A stochastic epidemiological model to estimate the size of an outbreak at the first case identification

Peter Czuppon[1,2], François Blanquart[2,3,*], 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] Université de Paris, INSERM, IAME, F-75018 Paris

[*] equal contributions

Abstract. The identification of a first case (e.g. by a disease-related death or hospitalization event) raises the question of the actual size of a local outbreak. Quick estimates of the outbreak size are required to assess the necessary testing, contact tracing and potential containment effort. Using general branching processes and assuming that epidemic parameters (including the basic reproductive number) are constant over time, we characterize the distribution of the first hospitalization time and of the epidemic size at this random time. We find that previous estimates either overestimate or largely underestimate the actual epidemic size. In addition, we provide upper and lower bounds for the number of infectious individuals of the local outbreak over time. The upper bound is the cumulative epidemic size, and the lower bound is a constant fraction of it. Lastly, we compute the number of detectable cases if one were to test the whole local outbreak at a single point in time. In a growing epidemic, most individuals have been infected recently, which can strongly limit the detection of infected individuals when there is a delay between an infection and its potential detection. Overall, our analysis provides new analytical estimates about the epidemic size at identification of a first disease-related case. This piece of information is important to inform policy makers during the early stages of an epidemic outbreak.

Keywords: epidemic dynamics, first event statistics, emerging infectious diseases, COVID-19, stochastic modeling, general branching processes

1 Introduction

Newly emerging infectious diseases, like the severe acute respiratory syndrome (SARS) or the 2009 H1N1 flu, have been a major concern in the 21st century, and especially in 2020 with the coronavirus disease 2019 (COVID-19). In addition, recurrent outbreaks of diseases as Ebola or cholera are responsible for large numbers of fatalities and thus pose a permanent threat. To limit the spread of any of these diseases, early detection of a new local outbreak is of great importance – especially to estimate the level of urgency for containment actions –, but is reliant on active surveillance. Exact prevalence numbers are hard to come by, particularly during the early phase of an outbreak. This is notably the case when there is asymptomatic transmission, or if suitable rapid tests are not yet available. Theoretical predictions may overcome this problem by providing estimates of the epidemic sizes and corresponding confidence intervals. These estimates are important to inform control policies: for example, to decide on the intensity of contact tracing, or to inform indirectly about the scale of the testing effort necessary to find infected individuals, from the immediate surroundings of the first case to much larger scales like neighborhoods or even cities.

The case fatality ratio (CFR) and the probability of hospitalization provide estimates for the size of an outbreak at the first death or hospitalization event, respectively. The idea is the following: The epidemic must have reached size 1/CFR on average – the mean of a geometric distribution with probability CFR – when the first disease-caused fatality can be identified. (We focus here on the first identified death, but the same reasoning holds for an identification at the first hospitalization and the use of an hospitalization probability.) We will refer to this estimate as the '*simple rule of thumb*'. This method underestimates the actual size of the epidemic at the identification of the first case, because it corresponds to the size of the epidemic when the individual that eventually died was infected. The method therefore omits all individuals that were infected between the infection time of the individual that will eventually die and the death of this individual.

This omission can however be compensated for with a method that we call '*improved rule of thumb*'. Taking into account the average time from infection to death together with the mean generation time and the mean number of secondary infections, one can derive the expected epidemic size at the first death event. This was done with a probabilistic framework by Jombart et al. (2020), who estimated the whole distribution of epidemic sizes over multiple realizations of the stochastic epidemiological processes. The authors developed a stochastic simulation algorithm with random secondary infections, random transmission times and random death times to estimate the epidemic sizes of COVID-19 in France, Italy and Spain at the first death event. This improved rule of thumb relies on the assumption that the current epidemic size at the infection of the first detected case is on average 1/CFR. It does not take into account the history of the epidemic. The error would be substantial when the epidemic grows slowly. Here, we develop a method to address this shortcoming and estimate the size of an epidemic at the first identification of a case. We will see that ignoring the epidemic history, as done by the improved rule of thumb, overestimates the actual epidemic size. In contrast, the simple rule of thumb may largely underestimate the epidemic size, especially when the epidemic grows fast. In particular, it is the cumulative epidemic size and not the current epidemic size which on average is of size 1/CFR.

We additionally provide analytical formulas for the distribution of the first identification time and the distribution of the epidemic size at that random time. The previous analyses that we are aware of did not allow for an estimation of the identification time, neither did they derive analytical formulas for the *distribution* of the epidemic size at the time of identification.

We first introduce the model, a general branching process, to describe the early phase of an epidemic. We assume a gamma distribution for the infection times and model the number of secondary infections

by either a Poisson or a negative binomial distribution. Using asymptotic results on general branching processes, we estimate the cumulative epidemic size and give a lower bound on the number of infectious individuals over time. The upper bound is the cumulative epidemic size. Additionally, we study the number of cases detectable by reverse transcriptase polymerase chain reaction (RT-PCR), if we were to sample the whole population at a single point in time. Lastly, we use our results on the cumulative epidemic size to derive an approximation of the first identification time and the epidemic size at this event. For the sake of illustration, we will consider the event 'first hospitalization', but keep in mind that our analysis carries over to any type of quantifiable event, including the first disease-related death.

2 Model

We model the early phase of a newly emerging disease in a fully susceptible population. In the analysis, we assume that the number of susceptible individuals is not limiting the spread of the disease, that is, that the fraction of susceptible individuals in the population remains close to one.

The epidemic starts with a single infected individual at time $t = 0$. Each infected individual i is assigned an infection age a_i measuring the time since the individual was infected. The age of infection determines the transmission potential of an individual through time. We use a gamma distribution to model the transmission rate over time, but any distribution would work in principle. In particular, an exponential distribution for the transmission rate (the memory-less distribution) would translate our model to the framework of ordinary differential equations (ODEs) and the classical 'Susceptible-Infected-Recovered' (SIR) epidemic model.

Every infectious individual infects R_0 other individuals on average, where R_0 is the basic reproduction number (we assume that population immunity is low). The actual number of secondary infections, which we will also refer to as 'offspring', can vary strongly between infected individuals. For example, recent estimates for COVID-19 indicate that about 20% of infected individuals are responsible for about 80% of secondary transmissions (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. In our analysis, we will use both the Poisson and the negative binomial distribution to model the number of secondary infections, and we will highlight the resulting differences for the epidemic dynamics, as done e.g. in Althouse et al. (2020).

We are interested in the population composition at a certain event, the first identification of an infected individual, e.g. the first disease-related death or hospitalization. Here, we parameterize our model with respect to the first hospitalization event. In our model, each infected individual has a certain probability p_{hosp} to eventually be admitted to a hospital. Hospitalization happens at a random time after infection of the individual, denoted t_{hosp}. We model the hospitalization time by a gamma distribution (though again any distribution would be possible).

Fig. 1 illustrates the epidemic dynamics at the individual level.

2.1 Parameterization of the model

We base our simulations on parameters that are, if available, estimated from data of the French COVID-19 epidemic in early 2020. The number of secondary infections (pre-lockdown) was found to be $R_0 = 2.9$ with dispersion parameter $\kappa = 0.57$ (Salje et al., 2020), which is in line with other studies (e.g. Di Domenico et al., 2020; Foutel-Rodier et al., 2020; Roques et al., 2020; Sofonea et al., 2020). The

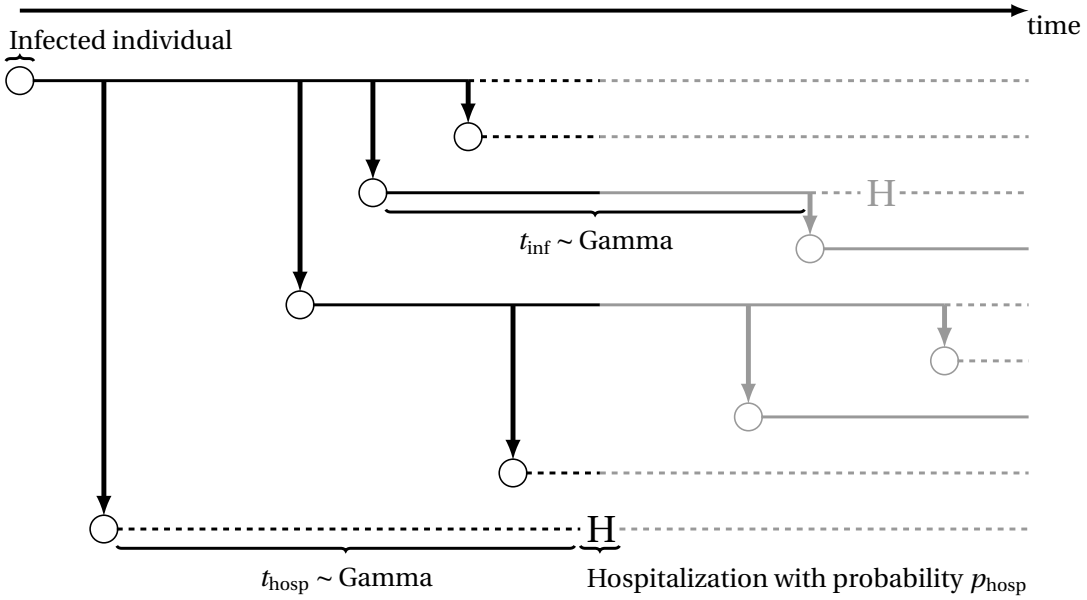


Figure 1: **Schematic view of the model describing the early stage of the epidemic.** Infectious individuals (circles) transmit the disease to a random number of individuals (vertical arrows) at transmission times t_{inf} that are distributed according to a gamma distribution. An infected individual will eventually be hospitalized (depicted by an 'H') with probability p_{hosp}. Hospitalization happens at a gamma-distributed random time t_{hosp} after infection. Individuals that do not transmit the disease anymore are labeled 'non-infectious' (dashed lines); 'infectious' individuals will transmit the disease at some point in the future (solid lines). Dynamics that occur at times after the first hospitalization event (left 'H') are shaded.

infection times t_{inf} are drawn from a gamma distribution with mean 5.5 days and standard deviation 2.14 days (Cheng et al., 2020; Ferretti et al., 2020; He et al., 2020).

Each infected individual has a probability $p_{\text{hosp}} = 0.029$ (Salje et al., 2020) to be hospitalized during the course of their infection. In the case of hospitalization, we draw a hospitalization time t_{hosp} from a gamma distribution with mean 14.4 days and standard deviation 2.58 days estimated from hospitalization events in France (Foutel-Rodier et al., 2020). Similar estimates have been obtained for other data sets (e.g. Faes et al. (2020) for Belgium and Linton et al. (2020) for Wuhan in China).

Table 1 summarizes all the random variables together with their probability distributions and parameter values, as used in the simulations.

3 Results

We use general branching process theory to derive estimates of the mean epidemic size over time and a lower bound on the number of infectious individuals. Additionally, we estimate the number of RT-PCR positive individuals who would be detected if one were to test the whole local population at a single point in time. We then use the analytical result on the epidemic size to estimate the distribution of the first hospitalization time within an epidemic cluster. Finally, we study the *distribution* of the epidemic size at the first hospitalization time and compare it to estimates obtained with the simple or improved rules of thumb.

Variable	Interpretation	Distribution	Parameters	Reference
Y_i	number of secondary infections by individual i	Poisson (negative binomial)	mean: $R_0 = 2.9$ (dispersion: $\kappa = 0.57$)	Salje et al. (2020)
t_{inf}	time of secondary infection	Gamma (density: $\mu(t)$)	shape: 6.6, scale: 0.833 (mean: 5.5 days)	Github - OpenABM-Covid19 (Hinch et al., 2020)
t_{hosp}	time from infection to hospitalization	Gamma (density: $f_{\text{hosp}}(t)$)	shape: 31, scale: 0.463 (mean: 14.4 days)	Foutel-Rodier et al. (2020)
–	hospitalization event	Bernoulli	success probability: $p_{\text{hosp}} = 0.029$	Salje et al. (2020)

Table 1: **Definitions of random variables with corresponding probability distributions and parameter values.**

3.1 The size of the epidemic

Our analysis requires a quantification of the expected size of the epidemic over time. For this, we use the theory of supercritical branching processes. The population size of a supercritical branching process, conditioned on survival, grows exponentially with rate α. This parameter α is referred to as the Malthusian parameter, as described for instance in the books by Athreya and Ney (1972) or Haccou et al. (2005). With our notation, it is implicitly defined through

$$\frac{1}{R_0} = \int_0^\infty e^{-\alpha t}\mu(t)dt\,, \tag{1}$$

where μ is the density of the transmission distribution (cf. Table 1).

It follows from results of supercritical general branching processes and renewal theory, e.g. Haccou et al. (2005, Section 3.3.1), that the expected cumulative epidemic size is, for asymptotically large times t, given by

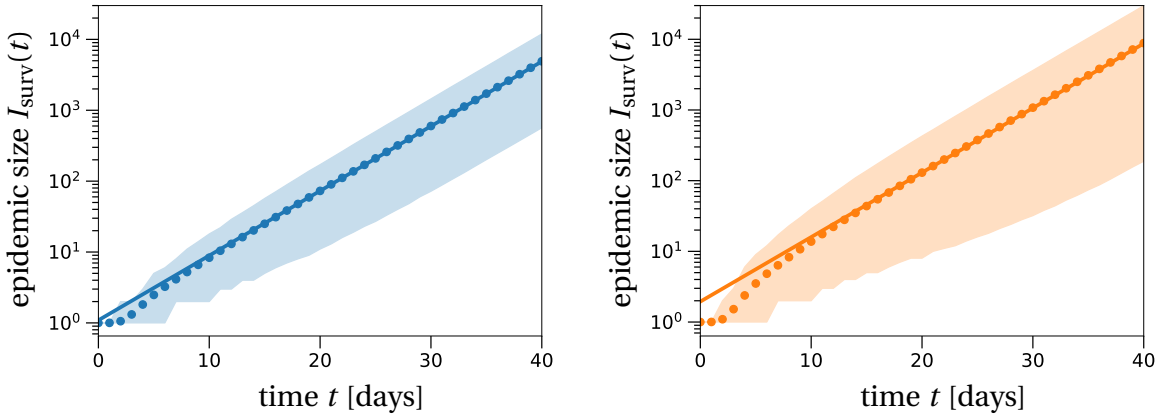
$$I_{\text{tot}}(t) = I(0)\frac{e^{\alpha t}}{\alpha\int_0^\infty R_0 e^{-\alpha s}s\mu(s)ds}\,. \tag{2}$$

The integral in the denominator is the mean generation time of the Malthusian process. If the transmission distribution $\mu(t)$ were exponential, the integral would be $1/\alpha$ and the epidemic would grow as the solution of an ODE, $I_{\text{tot}}(t) = I(0)e^{\alpha t}$.

With the parameter set specific to COVID-19 (Table 1), we obtain $\alpha \approx 0.2$, which corresponds to a doubling time of 5 days. Note that this expected cumulative epidemic size accounts for epidemics that eventually die out. Since we are only interested in epidemic clusters that eventually result in an epidemic outbreak, we need to rescale the initial epidemic size by dividing by the survival probability p_{surv}:

$$I_{\text{surv}}(t) = \frac{I(t)}{p_{\text{surv}}} = \frac{I(0)}{p_{\text{surv}}}\,\frac{e^{\alpha t}}{\alpha\int_0^\infty R_0 e^{-\alpha s}s\mu(s)ds}\,. \tag{3}$$

This rescaling reflects conditioning the epidemic process on survival. The survival probability p_{surv} can be found by numerically computing the fixed point of the probability generating function of the offspring distribution (as detailed in Appendix A).



(a) $Y_i \sim$ Poisson

(b) $Y_i \sim$ negative binomial

Figure 2: **The total number of infected individuals over time.** The shaded region shows the 90% confidence interval obtained from 10,000 stochastic simulations that resulted in an epidemic outbreak. Dots represent the average of these simulations over time. The theoretical prediction (solid line calculated from Eq. (3)), even though formally only valid for asymptotically large times, agrees well with the simulated averages, even for relatively small times. 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.

With our parameter set, we find that from time $t = 10$ days on, the deterministic equation in Eq. (3) provides an accurate prediction of the mean of the individual-based model conditioned on survival (Fig. 2). Epidemic sizes vary much more with a negative binomial distribution of offspring number (Fig. 2b) than with a Poisson distribution (Fig. 2a). For example, at day 20, the number of infected individuals could be between 2 and 404 in the Poisson scenario and between 2 and 1373 with a negative binomial offspring distribution. This is due to the greater variance in offspring number under a negative binomial distribution than under a Poisson distribution (of the same mean). First, super-spreading events are more likely; 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.

3.1.1 The number of infectious individuals over time – a lower bound

Instead of considering the cumulative epidemic size, a more relevant measure may be the number of currently infectious individuals in the population. To estimate this number, we now only count individuals that will at some point in the future infect a susceptible individual. More specifically, we consider as infectious at time t individuals that are infecting other individuals at time $t' > t$. We thereby consider as non-infectious individuals without any secondary transmission, and individuals after their last transmission event (dashed lines in Fig. 1). Potentially infectious individuals who do not effectively transmit are not counted by this method. Thus, our estimate is a lower bound: we compute the minimal number of infectious individuals in the population over time.

To model the number of infectious individuals as defined above, we need a function defining the time at which infectious individuals are declared non-infectious. We denote by $L(a)$ the probability of being infectious at infection age a, i.e. the probability that the individual will infect a susceptible in the future. It is given by

$$L(a) = \begin{cases} \sum_{k=1}^{\infty} \mathbf{P}(Y=k)\mathbf{P}\left(\max_{j=1,\dots,k} t_{\text{inf}}^{(j)} \geq a\right), & \text{for } a > 0, \\ 1 - \mathbf{P}(Y=0), & \text{for } a = 0. \end{cases} \tag{4}$$

The summands in the first line are composed of the probability of having k offspring and the probability that the latest of these k transmission events occurs after infection age a. The second line implies that an individual who just got infected ($a = 0$) is considered as infectious in our framework only if it will ever infect others.

Again using general branching processes theory (Athreya and Ney, 1972; Haccou et al., 2005), we find that the number of infectious individuals in the population at time t is given by

$$I_{\text{inf}}(t) = (1-\ell)\, I_{\text{surv}}(t), \tag{5}$$

with $I_{\text{surv}}(t)$ given in Eq. (3) and where ℓ is defined as

$$\ell = \int_0^{\infty} (1 - L(a)) \alpha e^{-\alpha a} da. \tag{6}$$

Eq. (5) is simply Eq. (2) with an additional factor $(1-\ell)$ accounting for the removal of non-infectious individuals from the currently infected individuals. This factor is obtained from Eq. (6), which computes the fraction of individuals that are not infectious anymore (term $(1-L(a))$) over the stationary infection age distribution (term $\alpha e^{-\alpha a}$) of the population. Under the assumption that the basic reproduction number R_0 does not change over time, the growth rate α will remain constant as well, cf. Eq. (1). In this case, the infection age distribution of individuals will converge to a stationary distribution. The population is growing exponentially with rate α, the Malthusian parameter. The stationary age distribution reflects this exponential growth. For instance, the larger the population growth rate α, the faster the population grows, and consequently, the larger the proportion of recent infections. Note that the stationary infection age distribution, $\alpha e^{-\alpha a}$, only depends on the Malthusian parameter α, but not on the actual distribution of offspring number. Formally, the stationary infection age distribution of the population is derived from the renewal equation corresponding to our model. For more details, we once again refer to the book by Haccou et al. (2005, Section 3.4).

With our COVID-19-specific parameter set, a Poisson-distributed number of secondary infections gives $(1-\ell) \approx 0.71$, i.e. 71% of the total epidemic size is infectious at any point in time once the stationary distribution of infection ages is reached. The negative binomial offspring distribution gives $(1-\ell) \approx 0.48$, i.e., just 48% of the total epidemic size is infectious. For smaller values of R_0 these values decrease, e.g. for $R_0 = 1.3$ we have 19% and 15%, respectively. The distribution of secondary transmission events has a large impact on the fraction of infectious individuals in the population (given our definition of infectiousness). This is explained by the large number of individuals that do not transmit the disease under negative binomial transmission sampling. By our definition of infectiousness, these individuals that do not transmit are 'removed' from the infectious population immediately after their infection. Therefore, the lower bound of infectious individuals in the case of a negative binomial offspring distribution is much lower than the corresponding lower bound with Poisson sampling.

3.1.2 The number of RT-PCR-detectable cases

Another measure to evaluate the prevalence of the disease in the 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 8$^{\text{th}}$ of May across a large number of countries was around 30% (Belloir and Blanquart, 2020; Russell et al., 2020). However, test capacity is not the only limiting factor to detect infected individuals. The probability to test positive, e.g. with a reverse transcriptase polymerase chain reaction (RT-PCR) test, given that the tested person is infected with the virus causing COVID-19 (SARS-CoV-2), also depends on the infection age of an individual (Borremans et al., 2020; Kucirka et al., 2020). Using these probabilities for testing positive by RT-PCR, we derive an upper bound for the number of RT-PCR-detectable cases within a local outbreak. That is, if we were to test all infected individuals in the epidemic cluster at the same point in time, e.g. on day 20 after the first infection in the local outbreak, how many people would we expect to test positive by RT-PCR?

To answer this question, we repeat the steps outlined in the analysis of the number of currently infectious individuals, i.e., we first need to define the age-dependent probabilities to test positive, denoted $Q(a)$. We use the estimates obtained in a meta-analysis for RT-PCR tests, which we re-plot in Fig. 3a (upper panel in Fig. 2 in Kucirka et al. (2020)). Then, similar to Eq. (5), we find

$$I_{\text{detect}}(t) = (1-q)\, I_{\text{surv}}(t)\,, \tag{7}$$

with

$$q = \int_0^\infty (1-Q(a))\alpha e^{-\alpha a} da \tag{8}$$

With our parameter set, $q \approx 0.7$ (evaluating a discretized version of the integral), so that we would expect that only about 30% of the infected individuals would test positive at each point in time (Fig. 3b; here this proportion is independent of the distribution of secondary infections). Again, this estimate strongly depends on the stationary infection age distribution. With $R_0 = 1.3$, we find that about 65% of the epidemic cluster would test positive by RT-PCR.

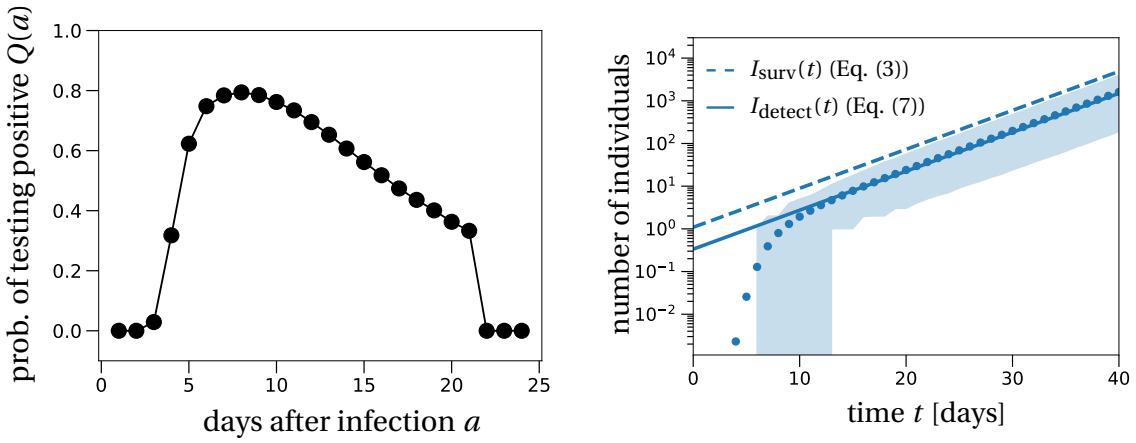
3.2 Time distribution of the first hospitalization event

We now study the timing of the first hospitalized case within the epidemic cluster. The estimate of the mean epidemic size over time (Eq. (3)) provides the deterministic time at which a certain infected population size is reached, and is therefore essential in deriving the first hospitalization time distribution. More specifically, the number of infection events till the infection of the first individual to be hospitalized happens, is a geometrically distributed number with probability p_{hosp}. That is, counting the infections until hospitalization resembles a discrete waiting process with 'success' probability p_{hosp}. We denote the number of infections prior to the infection of the eventually hospitalized individual by N_{inf} (N_{inf} includes the eventually hospitalized individual). Note that if we were interested in the j^{th} hospitalization event, the number of infected individuals until the j^{th} hospitalization event would be distributed according to a negative binomial distribution.

We now 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. (3)). We denote this time by $t^{\text{det}}_{N_{\text{inf}}}$. Lastly, we add the probability density of the time from infection to hospitalization, t_{hosp}, to this deterministically computed time. Denoting by T_{hosp} the time of first hospitalization, its density is given by:

$$\begin{aligned} h_{\text{hosp}}(t) &:= \lim_{dt\to 0} \mathbf{P}\left(T_{\text{hosp}} \in (t-dt, t+dt)\right) \\ &\approx \sum_{i=1}^{\infty} \mathbf{P}\left(N_{\text{inf}} = i\right)\; f_{\text{hosp}}\left(t - t_i^{\text{det}}\right) = \sum_{i=1}^{\infty} p_{\text{hosp}}\left(1-p_{\text{hosp}}\right)^{i-1} f_{\text{hosp}}\left(t - t_i^{\text{det}}\right), \end{aligned} \tag{9}$$

where $f_{\text{hosp}}(t)$ denotes the probability density of the time from infection to hospitalization t_{hosp} evaluated at time t (see Table 1). It is important to keep in mind that the density of the first hospitalization



(a) Probability of testing positive

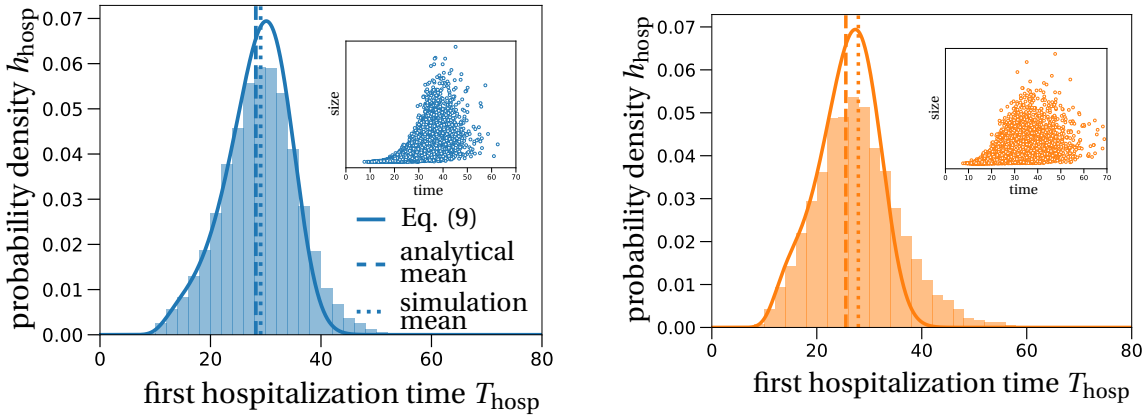
(b) Number of detectable cases

Figure 3: **The number of RT-PCR-detectable cases.** (a) The probability for an infected individual to test positive by an RT-PCR test is zero during the first two days after infection. It then increases up to approximately 0.8 on day 8 after infection before decreasing until three weeks after infection. The plot is based on data from Fig. 2, upper panel in Kucirka et al. (2020) (data points for days 22 to 24 are arbitrarily set to 0). (b) The expected number of detectable cases when testing the whole epidemic cluster on day t (solid line, Eq. (7)) is about 30% of the total size of the epidemic cluster (dashed line, Eq. (3)). The shaded region represents the 90% confidence interval of 10,000 individual-based simulations with a Poisson-distributed number of secondary infections, with averages depicted as dots.

time $h_{\text{hosp}}(t)$ is an approximation, because it is based on the mean epidemic size and not the whole distribution of epidemic sizes. The mean epidemic size directly provides the deterministic hitting time t_i^{det}, neglecting the whole distribution of the epidemic size.

With a Poisson distributed number of offspring, our analytical approximation is accurate for low hospitalization times. However the analysis underestimates the probability of large times to first hospitalization (those larger than ~ 28 days (Fig. 4a)). For negative binomial offspring numbers, our approximation underestimates the right tail of the hospitalization time distribution even more. This discrepancy can be explained as follows: Our approximation in Eq. (9) puts too much probability mass on hospitalization times close to the average hospitalization time, because we approximate the epidemic size using the deterministic hitting time. Our fit is poorer for negative binomially distributed offspring sizes (Fig. 4b), because the variance in epidemic sizes is larger than with Poisson-distributed offspring numbers. In particular, trajectories that remain at low prevalence for long times (inset in Fig. 4b) are responsible for the more pronounced tail when compared to the Poisson scenario in this parameter set. For lower values of R_0, our estimate is less good because of the slower convergence of the theoretical prediction of the average epidemic size, as shown in Section S2 in the Supplementary Information (SI).

While the simulated distributions of first hospitalization time differ with the two distributions, their means and standard deviations are similar. The mean (standard deviation) of the time of the first hospitalization, as obtained from 10,000 individual-based simulations, are ~ 29 days (~ 7 days) in the case of Poisson-distributed offspring numbers and ~ 28 days (~ 8 days) for negative binomial offspring numbers, respectively. The similarity of these values indicates that the distribution of offspring does not play a substantial role in the timing of the first hospitalized patient within an epidemic cluster. For



(a) $Y_i \sim$ Poisson

(b) $Y_i \sim$ negative binomial

Figure 4: **Distribution of the first hospitalization time.** The histograms are obtained from 10,000 stochastic simulations, and the solid lines are computed using Eq. (9). The insets are scatterplots where each point depicts a simulation which finished at a certain hospitalization time (x-axis) with a certain epidemic size (y-axis). The Poisson distributed offspring numbers in subfigure (a) result in less dispersed hospitalization times and are thus better approximated by our mean field type approach (compare dashed and dotted lines). Negative binomially distributed offspring numbers, subfigure (b), show a more pronounced distribution tail resulting from epidemics that remain relatively small for a long time (compare also the scatterplots in the insets of the panels).

the parameter set studied here, our analytical estimate captures the mean of the first hospitalization time reasonably well (dashed and dotted vertical lines in Fig. 4). Only for the negative binomial distribution does our analytical solution slightly underestimate the simulated average. The reason, as before, is the excess of trajectories that remain at low prevalence for long times, which are not well captured by our approximation.

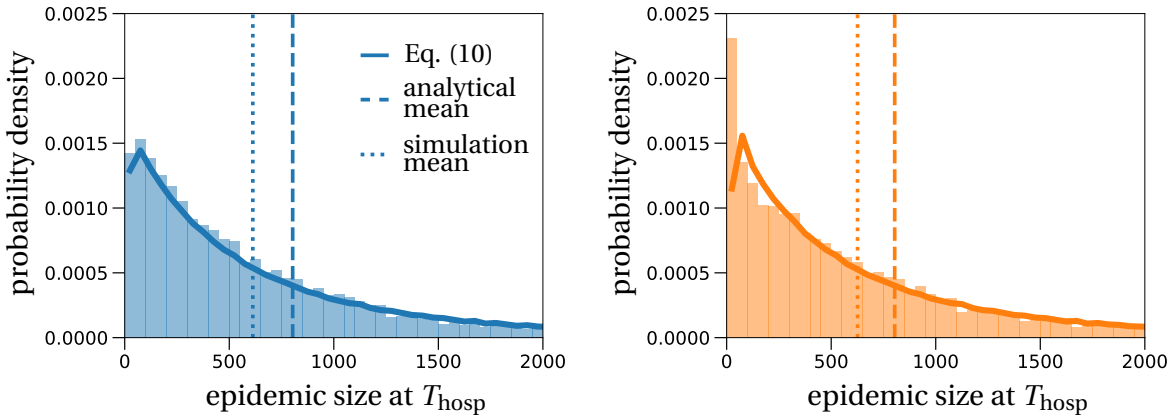
For lower values of R_0, we find that negative binomially distributed offspring sizes result in much lower average hospitalization times than Poisson distributed offspring numbers (Section S2, SI). The reason is that extinction is much more likely in the negative binomial offspring scenario (58% in the Poisson case, 83% in the negative binomial case). Therefore, trajectories that result in a hospitalization have grown faster to escape extinction than in the Poisson case and consequently, the first hospitalization will occur earlier (Fig. S2 in the SI).

3.3 Epidemic size at the first hospitalization event

In a next step, we use the distribution of the first hospitalization time to infer the size of the epidemic at that random time. Therefore, we combine Eqs. (3) and (9) and obtain the following probability mass function:

$$\mathbf{P}\left(I_{\text{surv}}(T_{\text{hosp}}) = k\right) = \int_0^\infty h_{\text{hosp}}(t)\mathbb{1}_{\{I_{\text{surv}}(t)\in[k-1/2,k+1/2)\}}\,dt\,. \tag{10}$$

This estimate of the epidemic size distribution accurately describes the simulated data (Fig. 5). The only difference occurs for low epidemic sizes in the case of a negative binomial offspring distribution. This excess in the simulated data is likely due to trajectories that remain at low prevalence for long times, which are due to the large number of non-transmitting infected individuals.



(a) $Y_i \sim$ Poisson

(b) $Y_i \sim$ negative binomial

Figure 5: **Epidemic size at the first hospitalization event.** The histograms are obtained from 10,000 stochastic simulations and represent the epidemic size at the first hospitalization event at time T_{hosp}. For the negative binomial offspring distribution, we observe an excess of low epidemic sizes compared to the theoretical prediction (solid lines, Eq. (10)). Otherwise, the two offspring distributions result in very similar shapes, which is also reflected by similar distribution statistics, e.g. for the 5- and 95-percentile (Table 2).

Similarly to our observations in the case of the first hospitalization time, there are no large differences between the empirical epidemic size distributions of the two offspring distributions. The mean epidemic sizes are ~ 610 for Poisson and ~ 625 for negative binomially distributed secondary infections. Even when comparing the 5 and 95 percentiles, the difference remains small, with the negative binomial distribution showing a bit more variance in the data (Table 2). In general, our analytical estimate overestimates the mean, which is most likely explained by the long tail of the theoretical distributions. For lower values of R_0, the tail of the theoretical distribution is much shorter and the empirical and theoretical mean of the epidemic size at hospitalization agree (Section S3, SI).

3.4 Comparison to the improved rule of thumb

Lastly, we compare our results to the improved rule of thumb, which goes as follows: at the first hospitalization event, go back a random number of days it takes from infection to hospitalization (drawn from the hospitalization time distribution f_{hosp}) and start the epidemic with a geometrically distributed number of infected individuals (on average $1/p_{\text{hosp}}$) at that time. Then the epidemic gets propagated by drawing a random number of secondary infections for each of these infected individuals and corresponding transmission times. This step is repeated for all infections before the first hospitalization event.

Our approach differs from this procedure by taking into account the entire epidemic history. Instead of initializing the dynamics with on average $1/p_{\text{hosp}}$ infectious individuals at a single time point, we accumulate the infectious population over time. This implementation results in smaller epidemic sizes at the first hospitalization event compared to the improved rule of thumb, as visualized in Fig. 6 and quantified in Table 2. The difference between the improved rule of thumb and the actual epidemic size becomes larger the longer one has to wait for the first hospitalization to occur. For example, if we compare the 5-percentiles of the epidemic size distributions with each other, i.e., hospitalizations that occurred early, there is essentially no difference between the two methods (35 infected individuals

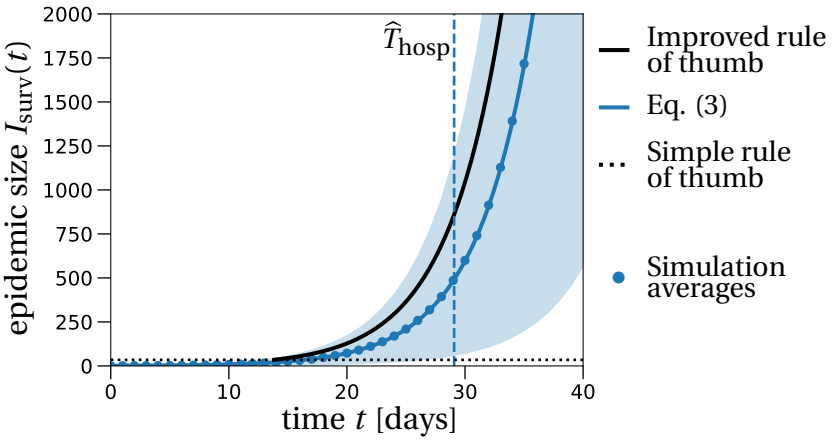


Figure 6: **Comparison of the epidemic size with the improved rule of thumb.** The simulation averages (dots) are obtained from 10,000 individual-based simulations, the shaded region is the 90% confidence interval of the simulated data and solid lines are the averages of the theoretical predictions of the respective models. The average of the improved rule of thumb is computed by initializing the epidemic with $1/p_{\mathrm{hosp}}$ infected individuals at time $\widehat{T}_{\mathrm{hosp}} - \widehat{t}_{\mathrm{hosp}}$, where $\widehat{T}_{\mathrm{hosp}}$ is the average time of the first hospitalization event (dashed line) and $\widehat{t}_{\mathrm{hosp}}$ the average of the time from infection to hospitalization. The epidemic is then propagated by Eq. (3) with the adjusted survival probability of the epidemic, $p_{\mathrm{surv}} = 1 - (p_{\mathrm{ext}})^{1/p_{\mathrm{hosp}}}$, where p_{ext} is the probability for an epidemic to die out if started with a single individual. The improved rule of thumb overestimates the actual epidemic size, the simple rule of thumb (dotted line) underestimates it.

in our model, 34 in the improved rule of thumb scenario; Table 2). However, comparing the mean and the 95-percentile between the two models, there are substantial differences. While the mean is overestimated by approximately 200 cases, the 95-percentile estimate differs by ~900 individuals. This inflation of difference is not surprising in view of the exponential growth of the number of cases; a small difference in the initial condition will result in large disagreement of the predictions two weeks after, corresponding to the average time of hospitalization.

For lower values of R_0, this overestimate is worse in relative terms, i.e., for $R_0 = 1.3$ the average value of the improved rule of thumb even exceeds the 95-percentile of the stochastic simulations (Section S4, SI). Yet, in absolute numbers, the improved rule of thumb still overstimates the epidemic size by around 200 infected individuals, thus not substantially changing the order of magnitude.

4 Discussion

We have developed analytical approximations to study the early dynamics of a local outbreak that is initiated by a single infected individual. Importantly, we derived the theoretical distribution of the first identification time, in our case the first hospitalization of an infected individual, and the distribution of the epidemic size at this time. We find that with our COVID-19-specific pre-lockdown parameter set, the choice of distribution of secondary transmission events, typically either Poisson or negative binomial, does not substantially affect these distributions – neither the analytical approximation nor the distribution derived from stochastic simulations. This is somewhat surprising in the case of the stochastic simulations, since the variance of the cumulative epidemic sizes, as shown in Fig. 2, varies

Model	Epidemic size (cumulative)			Number of infectious ind. – lower bound		
	Mean	5% quantile	95% quantile	Mean	5% quantile	95% quantile
Poisson	613	35	1844	290	16	878
Negative binomial	626	10	1974	164	1	516
Improved rule of thumb (Poisson)	807	34	2774	460*	19	1581
Improved rule of thumb (negative binomial)	803	19	2784	290*	5	1022

Table 2: **Statistics of the epidemic at the first hospitalization event.** The values are based on 10,000 individual-based simulations either with a Poisson or with a negative binomial offspring distribution. In the last two rows, we re-implemented the procedure described in Jombart et al. (2020) with a Poisson and a negative binomial offspring distribution. The epidemic size derived from the simple rule of thumb is 34 and thus largely underestimates the real epidemic size. *We have added the lower bounds of infectious individuals even though these numbers were not computed in Jombart et al. (2020).

considerably between these two offspring distributions. For smaller values of the basic reproduction number R_0, the estimates of the first hospitalization time differ between the two offspring distributions, the negative binomial distribution resulting in lower average hospitalization times than the Poisson distribution, as shown for $R_0 = 1.3$ in Section S2 in the SI. The reason is the much smaller establishment probability in case of a negative binomial offspring distribution, so that trajectories that result in a hospitalization grow faster early on to escape extinction. However the distribution of epidemic *size* is still the same for the two offspring distributions (Section S3, SI).

Previous estimates of the epidemic size at the first disease-caused death (Jombart et al., 2020) assume that all infections, from which one individual eventually dies, occur at the same time. This accumulation of infectious individuals at a single time point does not account for the epidemic history of the cluster and overestimates the actual epidemic sizes (Fig. 6 and Fig. S4 in the SI). While this effect is small if the identification time is small, it becomes considerable for late detection times. The later the identification of the first case, measured in the age of the epidemic cluster, the more does the improved rule of thumb overestimate the actual epidemic size. Additionally, with the previous methods, it is not possible to derive a distribution of the first event time, e.g. the first death or hospitalization event. In contrast, our analysis provides such an estimate.

The distribution of first event times can also be used to infer the time of the first infection within the cluster. However, this is only true in situations where the immigration of new cases is negligible so that the epidemic dynamics are indeed local and explained by a single cluster. Therefore, our analysis is also only valid during the early phase of an epidemic outbreak or if multiple introductions within a local community are unlikely. This seems to be in contrast to our theoretical results being derived for asymptotically large times. In the application though, our estimates converge very quickly towards the simulated average values for a large number of secondary transmissions, e.g. for $R_0 = 2.9$ as estimated during the early phase of the COVID-19 epidemic in France. For lower values of R_0, e.g. $R_0 = 1.3$ as studied in the SI, convergence of the theoretical epidemic size average (Eq. (3)) to the simulated average takes much longer and care should be taken when interpreting the theoretical results. We recommend to run stochastic simulations of epidemic sizes to verify the speed of convergence of the

average epidemic size as stated in Eq. (3).

In addition, we approximated the number of RT-PCR-detectable cases if one were to test the whole population on a single day. Even in this ideal scenario, because most people have been infected recently, only 30% of the cumulative number of cases will test positive. This is due to very recent infections not being detectable (Kucirka et al., 2020); a few days later, a stunning 97% of the cases that tested negative would now be RT-PCR-positive. Conversely, only a fraction of 3% of the negatively tested cases corresponds to old infections that are no longer detectable. It is thus important to test about a week after risk contact to identify most positive cases.

For lower values of R_0, the proportion of RT-PCR-positive cases is much more encouraging. For example, we would expect that with $R_0 = 1.3$, approximately 65% of the cumulative epidemic size would test positive because there will be comparatively fewer very recent undetectable infections. It is therefore advantageous, not only for the overall epidemic situation, but also for an efficient testing effort, to maintain a low number of secondary transmission events. Similar proportions of detectable cases are expected when testing for antibody markers, because the detection probability over time for these markers is comparable to that of RT-PCR-based tests (Borremans et al., 2020, Fig. 2).

It is important to keep in mind that our estimate is based not only on the probability to test positive for RT-PCR tests (Kucirka et al., 2020), but also on a constant basic reproduction number R_0. As we just mentioned, R_0 has a large effect on the detection rate because it affects the growth rate of the epidemic size, i.e. the Malthusian growth rate α (Eq. 1). This holds because the growth rate itself is linked to the stationary infection age distribution of the population, which is an exponential distribution with parameter α. If R_0 varies too fast, this impedes convergence of the infection age distribution towards the stationary infection age distribution and our analysis cannot be carried out anymore. The infection age distribution is therefore not stationary in reality, yet during the pre-lockdown phase of the epidemic in France the epidemic dynamics were well described by the theoretically computed stationary age distribution (Foutel-Rodier et al., 2020). The higher the reproduction number R_0 is, the faster is the convergence to the stationary age distribution and the better is this approximation justified.

A further restriction is that during the early phase of a newly emerging epidemic, estimates for hospitalization rates, transmission or hospitalization time distributions may not be available yet. Since evaluation by analytical formulas, or even running the stochastic simulations a large number of times, is very fast, it is possible and therefore recommended to evaluate a number of different scenarios. Moreover, we assume that the population-wide hospitalization probability is about the same in our setting of interest as where it was estimated. This assumption can break, for example, if the hospitalization probability varies across age classes and the age structure of infected individuals is different in our setting.

Apart from these limitations, our derived equations are valid for any emerging disease that is described by an exponential growth curve. This is mostly the case for airborne diseases like SARS or COVID-19. For infectious diseases like Ebola or HIV, the initial growth can be described as sub-exponential (Chowell et al., 2016; Viboud et al., 2016). While there is not yet a mechanistic description of this growth behavior, except for a connection to renewal equations pointed out in House (2016), a phenomenological description (Chowell et al., 2016; Viboud et al., 2016) should be sufficient to translate our analytical results into this context, at least if a stationary infection age distribution is accessible.

In conclusion, our analytical distributions for the first hospitalization time and the epidemic size distribution at that time can be useful to adapt the intensity and scale of interventions to contain a local outbreak. Additionally, the detection of the first case contains information about the introduction time of the disease into the local community, at least under the assumption that multiple introductions are unlikely. These results improve previous estimates on the epidemic size at the first detection. Previous

methods neglect the epidemic history of the epidemic cluster, which results in an overestimation of the actual number of infected individuals in the case of the improved rule of thumb, or an underestimation in the case of the simple rule of thumb. Our analysis shows that accounting for the whole epidemic history is important and necessary to well describe the entire distribution of the epidemic size.

Data availability

The C++ codes, data files and Python scripts used to generate the figures are available at `https://github.com/pczuppon/early_epidemic_inference`.

Acknowledgements

We thank Luca Ferretti for pointing us to the parameter estimates provided in the github repository of the agent-based model describing COVID-19 dynamics. 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.

A Computing the survival probability

We briefly outline how to compute the survival probability of a branching process. In general, for a branching process, the extinction probability is given by the smallest positive fixed point of the probability generating function of the offspring distribution; for more details we refer to the book by Haccou et al. (2005). The survival probability is simply the complement of the extinction probability, i.e., one minus the value of the root.

The probability generating function for the Poisson distribution with rate λ reads

$$G_P(z) = \mathbf{E}[z^X] = e^{\lambda(z-1)}, \tag{11}$$

where X is a Poisson distributed random variable. The probability generating function of a negative binomial distribution with success probability p and number of successes κ is given by

$$G_{NB}(z) = \left(\frac{p}{1-(1-p)z}\right)^{\kappa}. \tag{12}$$

The number of successes κ is also known as the dispersion parameter (and can formally be any real positive number) and the success probability p is given by $p = \kappa/(\kappa + R_0)$, where R_0 is the basic reproduction number.

Unfortunately, analytical solutions for the smallest positive fixed point of these two generating functions are not accessible. For the computation of the theoretical predictions in the figures, we determined the fixed points numerically.

References

Althouse, B. M., Wenger, E. A., Miller, J. C., Scarpino, S. V., Allard, A., Hébert-Dufresne, L., and Hu, H. Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLOS Biology*, 18(11):e3000897, 2020. doi: 10.1371/journal.pbio.3000897.

Athreya, K. B. and Ney, P. E. *Branching Processes*. Springer Berlin Heidelberg, 1972. doi: 10.1007/978-3-642-65371-1.

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. *medRXiv preprint*, 2020. doi: 10.1101/2020.09.10.20192120.

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.

Cheng, H.-Y., Jian, S.-W., Liu, D.-P., Ng, T.-C., Huang, W.-T., and and, H.-H. L. Contact tracing assessment of COVID-19 transmission dynamics in taiwan and risk at different exposure periods before and after symptom onset. *JAMA Internal Medicine*, 2020. doi: 10.1001/jamainternmed.2020.2020.

Chowell, G., Viboud, C., Simonsen, L., and Moghadas, S. M. Characterizing the reproduction number of epidemics with early subexponential growth dynamics. *Journal of The Royal Society Interface*, 13 (123):20160659, 2016. doi: 10.1098/rsif.2016.0659.

Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P.-Y., and Colizza, V. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC Medicine*, 18(1), 2020. doi: 10.1186/s12916-020-01698-4.

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.

Faes, C., Abrams, S., Beckhoven, D. V., Meyfroidt, G., Vlieghe, E., and Hens, N. Time between symptom onset, hospitalisation and recovery or death: a statistical analysis of different time-delay distributions in Belgian COVID-19 patients. *medRXiv preprint*, 2020. doi: 10.1101/2020.07.18.20156307.

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., and Fraser, C. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), 2020. doi: 10.1126/science.abb6936.

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.

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

He, X., Lau, E. H. Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B. J., Li, F., and Leung, G. M. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26(5):672–675, 2020. doi: 10.1038/s41591-020-0869-5.

Hinch, R., Probert, W. J. M., Nurtay, A., Kendall, M., Wymatt, C., Hall, M., Lythgoe, K., Cruz, A. B., Zhao, L., Stewart, A., Ferretti, L., Montero, D., Warren, J., Mather, N., Abueg, M., Wu, N., Finkelstein, A., Bonsall, D. G., Abeler-Dorner, L., and Fraser, C. OpenABM-covid19 - an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *medRXiv preprint*, 2020. doi: 10.1101/2020.09.16.20195925.

House, T. A general theory of early growth? *Physics of Life Reviews*, 18:109–111, 2016. doi: 10.1016/j.plrev.2016.08.006.

Jombart, T., van Zandvoort, K., Russell, T. W., Jarvis, C. I., Gimma, A., Abbott, S., Clifford, S., Funk, S., Gibbs, H., Liu, Y., Pearson, C. A. B., Bosse, N. I., Eggo, R. M., Kucharski, A. J., and Edmunds, W. J. Inferring the number of COVID-19 cases from recently reported deaths. *Wellcome Open Research*, 5: 78, 2020. doi: 10.12688/wellcomeopenres.15786.1.

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.

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.

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.

Roques, L., Klein, E. K., Papaïx, J., Sar, A., and Soubeyrand, S. Using early data to estimate the actual infection fatality ratio from COVID-19 in France. *Biology*, 9(5):97, 2020. doi: 10.3390/biology9050097.

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., Le Strat, Y., Lessler, J., Levy-Bruhl, D., Fontanet, A., Opatowski, L., Boëlle, P.-Y., and Cauchemez, S. Estimating the burden of SARS-CoV-2 in France. *Science*, 2020. doi: 10.1126/science.abc3517.

Sofonea, M. T., Reyné, B., Elie, B., Djidjou-Demasse, R., Selinger, C., Michalakis, Y., and Alizon, S. Epidemiological monitoring and control perspectives: application of a parsimonious modelling framework to the COVID-19 dynamics in France. *medRXiv preprint*, 2020. doi: 10.1101/2020.05.22.20110593.

Viboud, C., Simonsen, L., and Chowell, G. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics*, 15:27–37, 2016. doi: 10.1016/j.epidem.2016.01.002.

Supplementary Information: "A stochastic epidemiological model to estimate the size of an outbreak at the first case identification"

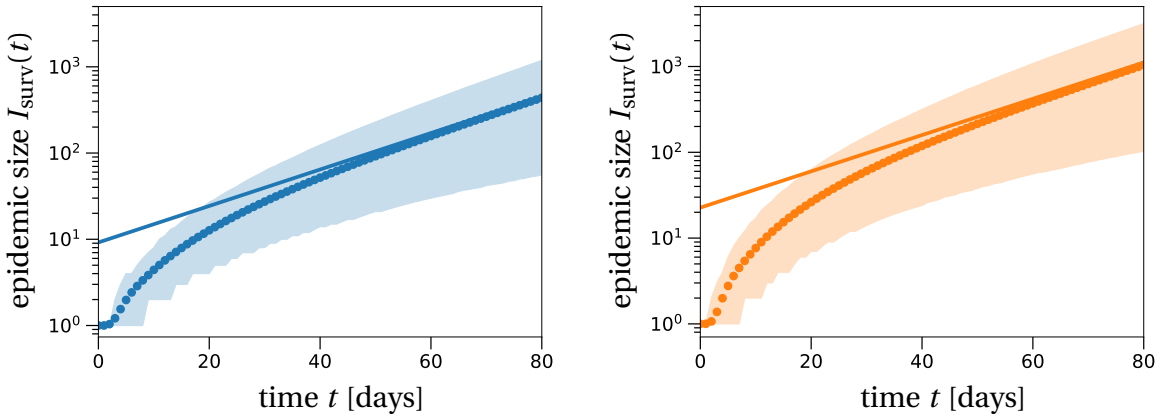
Peter Czuppon, François Blanquart, Florence Débarre

Table of Contents

A Epidemic size with $R_0 = 1.3$

We conduct the same analysis as in the main text, just with a reduced value of the basic reproduction number R_0. In the main text, we have analyzed the situation $R_0 = 2.9$ which corresponds to the pre-lockdown epidemic behavior in France (e.g. Salje et al., 2020). We now assume $R_0 = 1.3$, a value that is sligthly above the critical value $R_0 = 1$, to evaluate how well our proposed methodology compares to simulations for slower growing epidemics.

First, we plot the cumulative epidemic sizes over time (Fig. S1). In comparison to the main text (Fig. 2), the epidemic grows slower for the reduced value of R_0. The agreement between simulations and theory (Eq.(3)) starts later with $R_0 = 1.3$ than it does with $R_0 = 2.9$ (where Eq. (3) matched the simulations from around day 10 on). In general, the larger the reproductive number, R_0, the faster is the convergence of the simulations to the theoretical prediction. This follows from the general convergence theorem about the asymptotic growth of a supercritical branching process (Haccou et al., 2005, Section 3.3).



(a) $Y_i \sim$ Poisson

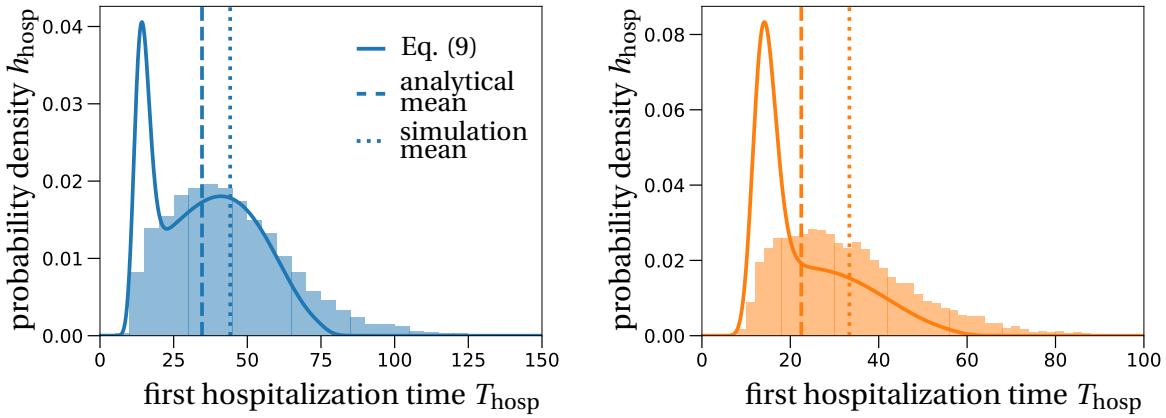
(b) $Y_i \sim$ negative binomial

Figure S1: **The total number of infected individuals over time with $R_0 = 1.3$.** The shaded region shows the 90% confidence interval obtained from 10,000 stochastic simulations that resulted in an epidemic outbreak. Dots represent the average of these simulations over time. The theoretical prediction (solid line calculated from Eq. (3) in the main text), even though formally only valid for asymptotically large times, agrees well with the simulated averages from around day 50 on.

B Hospitalization time in the case $R_0 = 1.3$

For $R_0 = 1.3$, the first hospitalization time has a broader distribution (Fig. S2) than with the larger value $R_0 = 2.9$ in the main text (Fig. 4). The simulated average values are larger for smaller values of R_0: for $R_0 = 1.3$ we have 44 days in the Poisson case and 33 days in the negative binomial offspring scenario. The corresponding values for $R_0 = 2.9$ are 29 and 28 days, respectively.

For $R_0 = 1.3$ and with both offspring distributions, our theoretical prediction (Eq. (9) in the main text) strongly overestimates early hospitalization times and underestimates larger hospitalization times. This is most likely explained by the relatively bad approximation of the epidemic size for very small times since the first infection of the epidemic cluster. As our theoretical approximation over-estimates epidemic size at early times (Fig. S1), it also over-estimates the probability of an hospitalization at an early time since the beginning of the local epidemic. As a result, our theoretical approximation puts too much weight on early hospitalization times and therefore under-estimates the hospitalization times obtained in simulations. For future research, it would be interesting to find more accurate approximations of the epidemic size for small times t in the context of general branching process.



(a) $Y_i \sim$ Poisson

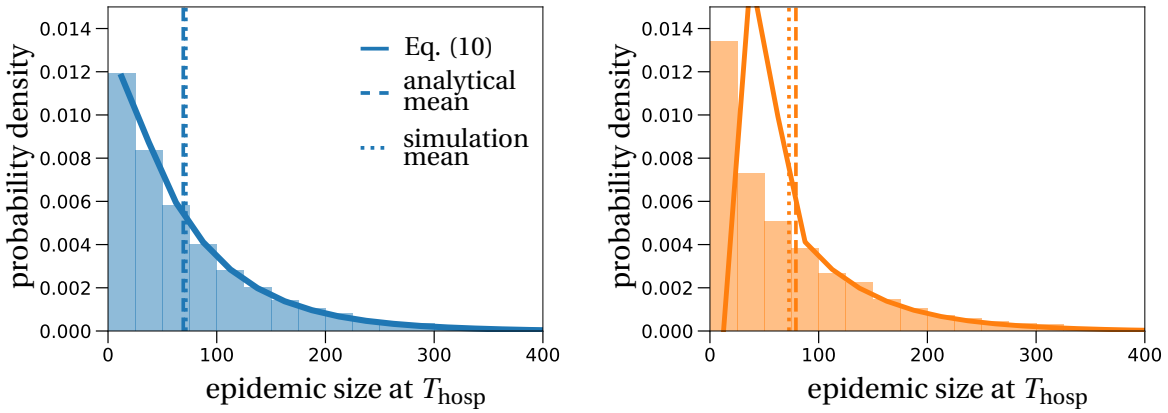
(b) $Y_i \sim$ negative binomial

Figure S2: **Distribution of the first hospitalization time with $R_0 = 1.3$.** The histograms are obtained from 10,000 stochastic simulations and the solid lines are computed by Eq. (9) in the main text. Here, both the Poisson (subfigure (a)) and negative binomial (subfigure (b)) distributed offspring numbers show a dispersed hospitalization time distribution, i.e., both have a heavy tail. In contrast to the main text, negative binomial distributed offspring sizes, subfigure (b), show a less pronounced distribution tail compared to the Poisson distribution. The theoretical predictions overestimate the early hospitalization times because of the bad approximation of epidemic sizes during the very early phase of the epidemic, cf. Fig. S1. Note the differing scalings of the axes between the two subfigures.

C Epidemic size at the first hospitalization for $R_0 = 1.3$

The epidemic size at the first hospitalization event is much smaller for $R_0 = 1.3$ than for the scenario considered in the main text ($R_0 = 2.9$). The average epidemic size at the first hospitalization as obtained from the simulations is 72 infected individuals in the Poisson case and 73 in the case of the negative binomial distribution for $R_0 = 1.3$. For $R_0 = 2.9$ the corresponding values are 610 and 625, respectively. The smaller sizes with a smaller R_0 are because the epidemic size grows much more slowly.

With the slower growing epidemic, the theoretical estimates agree reasonably well with the simulated distribution of epidemic sizes (Fig. S3). The only exception are very small values in case of the negative binomial distribution. The good estimation is also reflected by the overlap of theoretical and simulated average values (dashed and dotted vertical lines in Fig. S3).



(a) $Y_i \sim$ Poisson

(b) $Y_i \sim$ negative binomial

Figure S3: **Epidemic size at the first hospitalization event for $R_0 = 1.3$.** The histograms are obtained from 10,000 stochastic simulations and represent the epidemic size at the first hospitalization event at time T_{hosp}. For the negative binomial offspring distribution, the theoretical prediction (solid lines, Eq. (10) in the main text) overestimates the simulated data for low epidemic sizes. Otherwise, the two offspring distributions result in very similar shapes and are well approximated. This is also reflected by the close agreement of the analytical and simulated mean of the epidemic size at the first hospitalization event.

D Comparison to the improved rule of thumb in the case of $R_0 = 1.3$

In the main text, the improved rule of thumb (Jombart et al., 2020) overestimated the actual epidemic size, but was still within the 90%-confidence interval of the simulations. Here, with $R_0 = 1.3$, the improved rule of thumb exceeds even the 95-percentile of the simulation results. In absolute numbers, the improved rule of thumb exceeds the true epidemic size by approximately 200 infected individuals at the average time of hospitalization ($t \approx 44$): 267 infections on average with the improved rule of thumb vs. 65 infections in the simulations.

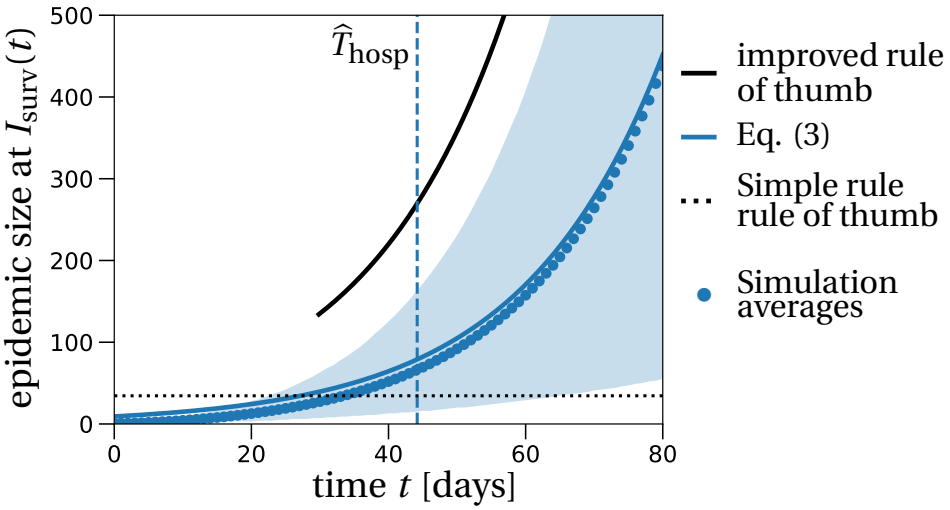


Figure S4: **Comparison of the epidemic size with the improved rule of thumb.** The simulation averages (dots) are obtained from 10,000 individual-based simulations, the shaded region is the 90% confidence interval of the simulated data and solid lines are the averages of the theoretical predictions of the respective models. The average of the improved rule of thumb is computed by initializing the epidemic with $1/p_{hosp}$ infected individuals at time $\widehat{T}_{hosp} - \widehat{t}_{hosp}$, where $\widehat{T}_{hosp}$ is the simulated average time of the first hospitalization event (dashed line) and $\widehat{t}_{hosp}$ the average of the time from infection to hospitalization. The epidemic is then propagated by Eq. (3) with the adjusted survival probability of the epidemic, $p_{surv} = 1 - (p_{ext})^{1/p_{hosp}}$, where p_{ext} is the probability for an epidemic to die out if started with a single individual. The improved rule of thumb overestimates the actual epidemic size, the simple rule of thumb (dotted line) underestimates it.