbreak. *Statistics in Medicine*, 30(9):984–994, 2011. doi: 10.1002/sim.4174.

[11] Rebuli N. P., Bean N., and Ross J. Estimating the basic reproductive number during the early stages of an emerging epidemic. *Theoretical Population Biology*, 119:26–36, 2018. doi: 10.1016/j.tpb.2017.10.004.

[12] Wallinga J. and Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2006. doi: 10.1098/rspb.2006.3754.

[13] Britton T. and Tomba G. S. Estimation in emerging epidemics: biases and remedies. *Journal of The Royal Society Interface*, 16(150):20180670, 2019. doi: 10.1098/rsif.2018.0670.

[14] Haccou P., Jagers P., and Vatutin V. A. *Branching Processes*. Cambridge University Press, 2005. doi: 10.1017/cbo9780511629136.

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[15] Svensson Å. A note on generation times in epidemic models. *Mathematical Biosciences*, 208(1): 300–311, 2007. doi: 10.1016/j.mbs.2006.10.010.

[16] Salje H., Tran Kiem C., Lefrancq N., Courtejoie N., Bosetti P., Paireau J., Andronico A., Hozé N., Richet J., Dubost C.-L., and et al. Estimating the burden of SARS-CoV-2 in France. *Science*, 2020. doi: 10.1126/science.abc3517.

[17] Kochańczyk M., Grabowski F., and Lipniacki T. Super-spreading events initiated the exponential growth phase of COVID-19 with $\mathscr{R}0$ higher than initially estimated. *Royal Society Open Science*, 7 (9):200786, 2020. doi: 10.1098/rsos.200786.

[18] Doob J. Conditional brownian motion and the boundary limits of harmonic functions. *Bulletin de la Société Mathématique de France*, 79:431–458, 1957. doi: 10.24033/bsmf.1494.

[19] Rambaut A., Loman N., Pybus O., Barclay W., Barrett J., Carabelli A., Connor T., Peacock T., Robertson D. L., Volz E., and COVID-19 Genomics Consortium UK (CoG-UK). *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*, 2020. URL `https://bit.ly/3sBZ5Y4`.

[20] Volz E., Mishra S., Chand M., Barrett J. C., Johnson R., Geidelberg L., Hinsley W. R., Laydon D. J., Dabrera G., O'Toole Á., and et al. Transmission of SARS-CoV-2 lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv preprint*, 2021. doi: 10.1101/2020.12.30.20249034.

[21] Davies N. G., Abbott S., Barnard R. C., Jarvis C. I., Kucharski A. J., Munday J. D., Pearson C. A. B., Russell T. W., Tully D. C., Washburne A. D., and et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372(6538):eabg3055, 2021. doi: 10.1126/science.abg3055.

[22] COVID-19 Genomics Consortium UK. *How do we collect and sequence SARS-CoV-2 samples?*, 2020 – accessed February 9th, 2021. URL `https://www.cogconsortium.uk/news_item/how-do-we-collect-and-sequence-sars-cov-2-samples/`.

[23] Colman E., Enright J., Puspitarani G. A., and Kao R. R. Estimating the proportion of SARS-CoV-2 infections reported through diagnostic testing. *medRxiv preprint*, 2021. doi: 10.1101/2021.02.09.21251411.

[24] Hinch R., Probert W. J. M., Nurtay A., Kendall M., Wymant C., Hall M., Lythgoe K., Cruz A. B., Zhao L., Stewart A., and et al. OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLOS Computational Biology*, 17(7): e1009146, 2021. doi: 10.1371/journal.pcbi.1009146.

[25] Borremans B., Gamble A., Prager K., Helman S. K., McClain A. M., Cox C., Savage V., and Lloyd-Smith J. O. Quantifying antibody kinetics and RNA detection during early-phase SARS-CoV-2 infection by time since symptom onset. *eLife*, 9, 2020. doi: 10.7554/elife.60122.

[26] Kucirka L. M., Lauer S. A., Laeyendecker O., Boon D., and Lessler J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine*, 173(4):262–267, 2020. doi: 10.7326/m20-1495.

[27] Hellewell J., Russell T. W., The SAFER Investigators and Field Study Team, The Crick COVID-19 Consortium, CMMID COVID-19 working group, Beale R., Kelly G., Houlihan C., Nastouli

medRxiv preprint doi: https://doi.org/10.1101/2020.11.17.20233403; this version posted September 17, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

E., and Kucharski A. J. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC Medicine*, 19(1), 2021. doi: 10.1186/s12916-021-01982-x.

[28] Hadfield J., Megill C., Bell S. M., Huddleston J., Potter B., Callender C., Sagulenko P., Bedford T., and Neher R. A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23): 4121–4123, 2018. doi: 10.1093/bioinformatics/bty407.

[29] Pekar J., Worobey M., Moshiri N., Scheffler K., and Wertheim J. O. Timing the SARS-CoV-2 index case in Hubei province. *Science*, 372(6540):412–417, 2021. doi: 10.1126/science.abf8003.

[30] Jombart T., van Zandvoort K., Russell T. W., Jarvis C. I., Gimma A., Abbott S., Clifford S., Funk S., Gibbs H., Liu Y., and et al. Inferring the number of COVID-19 cases from recently reported deaths. *Wellcome Open Research*, 5:78, 2020. doi: 10.12688/wellcomeopenres.15786.1.

[31] Trapman P. and Bootsma M. C. J. A useful relationship between epidemiology and queueing theory: The distribution of the number of infectives at the moment of the first detection. *Mathematical Biosciences*, 219(1):15–22, 2009. doi: 10.1016/j.mbs.2009.02.001.

[32] Lambert A. and Trapman P. Splitting trees stopped when the first clock rings and vervaat's transformation. *Journal of Applied Probability*, 50(1):208–227, 2013. doi: 10.1239/jap/1363784434.

[33] Brault V., Mallein B., and Rupprecht J.-F. Group testing as a strategy for COVID-19 epidemiological monitoring and community surveillance. *PLOS Computational Biology*, 17(3):e1008726, 2021. doi: 10.1371/journal.pcbi.1008726.

[34] Pavelka M., Van-Zandvoort K., Abbott S., Sherratt K., Majdan M., CMMID COVID-19 working group, Inštitút Zdravotných Analýz, Jarčuška P., Krajčí M., Flasche S., and Funk S. The impact of population-wide rapid antigen testing on SARS-CoV-2 prevalence in Slovakia. *Science*, 372(6542): 635–641, 2021. doi: 10.1126/science.abf9648.

[35] Bosetti P., Kiem C. T., Yazdanpanah Y., Fontanet A., Lina B., Colizza V., and Cauchemez S. Impact of mass testing during an epidemic rebound of SARS-CoV-2: a modelling study using the example of France. *Eurosurveillance*, 26(1), 2021. doi: 10.2807/1560-7917.es.2020.26.1.2001978.

Electronic Supplementary Material: The stochastic dynamics of early epidemics: probability of establishment, initial growth rate, and infection cluster size at first detection

Peter Czuppon[1,2,3]
Emmanuel Schertzer[4]
François Blanquart[2,5,*]
Florence Débarre[1,*]

[1] Institute of Ecology and Environmental Sciences of Paris (iEES-Paris, UMR 7618), Sorbonne Université, CNRS, UPEC, IRD, INRAE, 75252 Paris, France

[2] Center for Interdisciplinary Research in Biology, CNRS, Collège de France, PSL Research University, 75005 Paris, France

[3] Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany

[4] Faculty of Mathematics, University of Vienna, 1090 Wien, Austria

[5] Université de Paris, INSERM, IAME, 75018 Paris, France

$*$ equal contribution.

S1 Probability of establishment

Introductions of infected individuals into a susceptible population do not always result in a local infection cluster because of random extinction events. Here, we quantify the probability that the introduction of a single infected individual results in the establishment of a local cluster. The establishment probability of a local cluster depends on the probability distribution of the number of secondary infections, which also determines the chance of superspreading events. Similarly, the initial spread of new variants can be modeled just as emerging clusters. In the following, we briefly outline how to compute the probability of establishment and study how it is affected by different types of transmission distributions.

In our epidemiological model, we assume that the number of susceptible individuals is not limiting the spread of the disease, i.e., that the fraction of susceptible individuals in the population remains close to one. Every infected individual transmits the disease to R other individuals on average, where R is the effective reproduction number. The actual number of secondary infections can vary strongly between infected individuals. For example, estimates for COVID-19 indicate that about 20% of infected individuals are responsible for about 80% of secondary infections (Endo et al., 2020). These superspreaders (or superspreading events) cannot be captured by a Poisson-distributed number of secondary infections (Lloyd-Smith et al., 2005). A more dispersed distribution, i.e. with a larger variance, is the negative binomial distribution, where most of infected individuals do not transmit the disease at all. Its variance is typically quantified by the dispersion parameter $\kappa > 0$. The smaller the value of κ, the more variance has the negative binomial distribution.

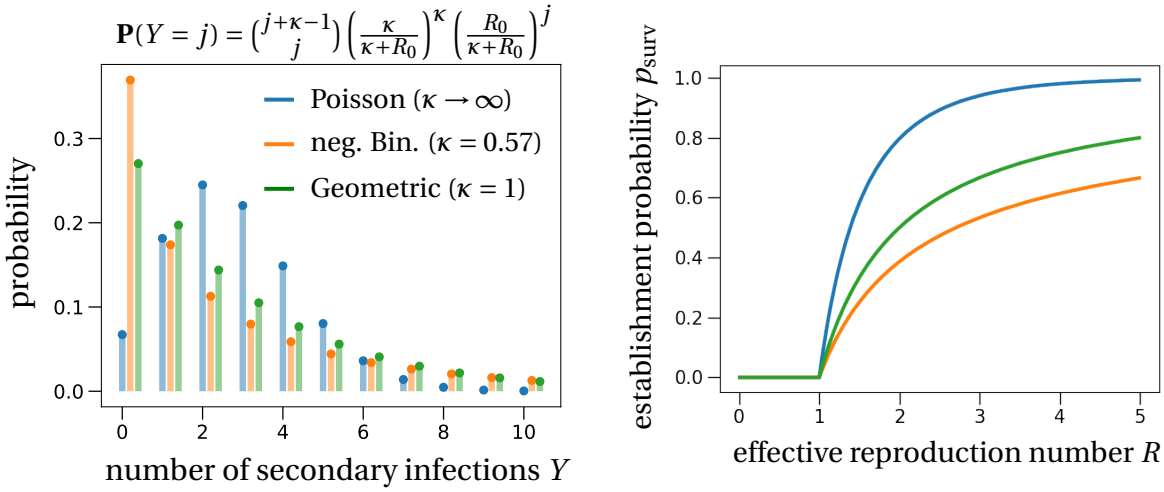
The establishment probability of a local cluster can be computed by following the transmission chain from one generation of infected individuals to the next. Specifically, if Y is the random number of secondary infections due to one infected individual, the probability of extinction, i.e., 1 minus the probability of survival, is computed by summing over the possible numbers of secondary infections times the probability that all corresponding chains of transmission do not survive:

$$p_{\text{ext}} = \underbrace{\mathbf{P}(Y=0)}_{\substack{\text{no secondary}\\ \text{infection}}} + \underbrace{p_{\text{ext}}\mathbf{P}(Y=1)}_{\substack{\text{one secondary}\\ \text{infection}}} + \underbrace{p_{\text{ext}}^2\mathbf{P}(Y=2)}_{\substack{\text{two secondary}\\ \text{infections}}} + p_{\text{ext}}^3\mathbf{P}(Y=3) + \cdots = \mathbf{E}[p_{\text{ext}}^Y]\,. \tag{S1}$$

The expectation on the right, $\mathbf{E}[z^Y]$, is called the probability generating function and can be computed for several distributions explicitly. Eq. (S1) then shows that the probability of extinction p_{ext} of a cluster started with a single infected individual, is given by the smallest positive fixed point of the probability generating function, i.e., $\mathbf{E}[z^Y] = z$ (Haccou et al., 2005). If the epidemic is started with k infected individuals, the extinction probability is simply given by p_{ext}^k, which yields a survival probability of $p_{\text{surv}} = 1 - p_{\text{ext}}^k$. Here, and also in the first equality of Eq. (S1), we have used the assumption that the transmission chains of the k initially infected individuals are independent of each other.

Plugging in different distributions for the number of secondary infections, we can numerically compute the probability of extinction. In the particular case of a geometric distribution, which is a negative binomial distribution with dispersion parameter $\kappa = 1$ (compare to Eq. (S3) below), the fixed point equation can be solved analytically by

$$\mathbf{E}[z^Y] = \frac{p}{1-(1-p)z} \quad \Rightarrow \quad p_{\text{ext}} = \frac{p}{1-p} = \frac{1}{R}, \tag{S2}$$



(a) Distribution of secondary infections

(b) Establishment probability

Figure S1: **Probability distribution and establishment probability of an epidemic for different distributions of the number of secondary infections.** (a) The probability for an infected individual to transmit the disease to Y susceptible individuals depends on the distribution of the number of secondary infections. The probability to not transmit the disease ($Y = 0$) is highest for the negative binomial distribution with $\kappa = 0.57$ (orange bars; value estimated for the French COVID-19 data pre-lockdown 2020; Salje et al. (2020)), i.e., the distribution with the highest variance. The probability of establishment is smallest for the offspring distribution with the largest variance, as shown in (b). For $\kappa = 1$, the geometric distribution (green curve), the establishment probability is explicit: $p_{\mathrm{surv}} = 1 - 1/R$. The largest establishment probability is found for a Poisson-distributed number of secondary transmission events (blue curve).

where p is the success probability of the geometric distribution and R is the average number of secondary infections (or effective reproduction number). The last equality in Eq. (S2) is obtained by noting that R is equal to the mean of the distribution, and that the mean of a geometric distribution is $(1-p)/p$.

The probability generating functions for the negative binomial and the Poisson distributions are given by

$$\mathbf{E}[z^Y] = \begin{cases} \left(\frac{p}{1-(1-p)z}\right)^{\kappa}, & \text{neg. binomial dist. with mean } R \text{ and dispersion parameter } \kappa \\ & \text{(success probability } p = \frac{\kappa}{\kappa+R}\text{)}, \\ e^{R(z-1)}, & \text{Poisson distribution with mean } R\,. \end{cases} \tag{S3}$$

Note that the Poisson distribution is obtained from the negative binomial distribution in the limit $\kappa \to \infty$. The fixed point of the probability generating functions of a negative binomial or a Poisson distribution cannot be computed analytically.

In Fig. S1, we plot the distributions and establishment probabilities as a function of R for the three different distributions of number of secondary infections. For R smaller than 1, the epidemic will not establish, so that the establishment probability is 0. For values of the effective reproduction number R greater than 1, the establishment probability becomes positive and the epidemic has a chance to establish. In general, the smaller the offspring variance, i.e., the larger the dispersion

parameter κ of the negative binomial distribution, the larger the probability of establishment. In applications, the analytically exact result $1/R$ (Eq.(S2)) is often used to approximate the extinction probability of an epidemic. Fig. S1 shows that this is indeed a reasonably good approximation for overdispersed offspring distributions such as the negative binomial distribution with a parameter $\kappa < 1$, at least as long as R does not become too large.

S2 Time-since-infection dependent probability of extinction

In this section, we compute the extinction probability of the epidemic when started with a single infected individual with time since infection a (also called 'age of infection'). In the previous section, we have only considered the special case $a = 0$.

We consider a branching process approximation of the epidemic. We assume that each individual is characterized by its time since infection and we set $\tau(a)$ as the mean infectiousness at age a, i.e.,

$$\tau(a)da \;=\; \mathbf{P}(\text{individual with time since infection in } [a, a+da] \text{ infects a new individual}). \tag{S4}$$

For $a > 0$, let $p_{\text{ext}}(a)$ be the probability that the epidemic starting with a single individual with time since infection a does not establish (in the previous section we have computed $p_{\text{ext}}(0)$). More formally, if Z_t denotes the number of individuals infected at time t, then

$$p_{\text{ext}}(a) \;=\; \mathbf{P}_a\left(\limsup Z_t < \infty\right) = \sum_{k=0}^{\infty} p_{\text{ext}}(0)^k \mathbf{P}(Y_a = k), \tag{S5}$$

where $\mathbf{P}_a$ is the probability distribution of the branching process starting from a single individual with age a and Y_a denotes the number of future secondary infections of an individual with age a. From this, in the case of a Poisson-distributed number of secondary infections, we find

$$\frac{d}{da} p_{\text{ext}}(a) \;=\; \tau(a)\,(1 - p_{\text{ext}}(0))\, p_{\text{ext}}(a). \tag{S6}$$

As a consequence

$$p_{\text{ext}}(a) \;=\; p_{\text{ext}}(0) \exp\left((1 - p_{\text{ext}}(0)) \int_0^a \tau(s)ds\right), \tag{S7}$$

where $p_{\text{ext}}(0)$ can be determined by the boundary condition

$$\lim_{a\to\infty} p_{\text{ext}}(a) = 1\,, \tag{S8}$$

which is ensured by the fact that $R < \infty$. (Intuitively, an individual infected a long time ago is not infectious anymore.) We then find $p_{\text{ext}}(0)$ as the unique solution of

$$p_{\text{ext}}(0) \exp\left((1 - p_{\text{ext}}(0))R\right) = 1, \text{ where } p_{\text{ext}}(0) < 1. \tag{S9}$$

Note that the expression of $p_{\text{ext}}(0)$ is the same as the one stated the previous section in the case of a Poisson-distributed number of secondary infections.

S3 Derivation of the asymptotic limit

We consider a Crump-Mode-Jagers process, also referred to as a general branching process, with an unspecified distribution of secondary infections to model the number of infected individuals over time. The random variable for the number of secondary infections is denoted by Y. We define τ as the intensity measure of the process, i.e., such that for every test function f

$$\mathbf{E}\left[\sum_{i=1}^{Y} f(t_i)\right] = \int_0^\infty f(u)\tau(u)du \tag{S10}$$

and the correlation intensity measure c such that

$$\mathbf{E}\left[\mathbb{1}_{Y\geq 2}\sum_{i\neq j} f(t_i,t_j)\right] = \int_0^\infty f(u,v)c(u,v)dudv\,. \tag{S11}$$

We assume the Kesten-Stigum condition

$$\int_0^\infty u\ln(u)\tau(u)du < \infty\,. \tag{S12}$$

Under this assumption, there exists a random variably W_∞ such that

$$\exp(-rt)Z_t \;\to\; W_\infty \;\text{ almost surely,} \tag{S13}$$

where r is the Malthusian parameter as defined in Eq. (4) in the main text, and Z_t is the number of infected individuals at time t.

Let A be the non-extinction event. We know that $\mathbf{E}[W_\infty] = 1$ and that almost surely $A = \{W_\infty > 0\}$. It follows that

$$\mathbf{E}[W_\infty|A] = \frac{1}{1-p_{\text{ext}}}\,, \tag{S14}$$

where p_{ext} is the extinction probability, which is the smallest root of the probability generating function as defined in Eq. (S1) in the main text:

$$p_{\text{ext}} = \mathbf{E}[p_{\text{ext}}^Y]. \tag{S15}$$

Combining Eqs. (S13) and (S14) gives the formal justification for the correction of the initial epidemic size by $1/(1-p_{\text{ext}})$ in Eq. (7) in the main text.

Further, if $\psi(t) = \mathbf{E}[e^{tY}]$ is the Laplace transform (or moment-generating function) of W_∞, then ψ satisfies the fixed point problem (Cohn, 1985, Theorem 4.1)

$$\text{for all } u\geq 0,\;\; \psi(u) = \mathbf{E}\left[\prod_{i=1}^{Y}\psi(ue^{-rt_i})\right]. \tag{S16}$$

Let us now evaluate the variance of W_∞ conditional on non-extinction. Differentiating the Laplace transform ψ twice with respect to u, we get

$$\begin{aligned}\psi''(u) &= \mathbf{E}\left[\mathbb{1}_{Y\geq 1}\sum_i e^{-2rt_i}\psi''(ue^{-rt_i})\prod_{j\neq i}\psi(ue^{-rt_j})\right] + \\ &\quad \mathbf{E}\left[\mathbb{1}_{Y\geq 2}\sum_{i\neq j} e^{-rt_i}e^{-rt_j}\psi'(ue^{-rt_i})\psi'(ue^{-rt_j})\prod_{k\neq i,j}\psi(ue^{-rt_k})\right].\end{aligned}$$

Evaluating at $u = 0$, we find

$$\mathbf{E}[W_\infty^2] \;=\; \mathbf{E}[W_\infty^2]\mathbf{E}\left[\mathbb{1}_{Y\geq 1}\;\sum_i e^{-2rt_i}\right] \;+\; \psi'(0)^2\mathbf{E}\left[\mathbb{1}_{Y\geq 2}\;\sum_{i\neq j} e^{-rt_i}e^{-rt_j}\right].$$

Since $\psi'(0) \;=\; 1/(1-p_{\text{ext}})$, this yields

$$\mathbf{E}[W_\infty^2] \;=\; \frac{1}{(1-p_{\text{ext}})^2}\frac{\mathbf{E}\left[\mathbb{1}_{Y\geq 2}\;\sum_{i\neq j} e^{-rt_i}e^{-rt_j}\right]}{1-\mathbf{E}\left[\mathbb{1}_{Y\geq 1}\;\sum_i e^{-2rt_i}\right]}.$$

With $A = \{W_\infty > 0\}$ as above, which coincides with the event of non-extinction, we find

$$\begin{aligned}\mathbf{E}[W_\infty^2|A] \;&=\; \left(\frac{1}{1-p_{\text{ext}}}\right)\frac{\mathbf{E}\left[\mathbb{1}_{Y\geq 2}\;\sum_{i\neq j} e^{-rt_i}e^{-rt_j}\right]}{1-\mathbf{E}\left[\mathbb{1}_{Y\geq 1}\;\sum_i e^{-2rt_i}\right]}\\ &=\; \left(\frac{1}{1-p_{\text{ext}}}\right)\;\frac{\int_0^\infty \exp(-r(t+u))c(t,u)dtdu}{1-\int_0^\infty \exp(-2rt)\tau(t)dt}.\end{aligned}$$

Then

$$\mathbf{Var}[W_\infty^2|A] \;=\; \frac{1}{1-p_{\text{ext}}}\left(\frac{\int_0^\infty \exp(-r(t+u))c(t,u)dtdu}{1-\int_0^\infty \exp(-2rt)\tau(t)dt} - \frac{1}{1-p_{\text{ext}}}\right).$$

S4 Renewal equation in the absence of conditioning

We informally derive the renewal equation of the incidence and the epidemic size as given in Eqs. (2) and (3) in the main text.

In the following, σ_x denotes the time of infection of individual x and $\mathcal{F}_t$ denotes the history of the infection process up to time t. (In the probabilistic jargon, $(\mathcal{F}_t)_{t\geq 0}$ is the natural filtration associated with the infection process). Let us define the random empirical measure, which counts the number of infections in a small time interval $[t, t+dt]$

$$db(t) \;=\; \sum_{x:\sigma_x<\infty} \delta_{\sigma_x}(dt), \tag{S17}$$

where δ_{σ_x} is the Dirac measure. The random empirical measure db records the infection times along the course of the epidemic. Each atom of the random measure db corresponds to a time of infection. Note that since the epidemic is triggered by a single individual with time since infection (or 'age') 0 at time 0, db has a Dirac mass at 0. For $t > 0$, we can assume without loss of generality that $\mathbf{E}[db]$ has no atom and we write

$$\text{for all } t > 0, \;\; i(t) := \frac{\mathbf{E}[db(t)]}{dt}\,, \tag{S18}$$

where $i(t)$ is the overall rate at which new individuals are infected in a small time interval $[t, t+dt]$. In other words, $i(t)dt$ models the incidence in the time interval $[t, t+dt]$.

Our goal is to show that i satisfies the renewal equation

$$i(t) \;=\; \tau(t) \;+\; \int_0^t \tau(t-u)i(u)du\,. \tag{S19}$$

Informally, we have

$$\mathbf{E}\Big[db(t) \mid \mathcal{F}_t\Big] \;=\; \sum_{\sigma_x<t} dt \times \tau(t-\sigma_x)\,. \tag{S20}$$

In words, if we condition on the history of the process up to time t, the expected number of new infections in the time interval $[t, t+dt]$ can be computed by adding up the expected contribution of every infected individual before time t. For individual x, this expected contribution is $dt \times \tau(t-\sigma_x)$ and the previous equation follows by adding up the contribution of every previously infected individual.

Next, the sum on the right hand side of the latter equation can be rewritten in terms of an integral with respect to the infection measure db in the following way

$$\sum_{\sigma_x<t} dt \times \tau(t-\sigma_x) \;=\; dt \int_{[0,t)} \tau(t-u)db(u)\,, \tag{S21}$$

so that

$$\mathbf{E}\Big[db(t) \mid \mathcal{F}_t\Big] \;=\; dt \int_{[0,t)} \tau(t-u)db(u)\,. \tag{S22}$$

Averaging over every realization of the process up to time t, we obtain

$$\begin{aligned} \mathbf{E}\Big[db(t)\Big] &= dt\;\mathbf{E}\Big[\int_{[0,t)} \tau(t-u)db(u)\Big] \\ &= \; dt \int_{[0,t)} \tau(t-u)\mathbf{E}[db(u)]\,. \end{aligned} \tag{S23}$$

Recalling our definition of $i(t)$ in Eq. (S18), this yields

$$i(t) \;=\; \tau(t) \;+\; \int_0^t \tau(t-u)i(u)du\,, \tag{S24}$$

where the term $\tau(t)$ on the right hand side follows from the fact that db has a Dirac mass at 0. This is Eq. (2) in the main text.

The cumulative epidemic size $I(t) := 1 + \int_0^t i(s)ds$ then satisfies the renewal equation

$$\begin{aligned} I(t) &= 1 + \int_0^t i(s)ds \\ &= 1 + \int_0^t \big(\tau(s) + \int_0^s i(s-u)\tau(u)du\big)ds \\ &= 1 + \int_0^t \tau(s)ds + \int_0^t \tau(u) \int_u^t i(s-u)ds\, du \\ &= 1 + \int_0^t \tau(s)ds + \int_0^t \tau(u) \underbrace{\int_0^{t-u} i(s)ds}_{=I(t-u)-1} du \\ &= 1 + \int_0^t \tau(a)I(t-u)du\,, \end{aligned} \tag{S25}$$

where we have used Fubini's theorem to change the order of integration. This explains Eq. (3) in the main text.

S5 Delay differential equation for the process conditioned on explosion

In this section, we derive a delay differential equation for the rate of new infections over time when the process is conditioned on non-extinction. This delay differential equation explains Eq. (9) in the main text.

Markov model. For simplicity, we assume that the infection process is Markovian in its age structure, where by age structure we refer to the time since infection composition in the infected population. More precisely, this means that the times since infection for a given individual are given by a Poisson point process with intensity τ. At time $t > 0$, we encode the population by a random vector $\mathbf{a} = (a_1, a_2, \cdots, a_n)$ where $a_1 > \cdots > a_n$ are the times since infection listed in decreasing order (individual 1 is the first infected individual etc.), and n is the number of infected individuals at time t. We denote by $|\mathbf{a}|$ the number of elements in $\mathbf{a}$. The infinitesimal generator of the infection process is then given by

$$Lf(\mathbf{a}) \;=\; \sum_{i=1}^{|\mathbf{a}|} \left[\tau(a_i) \Big(f((\mathbf{a}, 0)) - f(\mathbf{a}) \Big) \;+\; \partial_i f(\mathbf{a}) \right], \tag{S26}$$

where $(\mathbf{a}, 0) := (a_1, \cdots, a_{|\mathbf{a}|}, 0)$ and f is a continuous and bounded test function (e.g Ethier and Kurtz, 1986). Intuitively, one can think of the infinitesimal generator to be the expected rate of change of a Markov process, which generalizes the concept of the transition rate matrix in the case of a discrete state continuous time Markov chain. It encodes the infinitesimal dynamics of the branching process in the following way: on a time interval dt,

- an individual with time since infection a_i produces a new infection with a probability $\tau(a_i)dt$. Conditional on this event, the infection age (time since infection) structure of the population makes a transition from $\mathbf{a}$ to $(\mathbf{a}, 0)$, i.e., an individual with time since infection 0 is added to the population.
- the time since infection of every individual increases by dt (this corresponds the partial derivative of f in the generator).

The Doob h-transform. Let us now consider the probability of survival, which is equivalent to explosion ($Z_t \to \infty$ for $t \to \infty$),

$$h(\mathbf{a}) \;:=\; 1 - \prod_{i=1}^{|\mathbf{a}|} p_{\text{ext}}(a_i). \tag{S27}$$

for a population with initial time-since-infection structure $\mathbf{a}$. Applying Eq. (S6), we note that h is a harmonic function for the generator L, i.e., $Lh = 0$. Let us consider the Markov process with infinitesimal generator

$$\widetilde{L}f = \frac{1}{h} L(fh)\,. \tag{S28}$$

This is the Doob h-transform of the original process.

Branching process conditioned on explosion. The Markov process with generator $\widetilde{L}$ is known to be identical in law with the original process conditioned on explosion (Chetrite and Touchette, 2015; Doob, 1957). Using $Lh = 0$, the transformed infinitesimal generator computes to

$$\begin{aligned}\widetilde{L}f(\mathbf{a}) &= \sum_{i=1}^{|\mathbf{a}|}\left[\tau(a_i)\frac{h((\mathbf{a},0))}{h(\mathbf{a})}\Big(f((\mathbf{a},0)) - f(\mathbf{a})\Big) \;+\; \partial_i f(\mathbf{a})\right] \\ &= \sum_{i=1}^{|\mathbf{a}|}\left[\tau(a_i)\frac{1-p_{\text{ext}}(0)\prod_{i=1}^{|\mathbf{a}|}p_{\text{ext}}(a_i)}{1-\prod_{i=1}^{|\mathbf{a}|}p_{\text{ext}}(a_i)}\Big(f((\mathbf{a},0)) - f(\mathbf{a})\Big) \;+\; \partial_i f(\mathbf{a})\right].\end{aligned} \tag{S29}$$

The formula can now be interpreted as follows. For a given time-since-infection structure $\mathbf{a}$, conditioning on explosion induces an adjustment of the time since infection of the ith individual as follows

$$\widetilde{\tau}^{\mathbf{a}}(a_i) \;=\; \tau(a_i)\;\cdot\;\underbrace{\frac{1-p_{\text{ext}}(0)\prod_{j=1}^{|\mathbf{a}|}p_{\text{ext}}(a_j)}{1-\prod_{j=1}^{|\mathbf{a}|}p_{\text{ext}}(a_j)}}_{\text{adjustment}}. \tag{S30}$$

We note that the infection rate is increased for every individual, independent of their time since infection, by the same amount. The increase, however, depends on the time-since-infection structure of the entire infected population.

The delay equation. We keep the same notation as in Section S4. By the same reasoning as in Section S4, we have

$$\mathbf{E}[db(t)|\mathcal{F}_t] \;=\; \sum_{\sigma_x<t}\left(\frac{1-p_{\text{ext}}(0)\prod_{y:\sigma_y<t}p_{\text{ext}}(t-\sigma_y)}{1-\prod_{y:\sigma_y<t}p_{\text{ext}}(t-\sigma_y)}\right)dt\times\tau(t-\sigma_x)\,. \tag{S31}$$

In words, Eq. (S31) means that if we condition on the history of the process up to time t, the expected number of new infections in the time interval $[t, t+dt]$ can be computed by adding up the expected contribution of every infected individual before time t. According to Doob's h-transform formula, for an individual infected at time σ_x, the expected contribution of this individual is given by $dt \times \tau(t-\sigma_x)$ multiplied by the Doob's term

$$\left(\frac{1-p_{\text{ext}}(0)\prod_{y:\sigma_y<t}p_{\text{ext}}(t-\sigma_y)}{1-\prod_{y:\sigma_y<t}p_{\text{ext}}(t-\sigma_y)}\right), \tag{S32}$$

which depends on the whole time-since-infection structure of the population up to time t.

We now rewrite the adjustment factor in the following way

$$\begin{aligned}\frac{1-p_{\text{ext}}(0)\prod_{y:\sigma_y<t}p_{\text{ext}}(t-\sigma_y)}{1-\prod_{y:\sigma_y<t}p_{\text{ext}}(t-\sigma_y)} &= \frac{1-p_{\text{ext}}(0)\exp\Big(\sum_{y:\sigma_y<t}\ln(p_{\text{ext}}(t-\sigma_y))\Big)}{\exp\Big(\sum_{y:\sigma_y<t}\ln(p_{\text{ext}}(t-\sigma_y))\Big)} \\ &= \frac{1-p_{\text{ext}}(0)\exp\Big(\int_{[0,t)}\ln(p_{\text{ext}}(t-u))db(u)\Big)}{1-\exp\Big(\int_{[0,t)}\ln(p_{\text{ext}}(t-u))db(u)\Big)}.\end{aligned} \tag{S33}$$

As a consequence, Eq. (S31) can be rewritten as

$$\begin{aligned}\mathbf{E}[db(t)|\mathcal{F}_t] &= dt\left(\sum_{x:\sigma_x<t}\tau(t-\sigma_x)\right)\times\frac{1-p_{\text{ext}}(0)\exp\left(\int_{[0,t)}\ln(p_{\text{ext}}(t-u))db(u)\right)}{1-\exp\left(\int_{[0,t)}\ln(p_{\text{ext}}(t-u))db(u)\right)}\\ &= dt\int_{[0,t)}\tau(t-u)db(u)\times\frac{1-p_{\text{ext}}(0)\exp\left(\int_{[0,t)}\ln(p_{\text{ext}}(t-u))db(u)\right)}{1-\exp\left(\int_{[0,t)}\ln(p_{\text{ext}}(t-u))db(u)\right)}.\end{aligned} \tag{S34}$$

Averaging over the past of the infection process $\mathcal{F}_t$, we find that

$$\mathbf{E}[db(t)] = dt\ \mathbf{E}\left[\int_{[0,t)}\tau(t-a)db(a)\times\frac{1-p_{\text{ext}}(0)\exp\left(\int_{[0,t)}\ln(p_{\text{ext}}(t-a))db(a)\right)}{1-\exp\left(\int_{[0,t)}\ln(p_{\text{ext}}(t-a))db(a)\right)}\right]. \tag{S35}$$

Since db has a Dirac at 0 we have

$$\mathbf{E}[db(t)] = dt\mathbf{E}\left[\left(\tau(t)+\int_0^t\tau(t-u)db(u)\right)\times\frac{1-p_{\text{ext}}(0)p_{\text{ext}}(t)\exp\left(\int_0^t\ln(p_{\text{ext}}(t-u))db(u)\right)}{1-p_{\text{ext}}(t)\exp\left(\int_0^t\ln(p_{\text{ext}}(t-u))db(u)\right)}\right]. \tag{S36}$$

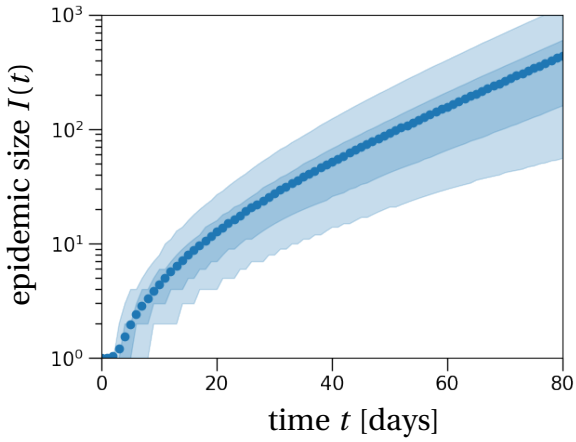
Let us now compare this formula with its analog in Eq. (S23) in the absence of conditioning the process on survival. In Eq. (S23), we can use Fubini's theorem to interchange the integral and the expected value. This provides the renewal equation for the average infection measure $i(t)$ – see Eq. (S19). In the presence of conditioning in Eq. (S36), the right hand side is not linear in db anymore and we cannot derive an autonomous equation for the density $i(t)$ as in the unconditioned case. Our approximation now consists in ignoring the non-linearities on the right hand side of Eq. (S36) and replacing $db(t)$ by its expected value. This gives

$$\begin{aligned}\widetilde{i}(t) &= \left(\tau(t)+\int_0^t\tau(t-u)\widetilde{i}(u)du\right)\times\frac{1-p_{\text{ext}}(0)p_{\text{ext}}(t)\exp\left(\int_0^t\ln(p_{\text{ext}}(t-u))\widetilde{i}(u)du\right)}{1-p_{\text{ext}}(t)\exp\left(\int_0^t\ln(p_{\text{ext}}(t-u))\widetilde{i}(u)du\right)}\\ &= F(\widetilde{i}(s);s\in[0,t]),\end{aligned} \tag{S37}$$

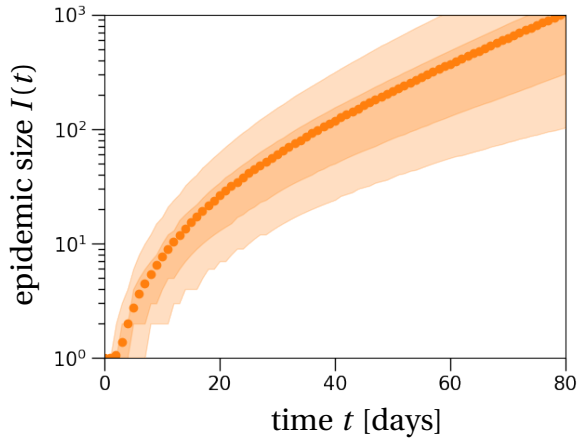
which explains Eq. (9) in the main text. The overall epidemic size of the conditioned process is then given by $\widetilde{I}(t)=1+\int_0^t\widetilde{i}(s)ds$.

S6 Cumulative epidemic size over time

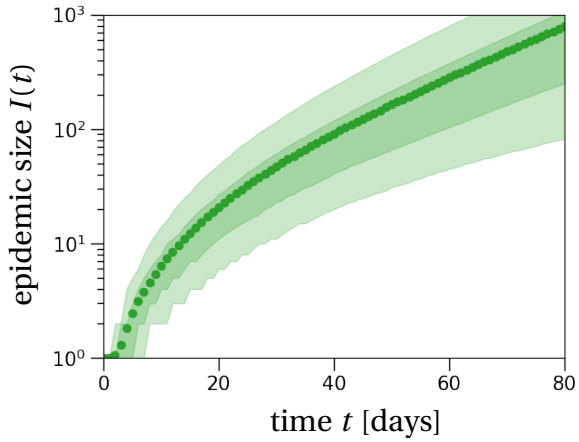
In the main text, we have seen that the cumulative epidemic size over time varies a lot between different stochastic simulations (Fig. 1 in the main text). The number of secondary infections was Poisson-distributed in the main text. The negative binomial and geometric distribution exhibit more variance than a Poisson distribution. Therefore, the variance of the cumulative epidemic size obtained from 10,000 stochastic simulations is even larger in these cases (compare the subfigures in Fig. S2). This can be explained as follows: First, super-spreading events are more likely and as a result, epidemic sizes can be much larger. Secondly, and conversely, a larger number of infected individuals do not transmit the infection at all; as a result, the total epidemic size can remain smaller.



(a) Poisson



(b) negative binomial



(c) geometric

Figure S2: **The cumulative number of infected individuals over time.** The light and dark shaded regions show the 90% and 50% inter-quantile ranges obtained from 10,000 stochastic simulations that resulted in establishment of an epidemic cluster. Dots show the average of these simulations over time. With a negative binomial offspring distribution (b), a lot of infected individuals do not transmit the disease, which results in lower epidemic sizes than compared to a Poisson offspring distribution (a). In contrast, few infected individuals with a lot of secondary infections early in the epidemic can generate much larger epidemic sizes when compared to the Poisson distribution. These two effects together explain the much larger variance in epidemic sizes for the negative binomial distribution compared to the Poisson distribution. For a geometric offspring distribution (c), the mean epidemic size and amount of variation are between the numbers found by the negative binomial and Poisson cases. This is explained by the dispersal parameter of the geometric distribution being $\kappa = 1$, which is between the respective values for our choice of the negative binomial offspring distribution ($\kappa = 0.57$) and the limit $\kappa \to \infty$ corresponding to the Poisson case. The effective reproduction number is set to $R = 1.3$ and the transmission density $\mu(t)$ is a Gamma distribution with the parameters as stated in Table 1 in the main text.

S7 Detection rate of a single mass testing effort

A good measure to evaluate the prevalence of the disease in a population is the number of detected cases, which depends on the testing effort. In the context of COVID-19, the average detection rate on the 8th of May 2020 across a large number of countries, mainly in Europe and North America, was inferred to be around 30% at best (Belloir and Blanquart, 2021; Russell et al., 2020). However, test capacity is not the only limiting factor to detect infected individuals. The probability for an infected individual to test positive by a reverse polymerase chain reaction or a lateral flow test (LFT), which we will more loosely refer to as rapid tests, varies over the course of the individual's infection (Borremans et al., 2020; Hellewell et al., 2021; Kucirka et al., 2020). Here, we will use the data obtained for rapid tests in Hellewell et al. (2021, Fig. 4B) and for reverse transcriptase polymerase chain reaction (RT-PCR) tests in Kucirka et al. (2020, Fig. 2), as plotted in Fig. S3. During the first three days of an infection, the viral load within the infected individual is too low to reliably detect the virus with both testing procedures. The probability of detection then reaches a maximum around day four post-infection and then gradually declines. The decline is much faster for rapid tests than for RT-PCR tests. In the following, we denote the probability to test positive on day a after infection by $Q(a)$.

Using the probabilities for testing positive by a rapid test and an RT-PCR test (Fig. S3), we compute an upper bound for the fraction of detectable infectious individuals within the epidemic outbreak. Note that this result does not rely on our stochastic correction of the deterministic dynamics, nor does it depend on the epidemic size over time. Instead, the result only depends on the stationary time-since-infection distribution (also called infection age distribution), which is determined by the Malthusian growth rate r. Therefore, this result is true for any exponentially growing epidemic.

Again using general branching process theory (Haccou et al., 2005; Jagers, 1969), we find that the number of detectable individuals in the population at time t is given as a fraction q of the overall epidemic size as given in Eq. (8) in the main text. The fraction q is defined as

$$q = \frac{1}{C} \int_0^T Q(a) r e^{-ra} da \,, \tag{S38}$$

where $C = \int_0^T r e^{-ra} da = 1 - e^{-Tr}$ is the proportion of individuals that have been infected less than T days ago. The integral in Eq. (S38) computes the fraction of individuals that are detectable (term $Q(a)$) integrated over the stationary time-since-infection distribution (term re^{-ra}) of the population, restricted to the time since infection being less than T days (constant C). For the numerical example, we assume that after $T = 14$ days infected individuals are not infectious anymore. The resulting truncated distributions for $R = 2.9$ and $R = 1.3$ are plotted in Fig. S3 (dashed and dotted line).

With a high reproduction number $R = 2.9$, we find $q \approx 0.255$ (evaluating a discretized version of the integral) when testing is done by rapid tests, so that we would expect that only about 25.5% of the infected individuals would test positive with a rapid test at each point in time. Using the RT-PCR data instead, we find $q = 0.29$. These estimates strongly depend on the combination of the probability to test positive $Q(a)$ and the stationary time-since-infection distribution. The stationary time-since-infection distribution itself depends on the Malthusian growth rate r (Eq. (4) in the main text). If we reduce the reproduction number to $R = 1.3$, the detection rates become $q = 0.265$ and $q = 0.48$ for rapid tests and RT-PCR tests, respectively. Interestingly, the fraction of

detectable cases does not change in the rapid test scenario substantially, which is a robust result for all possible reproduction numbers between 1.1 and 3. This seems to be due to our choice to only consider individuals with a time since infection less than 14 days. For other choices of this infectiousness threshold, the difference between the detection rates for the two reproduction numbers is more notable. In case of the RT-PCR test, the detection rate increased a lot with a lower reproduction number, from 29% for $R = 2.9$ to 48% for $R = 1.3$. The explanation for this increase in detection rate is that in slowly growing epidemics, more individuals are infected a long time ago and remain detectable with RT-PCR, whereas in faster growing epidemics many individuals are not yet detectable because they were infected within the last two days.

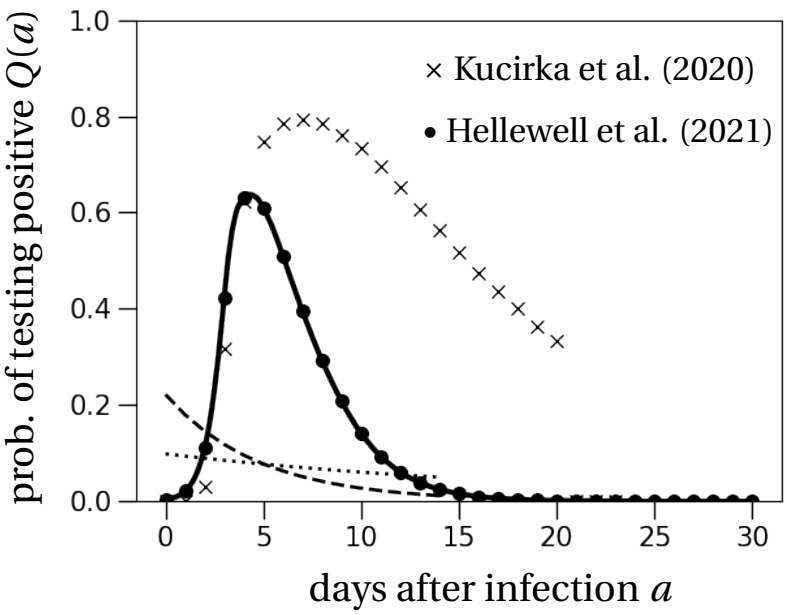


Figure S3: **Probability of testing positive.** The probability for an infected individual to test positive by a rapid test (bullet points) and a RT-PCR test (crosses) is close to zero during the first two days after infection. It then increases up to approximately 0.6 on day 4 after infection for rapid tests and up to 0.8 on day 8 for RT-PCR tests. The probability of testing positive then decreases. For rapid tests, the probability of testing positive approaches zero by day 21 after infection. For the RT-PCR test, the probability of testing positive remains relatively high at least until day 20, which corresponds to the last estimated time point in Fig. 2 (upper panel) from Kucirka et al. (2020). The probabilities for rapid tests are based on data from Fig. 4B in Hellewell et al. (2021). The dashed and dotted lines show the re-normalized stationary time-since-infection distributions for $R = 2.9$ and $R = 1.3$ (Eq. (S38)).

S8 Testing frequency and cluster size at detection

We study the average epidemic cluster size at detection when varying the daily testing frequency f in a population. Fig. S4 shows that our analytical prediction based on Eqs. (11) and (12) from the main text slightly overestimates the averages as obtained from stochastic simulations. Importantly, the epidemic size at detection reflects the exponential growth of the epidemic and declines exponentially with an increasing daily testing frequency. This indicates that there might be a clear optimal testing frequency that ensures clusters are relatively small when detected while not overwhelming test systems.

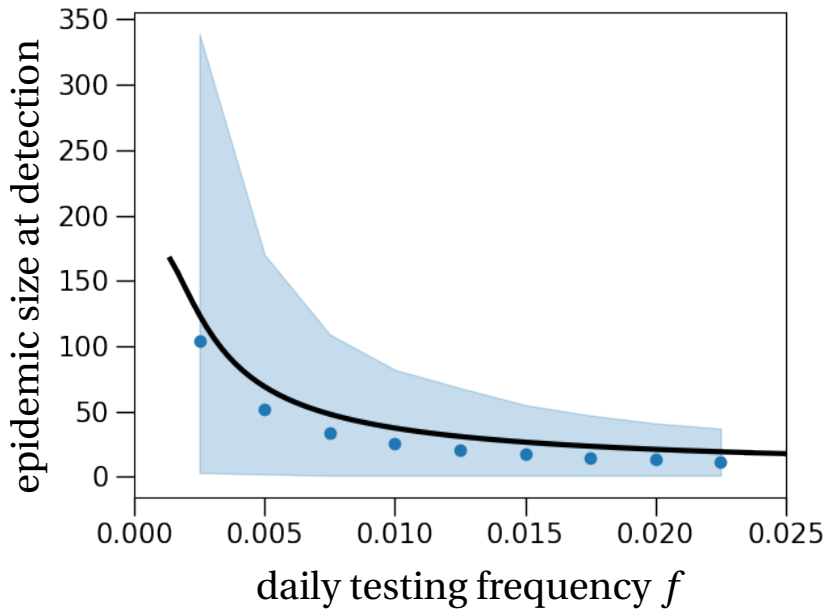


Figure S4: **Average epidemic cluster size at detection depending on the daily testing frequency.** The shaded region shows the 90% confidence interval of the cluster size obtained from 10,000 stochastic simulations that resulted in cluster establishment. Dots represent the average cluster sizes of these simulations at the first detection time. The theoretical prediction (black solid line) is obtained from Eq. (11) in the main text, adapted to the setting of testing. In particular, the distribution until detection is motivated by the probability to test positive in Fig. S3 in SI, Section S7, and translated to a detection probability by Eq. (12) in the main text. The effective reproduction number is set to $R = 1.1$ and the number of secondary infections is Poisson-distributed. The transmission density $\mu(t)$ is as stated in Table 1 in the main text.

References

Belloir, A. and Blanquart, F. Estimating the global reduction in transmission and rise in detection capacity of the novel coronavirus SARS-CoV-2 in early 2020. *Epidemics*, 35:100445, 2021. doi: 10.1016/j.epidem.2021.100445.

Borremans, B., Gamble, A., Prager, K., Helman, S. K., McClain, A. M., Cox, C., Savage, V., and Lloyd-Smith, J. O. Quantifying antibody kinetics and RNA detection during early-phase SARS-CoV-2 infection by time since symptom onset. *eLife*, 9, 2020. doi: 10.7554/elife.60122.

Chetrite, R. and Touchette, H. Nonequilibrium markov processes conditioned on large deviations. *Annales Henri Poincaré*, 16(9):2005–2057, 2015. doi: 10.1007/s00023-014-0375-8.

Cohn, H. A martingale approach to supercritical (CMJ) branching processes. *The Annals of Probability*, 13(4), 1985. doi: 10.1214/aop/1176992803.

Doob, J. Conditional brownian motion and the boundary limits of harmonic functions. *Bulletin de la Société Mathématique de France*, 79:431–458, 1957. doi: 10.24033/bsmf.1494.

Endo, A., Abbott, S., Kucharski, A. J., and Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*, 5:67, 2020. doi: 10.12688/wellcomeopenres.15842.3.

Ethier, S. N. and Kurtz, T. G. *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. J. Wiley & Sons, New York, Chichester, 1986.

Haccou, P., Jagers, P., and Vatutin, V. A. *Branching Processes*. Cambridge University Press, 2005. doi: 10.1017/cbo9780511629136.

Hellewell, J., Russell, T. W., The SAFER Investigators and Field Study Team, The Crick COVID-19 Consortium, CMMID COVID-19 working group, Beale, R., Kelly, G., Houlihan, C., Nastouli, E., and Kucharski, A. J. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC Medicine*, 19(1), 2021. doi: 10.1186/s12916-021-01982-x.

Jagers, P. A general stochastic model for population development. *Scandinavian Actuarial Journal*, 1969(1-2):84–103, 1969. doi: 10.1080/03461238.1969.10405220.

Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., and Lessler, J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine*, 173(4):262–267, 2020. doi: 10.7326/m20-1495.

Lloyd-Smith, J., Schreiber, S., Kopp, P., and Getz, W. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005. doi: 10.1038/nature04153.

Russell, T. W., Golding, N., Abbott, S., Hellewell, J., Pearson, C. A. B., van Zandvoort, K., Jarvis, C. I., Gibbs, H., Liu, Y., Eggo, R. M., Edmunds, J. W., and Kucharski, A. J. Reconstructing the global dynamics of under-ascertained COVID-19 cases and infections. *medRXiv preprint*, 2020. doi: 10.1101/2020.07.07.20148460.

Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Andronico, A., Hozé, N., Richet, J., Dubost, C.-L., and et al. Estimating the burden of SARS-CoV-2 in France. *Science*, 2020. doi: 10.1126/science.abc3517.