



**HAL**  
open science

## Estimating dates of origin and end of COVID-19 epidemics

Thomas Bénéteau, Elie ‘ Baptiste, Mircea T Sofonea, Samuel Alizon

► **To cite this version:**

Thomas Bénéteau, Elie ‘ Baptiste, Mircea T Sofonea, Samuel Alizon. Estimating dates of origin and end of COVID-19 epidemics. 2021. hal-03120811v1

**HAL Id: hal-03120811**

**<https://hal.science/hal-03120811v1>**

Preprint submitted on 25 Jan 2021 (v1), last revised 26 May 2022 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 Estimating dates of origin and end of COVID-19 epidemics

2 Thomas Bénéteau<sup>+,\*</sup>, Baptiste Elie<sup>+,\*</sup>, Mircea T. Sofonea, Samuel Alizon

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

4 <sup>+</sup> equal contribution

5 <sup>\*</sup> authors for correspondance: thomas.beneteau@ird.fr baptiste.elie@ird.fr

## 6 **Abstract**

7 Estimating the date at which an epidemic started in a country and the date at which it can end  
8 depending on interventions intensity are important to guide public health responses. Both are po-  
9 tentially shaped by similar factors including stochasticity (due to small population sizes), super-  
10 spreading events, and memory effects. Focusing on COVID-19 epidemics, we develop and analyse  
11 mathematical models to explore how these three factors may affect early and final epidemic dynam-  
12 ics. Regarding the date of origin, we find limited effects on the mean estimates, but strong effects on  
13 their variances. Regarding the date of extinction following lock-down onset, mean values decrease  
14 with stochasticity or with the presence of superspreading events. These results underline the impor-  
15 tance of accounting for heterogeneity in infection history and transmission patterns to make accurate  
16 predictions regarding epidemic temporal estimates.

# 17 1 Introduction

18 The ability to make robust epidemiological inferences or predictions strongly relies on the law of large  
19 numbers, which buffers the variability associated with individual processes. Most models of infectious  
20 diseases spread are deterministic and therefore assume that the number of infected hosts is large and  
21 above what has been termed the ‘outbreak threshold’ [10]. This assumption is violated at the beginning  
22 and end of an epidemic, where stochasticity may have a strong effect [4].

23 In this study, we tackle two issues. First, we wish to estimate the date of origin of an epidemic in  
24 a country, focusing on the case of COVID-19 outside China. This question is important because the  
25 infection being imported, some cases may be detected before the reported beginning of an epidemic  
26 wave, which is somehow counter-intuitive to an audience not familiar with stochasticity. Conversely,  
27 cryptic transmission can take place before an epidemic wave is detected, as observed thanks to SARS-  
28 CoV-2 genomic data in Washington state (USA) in Feb 2020 [3]. Second, we investigate how many days  
29 strict control measures need to last to ensure that the prevalence falls below key thresholds. Despite its  
30 public health implications, this latter question has rarely been investigated. There are some exceptions,  
31 for instance in the context of poliomyelitis [8], Ebola virus disease [23], and MERS [18] epidemics,  
32 but these neglect superspreading events and/or do not include non-Markovian effects (i.e. memory  
33 effects). Recently, however, it has been shown that incorporating secondary cases heterogeneity can  
34 significantly lower the delay until an Ebola virus disease outbreak can be considered to be over [7].

35 The COVID-19 pandemic has led to an unprecedented publication rate of mathematical models,  
36 several of which involve stochasticity. For instance, Hellewell *et al.* [12] analysed the initial steps  
37 of the outbreak to estimate the fraction of the transmission chains that had to be tracked to control  
38 the epidemics. Their results depend on the value of the basic reproduction number (denoted  $\mathcal{R}_0$ ),  
39 which corresponds to the mean number of secondary infections caused by an infected individual in  
40 an otherwise fully susceptible population [2], but also on individual heterogeneity. Indeed, if few  
41 individuals tend to cause a large number of secondary infections while the majority tends to cause none,  
42 the probability of outbreak emergence is much lower than if all individuals cause the same number of  
43 secondary infections [15]. Accounting for this property, a study used the early COVID-19 outbreaks  
44 incidence data in different countries to estimate the dispersion of the distribution of individual  $\mathcal{R}_0$  [9].  
45 Finally, Althouse *et al.* [1] have also used stochastic modelling to explore the role of super-spreading  
46 events in the pandemic and its consequences on control measures.

47 Here, we develop an original discrete stochastic (DS) model, which features some of the known  
48 characteristics of the COVID-19 epidemics. In particular, following earlier studies [12], we account for

49 the fact that not all hosts transmit on the same day post-infection. This is captured by assuming a dis-  
50 tribution for the serial interval, which is the time between the onset of the symptoms in the ‘infector’  
51 and that in the infected person [16, 11]. We also allow for heterogeneity in transmission patterns by  
52 assuming negative binomial distribution of the secondary cases. Furthermore, we reanalysed an earlier  
53 deterministic non-Markovian model [21] by setting the date of origin of the epidemic as the main free  
54 parameter. Finally, we analyse a classical deterministic Markovian model, which is commonly used to  
55 analyse COVID-19 epidemics [? ]. By comparing these models, we explore the importance of stochas-  
56 ticity, individual heterogeneity, and non-Markovian effects on the estimates of the dates of origin and  
57 end of a nation-wide COVID-19 epidemic, using France as a test case and mortality data because of its  
58 extensive sampling compared to case incidence data.

## 59 2 Methods

### 60 2.1 The Discrete Stochastic (DS) model

We assume that each infected individual causes on average  $\mathcal{R}_0$  secondary cases and that the host pop-  
ulation is homogeneously mixed (i.e. no spatial structure), an assumption that is relevant if a small  
fraction of the population is infected [24]. We model the number of new infected individuals per day  
(i.e. the daily incidence) as an iterative sequence following a Poisson distribution. Let  $(Y_t)_{t \in \mathbb{N}}$  be the  
random variable describing the incidence over time,  $t$  being the number of days since initialisation of  
the process. For all  $t \in \mathbb{N}$ , the sequence of  $(Y_{t+1})_{t \in \mathbb{N}}$  is such that

$$Y_{t+1} \sim \text{Poisson} \left( \eta_t \sum_{i=0}^t \omega_{t-i} \sum_{k=1}^{Y_i} F_{k,i} \right) \quad (1)$$

61 where  $\omega_{t-i}$  is the probability of infecting someone at time  $t$  ( $i$  days after being infectious),  $\eta_t$  is the av-  
62 erage contact rate in the population at day  $t$ , and  $F_{k,i}$  is the force of infection of individual  $k$ , infected  
63 at time  $i$ . The model is non-Markovian, which means that individual histories matter for the dynam-  
64 ics. More specifically, the probability that an event occurs (e.g. infecting another host) depends on the  
65 number of days spent in a state (e.g. being infected). Here, these non-Markovian aspects are captured  
66 through  $\omega$ , which is itself based on the generation time of the infection [16].

67 We consider two scenarios (a) without and (b) with individual heterogeneity. If we denote by  $\mathcal{F}$  the  
68 distribution of random variables  $(F_{x,y})_{(x,y) \in \mathbb{N}^2}$ , where  $F_{x,y}$  is the force of infection of an individual  $x$ ,  
69 infected at day  $y$ , then, in each scenario we assume that:

- a)  $\mathcal{F}$  is a Dirac distribution, noted  $\delta(\mathcal{R}_0)$ , implying that there is no heterogeneity and individuals

have the same infectivity and infection duration distribution. The sequence  $(Y_n)_{n \in \mathbb{N}}$  then simplifies into:

$$Y_{t+1} \sim \text{Poisson} \left( \mathcal{R}_0 \eta_t \sum_{i=0}^t \omega_{t-i} Y_i \right) \quad (2)$$

70 b)  $\mathcal{F}$  is a Gamma distribution with shape parameter  $k = 0.16$  and mean  $\mathcal{R}_0$ , implying that individuals are heterogeneous in infectivity and/or infection duration, which can lead to ‘superspreading’  
 71 events. We use the shape parameter ( $k$ ) value estimated for a SARS outbreak in 2003 [15], which  
 72 is consistent with early estimates for SARS-CoV-2 epidemics [9, 1, 14, 22].  
 73

74 To model the intensity of the control over the epidemic at time  $t$  such as, for instance, a national  
 75 lock-down, we vary the contact rate parameter  $\eta_t$ . We assume that  $\eta_t$  is piecewise constant and that its  
 76 discontinuities capture changes in public health policy (see Figure S6).

Overall, we define the temporal reproduction number ( $\mathcal{R}_t$ ) at time  $t$  such that

$$\mathcal{R}_t = \eta_t \mathbb{E}[\mathcal{F}] = \eta_t \mathcal{R}_0 \quad (3)$$

## 77 2.2 Beginning of the epidemic wave

78 To infer the starting date of the epidemic wave, we run our discrete stochastic (DS) algorithm starting  
 79 from one infected individual until the infection dynamic becomes deterministic, *i.e.* the law of large  
 80 numbers applies. We set the incidence threshold to 100 daily deaths, which was reached on March 23  
 81 in France; a value much higher than the outbreak threshold above which a stochastic fade out is unlikely  
 82 [10]. We use independent estimates for the other parameters and perform a sensitivity analysis, shown  
 83 in the Appendix.

To simulate death events in the DS model, we use the infection fatality ratio  $p$  and the delay from infection to death  $\theta$  previously estimated on French data of ICU and deaths [21] (Table S1). These estimates compare very well with other independent estimates made from contact tracing data [13]. More specifically, if we write  $X_t$  the number of individuals infected at time  $t$  who will die:

$$X_t \sim \text{Binomial}(Y_t, p) \quad (4)$$

84 We then chose the day of death for each individual of  $X_t$  by drawing a time from infection to death  
 85 following  $\theta$ .

86 We repeat the algorithm 10,000 times in order to obtain a stable distribution of starting dates and

87 discard epidemics that die out before reaching the threshold incidence.

88 To allow for comparison with empirical data, we first smooth out week-end under-reporting by  
89 computing a sliding average of this time series over a 7-days window.

90 Finally, we assume that the consequences of the lock-down, which was initiated in France on March  
91 17, did not affect the death incidence time series until the very end of March because of the delay  
92 between infection and death, which we estimate in France to be more than 11 days for 95% of the cases  
93 [21].

### 94 2.3 End of the epidemic wave

95 A national lock-down was established in France between Mar 17 and May 11, which drastically de-  
96 creased the spread of the epidemic with an estimated efficacy of  $1 - \eta_{FR} = 76\%$  [21]. On May 11,  
97 however, the virus was still circulating in France. Here, we estimate how many additional days of lock-  
98 down would have been necessary to reach epidemic extinction for various lock-down intensity post  
99 May 11. In the following we note by  $(\zeta)_{t>55}$ , the variation in the intensity of the lock-down after the 55  
100 days of the official lock-down (i.e. after May 11), defined as

$$\zeta_t = \frac{\eta_t - \eta_{FR}}{1 - \eta_{FR}} \quad (5)$$

101 where  $\eta_{FR} = 0.24$  represents the estimated contact rate of the population during the first lock-down.

102 To avoid the unnecessary accumulation of uncertainties, we initialise the model with incidence  
103 values obtained from a discrete-time non-Markovian model [21] on the period ranging from April 26  
104 to May 11. This interval is chosen because most of the infections after May 11 originate from infections  
105 that started less than 15 days ago (mathematically,  $\mathbb{P}[w_i \leq 15] \leq 0.999$  using the model calibration for  
106 the serial interval  $(w_i)_{i \in \mathbb{N}}$  in Table S1).

107 We then use a Monte-Carlo procedure to estimate key features of the sequence  $(Y_t)_t$ , such as the  
108 mean extinction time or the asymptotic extinction probability. This is done by running 10,000 inde-  
109 pendent and identically distributed simulations of our process for each set of parameters. We stock  
110 each of these 10,000 trajectories and then analyse these trajectories as follow. The scripts used for the  
111 simulations can be found in the supplementary materials.

112 First, we estimate the distribution of  $\tau$ , which is the minimal lock-down duration such that the  
113 incidence is always null afterwards for a given contact rate reduction post May 11. Mathematically,

$$\tau = \inf_{s \in \mathbb{N}} \{Y_k = 0; \forall k \geq s\} \quad (6)$$

114 The approximation of this distribution is obtained by assuming an infinitely long lock-down extension  
 115 under fixed contact reduction restrictions ( $(\zeta_t)_{t>55} = \alpha$ , with  $0 \leq \alpha \leq 1$ ).

116 Second, we study the effect of finite lock-down extensions on the probability of extinction and focus  
 117 on the risk of epidemic rebound upon lock-down lifting. For simplicity, we assume no control (i.e.  $\zeta_t =$   
 118 1) once the lock-down is over. The probability of having no new cases at time  $t$  ( $p_0(t)$ ) is estimated using  
 119 the following formula

$$p_0(t) = \frac{1}{N} \sum_{k=1}^N \mathbb{1}_{\{Y_t^k=0\}} \quad (7)$$

120 where  $N$  is the number of simulations and  $Y_t^k$  the number of newly infected individuals in the  $k$ -th  
 121 simulation at time  $t$ .

122 Third, we study the effect of initiating the lock-down one month or two weeks earlier in the epi-  
 123 demic (in France, February 17 or March 03 respectively) on the distribution of  $\tau$ . For comparison pur-  
 124 poses, we assume that the spread of the dynamic is equal to  $\eta_{FR} = 0.24$  for the first 55 days and then  
 125 extend the lock-down indefinitely with variable intensities to estimate  $\tau$  as described previously (see  
 126 equation 6).

## 127 2.4 Alternative models

128 To further study the effects of stochasticity, non-Markovian dynamics, and superspreading, we imple-  
 129 mented two additional models. The first is Markovian, i.e. memoryless, and is based on a simpler  
 130 model derived from a classical SEIR model. The second has a discrete-time structure, which allows to  
 131 capture non-Markovian dynamics [21].

### 132 The SEAIRHD model

133 In this classical compartment model, hosts can belong to seven states: susceptible to infection ( $S$ ), ex-  
 134 posed (i.e. infected but not infectious,  $E$ ), asymptomatic and infectious ( $A$ ), infectious and symptomatic  
 135 ( $I$ ), removed (i.e. recovered or isolated,  $R$ ), hospitalised who will die ( $H$ ), or dead ( $D$ ). The model is  
 136 described by a set of ODE detailed in the appendix. In the simulations, we assume that one exposed

137 individual starts the epidemic on day  $t_0$ .

138 This model is solved numerically using the Numpy package on Python 3.8.3 to obtain a deterministic  
139 trajectory with the parameters fitted to the empirical data, with a moving average of 7 days. We also  
140 simulate a stochastic version of this model 1,000 times using a Gillespie algorithm with the package  
141 TiPS [6] in R v.3.6.3 [20]. Scripts for the SEAIRHD model can be found in the supplementary materials.

## 142 **A non-Markovian deterministic model**

143 We estimate dates of origin and end of epidemics using an existing discrete-time model that has a  
144 similar structure to the continuous model mentioned above with an additional age-structure [21]. The  
145 serial interval is the same as in our model [17], and so is the use of non-exponential delays from infection  
146 to death. However, two major differences are that this earlier model is not stochastic and does not allow  
147 for superspreading events. We restricted the parameter inference to the daily death data described  
148 previously, with the main free parameter being the date of origin. We invite the reader to refer to [21]  
149 for the scripts and further details on this approach.

## 150 **2.5 Model calibration**

151 To allow for model comparison and improve estimates, we fixed some key parameters based on existing  
152 values, focusing on the French COVID-19 epidemic. Table S1 lists all the parameters used along with  
153 key references.

154 The likelihood of the deterministic SEAIRHD model was computed assuming a Poisson distribution  
155 of the daily mortality incidence data. Parameter inference with maximum likelihood was performed  
156 using the Powell algorithm implemented by Scipy.minimize function in Python.

# 157 **3 Results**

## 158 **3.1 Origin of the epidemic wave**

159 When neglecting host heterogeneity, using our DS algorithm, the median delay between the importa-  
160 tion of the first case of the epidemic wave and the time mortality incidence reaches 100 deaths per day  
161 is 67 days (equivalent to a first case on January 16 in France), with a 95% confidence interval (95% CI)  
162 between 62 and 79 days, *i.e.* between January 4 and 21 in France (Fig. 1). With this model, only 7% of  
163 the outbreaks die out before reaching the threshold.

164 Superspreading events, *i.e.* when the individual force of infection  $\mathcal{F}$  follows a Gamma distribution,  
165 seem to have limited effects on these results: the median delay drops slightly to 64 days (January 19



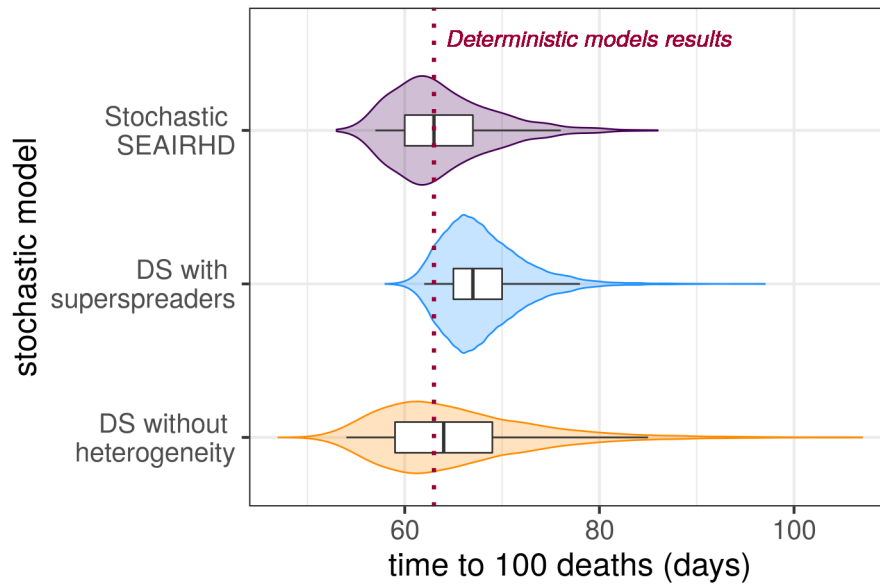


Figure 1 – **Estimated distribution of the number of days until daily mortality incidence reaches 100 deaths.** The boxplots and the whiskers indicate the 2.5%, 25%, 50%, 75%, and 97.5% quantiles out of the 10,000 simulations. The red dashed line shows the estimates using the deterministic models.

166 in France), although with a larger 95% CI, between 54 and 85 days. Moreover, as expected [15], we  
 167 observe a soar in the frequency of epidemic outbreaks dying out before reaching the threshold, which  
 168 represent 75% of our simulations.

169 When assuming a deterministic Markovian SEAIRHD model, the date of importation of the first  
 170 case of the epidemic wave that best fits the results is slightly later than the DS models estimates, with a  
 171 delay of 63 days until daily mortality incidence reaches 100 deaths. A stochastic implementation of the  
 172 same model yields the same median delay of 63 days, and a 95% confidence interval between 56 and  
 173 76 days, which is comparable to the DS model. However, consistently with earlier studies [21? ], the  
 174 ability of this memoryless model to capture the data is limited (Fig. S2 in the Appendix). Finally, the  
 175 maximum likelihood parameter estimates from a deterministic non-Markovian model [21], restricted  
 176 to the mortality data, indicates a similar delay of 63 days (January 20), with a 95%CI between 63 and 64  
 177 days.

178 We perform a sensitivity analysis of our results focusing on two of our parameters. First, we show  
 179 that the median delay for daily incidence to reach 100 deaths is decreased by 5 days when the serial  
 180 interval standard deviation is decreased by one third (Fig. S4). Those estimates therefore remain within  
 181 the confidence interval of our starting date. Second, increasing the number of initially imported cases  
 182 from 1 to 5 decreases the delay by 8 days. However, when assuming a more realistic scenario where  
 183 all those cases are not imported on the same day, we find a much more limited impact on the delay

184 (Fig. S5).

185 Overall, non-Markovian dynamics or stochasticity do not tend to significantly impact the estimate of  
186 the delay for an epidemic to reach daily mortality incidence of 100 deaths. Introducing super-spreading  
187 events, however, slightly decreases the delay estimated and greatly increases its variance. As expected,  
188 the initial number of imported cases can have an impact on the estimates.

## 189 3.2 End of the epidemic wave with lock-down

### 190 Time to eradication

191 We estimated the distribution of the minimal lock-down duration to eradicate the epidemic ( $\tau$ ). We  
192 first neglect superspreading events and start from the end of the first-wave lock-down in France on  
193 May 11 (orange violins in Figure 2). When maintaining the constraints on social interactions to their  
194 full intensity ( $\zeta_{t>55} = 0$ ), a total of at least 7.6 months of lock-down, including the 55 days between Mar  
195 17 and May 11, are required to reach a 95% extinction probability.

196 When accounting for individuals heterogeneity, we find that, everything else being equal, the quan-  
197 tiles of  $\tau$  are always lower than in homogeneous case. However, 6.9 months of lock-down at full in-  
198 tensity ( $\zeta_{t>55} = 0$ ) are still required to guarantee 95% chance of extinction (blue violins in Figure 2).  
199 Here, taking into account the individual heterogeneity reduces the variance of  $\tau$ . Indeed, transmis-  
200 sion heterogeneity implies that the majority of the infected people do not transmit, which increases the  
201 extinction probability.

202 The mean values of  $\tau$  increases with the decrease in the intensity of the lock-down constraints  
203 ( $\zeta_{t>55}$ ). As  $\zeta_t$  tends towards  $\frac{1-\eta_{FR}}{(1-\eta_{FR})\mathcal{R}_0}$  the mean values of  $\tau$  diverge towards infinity. The dynamical  
204 process is said to be critical (resp. super-critical) if  $\eta_t = \frac{1}{\mathcal{R}_0}$  (resp.  $\eta_t \geq \frac{1}{\mathcal{R}_0}$ ). This result holds true when  
205 assuming transmission heterogeneity.

### 206 Rebound risk

207 In our stochastic model, a newly infected individual may cause several secondary infections  $\delta$  days  
208 after being infectious. Therefore, the incidence at time  $t$  ( $(Y_t)_{t \in \mathbb{N}}$ ) can alternate between zero and non-  
209 zero values. To evaluate the risk of epidemic rebound, we implement a finite lock-down extension after  
210 which all constraints are released ( $\eta_t = 1 \Leftrightarrow \zeta_t = 1$ ). This allows us to calculate  $p_0(t)$ , the probability to  
211 have 0 new cases after time  $t$ . In Figure S7, we see a sharp decrease in  $p_0(t)$  a few days after lock-down  
212 release.

213 The rebound risk is directly linked to the random variable  $(F_{x,y})$  (the force of infection of an individ-

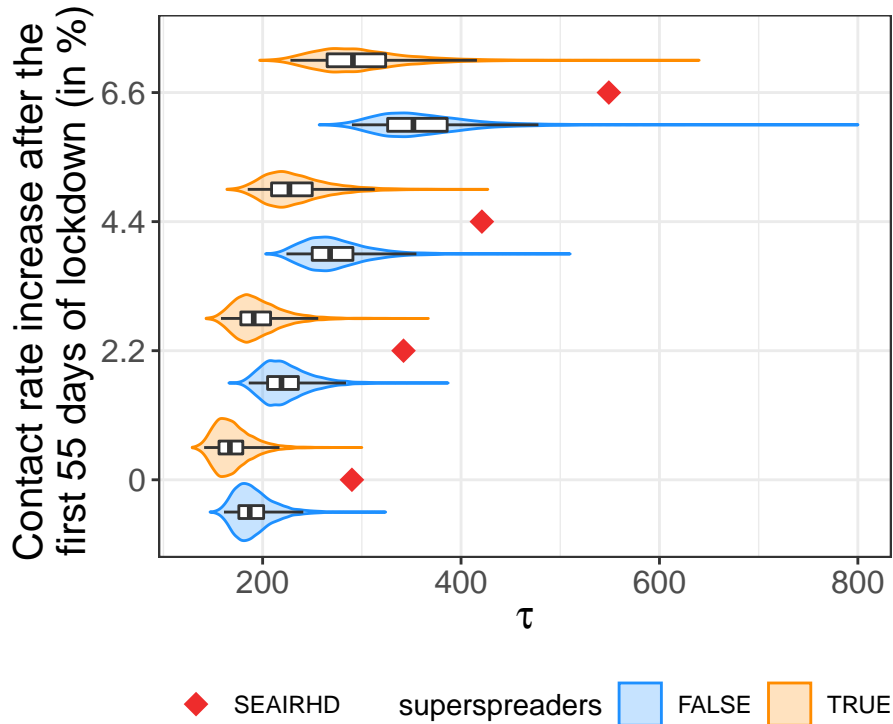


Figure 2 – Effect of the lock-down intensity, stochasticity, and superspreading events on the time to extinction ( $\tau$ ). The distributions of  $\tau$  (number of days since the start of the lock-down on Mar 17) for several lock-down intensities increase after the first 55 days (i.e. after May 11) are plotted on the Y-axis ( $\zeta_t$ ) using violin plots and boxplots. Results without transmission heterogeneity ( $\mathcal{F} = \delta(R_0)$ ) are in orange. In blue, we assume a Gamma distribution for  $\mathcal{F}$ . Red diamonds show results from the deterministic Markovian model. The box extends from the lower to upper quartiles of the data. The whiskers expand from the 2.5% to the 97.5% quantiles.

214 ual  $x$  infected  $y$  days after the start of the simulation). Assuming individual transmission heterogeneity  
 215 drastically reduces the risk of rebound, as it also implies that most infectees do not transmit the disease.

### 216 Eradication and lock-down initiation date

217 We now turn to the consequence of implementing a lock-down a month or two weeks earlier. In France,  
 218 this corresponds to Feb 17 and Mar 03 (at that time, a total of respectively 1 and 3 deaths were reported).

219 The results are shown in Figures 3 for the case without host heterogeneity and Fig. S8 with super-  
 220 spreading events. Initiating the lock-down one month earlier, i.e. for France approximately 33 days  
 221 after the onset of the epidemic wave, decreases the 95% quantile of  $\tau$  by 96 days without transmission  
 222 heterogeneity (92 days with heterogeneity) in the most restrictive scenario. If the onset of the lock-down  
 223 is brought forward by two weeks (Mar 03), i.e. in France approximately 48 days after the onset of the  
 224 epidemic, 95% of the extinction events occur before the 188<sup>th</sup> days of lock-down without transmission

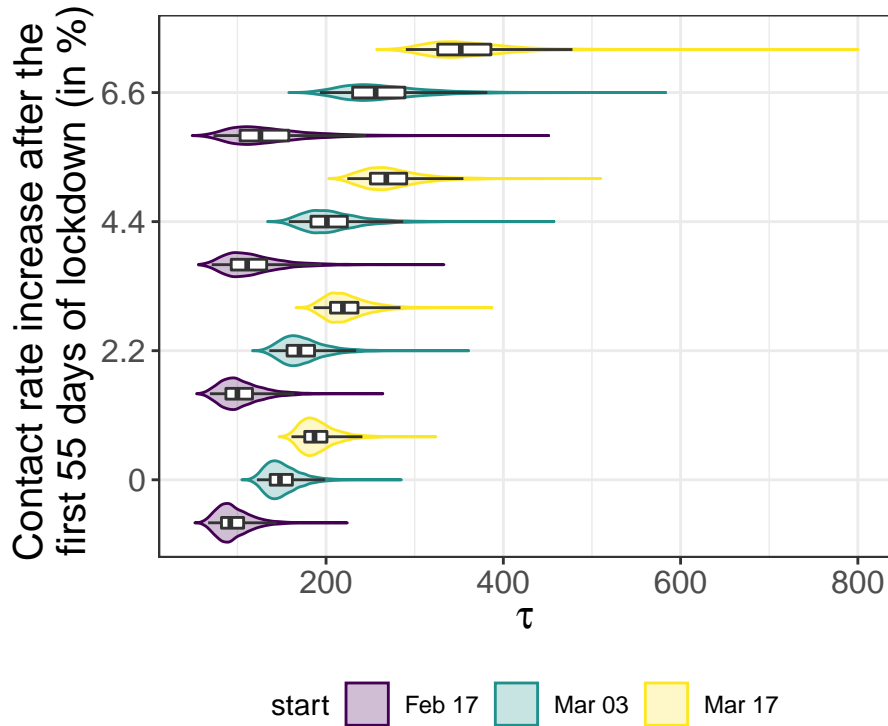


Figure 3 – **Effect of the lock-down intensity, stochasticity, and initiation date on the time to extinction ( $\tau$ ) without superspreading events.** The distributions of  $\tau$  (number of days since the start of the lock-down on Mar 17) for several lock-down intensities increase after the first 55 days are plotted on the Y-axis ( $\zeta_t$ ) using violin plots and boxplots. In this graph we assume there is no individual spreading heterogeneity. The colors indicate the different initiation date of the lock-down: in purple it starts on Feb 17, green Mar 03 and yellow on Mar 17 (official start). The box extends from the lower to upper quartiles of the data. The whiskers expand from the 2.5% to the 97.5% quantiles.

225 heterogeneity (169<sup>th</sup> days with heterogeneity). Hence a reduction of 41 (resp. 38) days of lock-down  
 226 could be expected compared to the later start (Mar 17).

227 These numbers increase with the easing of the constraints following the first 55 days of strict lock-  
 228 down ( $\eta_{FR}$ ). When assuming a lighter control in the following days (e.g.  $\zeta_{t>55} = 6.6\%$ ), one can notice  
 229 that the increase in the quantiles of  $\tau$  when starting the lock-down on Feb 17 is much lower than the  
 230 two other cases. Since the epidemic has not spread to same extent in the latter scenario, the first 55 days  
 231 of lock-down are decisive in the slow-down of the epidemic.

## 232 4 Discussion

233 In the early and final stages of an epidemic, stochastic forces may strongly affect transmission dynam-  
 234 ics because infection prevalence is low. Using stochastic mathematical modelling, we estimate the time

235 for a COVID-19 epidemic to reach an incidence of 100 deaths per day to be approximately 67 days,  
236 with a 95% probability between 62 and 79 days. In the case of France, where such incidence values  
237 were reached on Mar 23, this translates into an origin of the first epidemic around January 16, with 95%  
238 probability between January 4 and 21. This is consistent with estimates obtained using virus genome  
239 data, although these should be interpreted with caution due to the uncertainties regarding the molecu-  
240 lar clock estimates for the virus and the incomplete sampling in France [5].

241 Accounting for superspreading events does yield a later median date of origin (January 19 for  
242 France). This faster dynamic comes from the fact that simulated outbreaks that do not die out (and  
243 therefore are accounted in the results) are mostly due to early superspreading events, which can lead  
244 to a faster initial dynamic [15]. However, this difference is not significant.

245 In general, the 95% confidence intervals generated by our different models overlap. This could  
246 originate from our use of mortality data. Since death occurs after a mean delay of 23 days after infec-  
247 tion, by the time incidence starts to increase, transmission dynamics are largely deterministic. This also  
248 explains why introducing superspreading events mostly increases the origin date uncertainty. Unfortu-  
249 nately, hospital admission date is not available for France until Mar 18 and screening data was initially  
250 performed with a very low sampling rate in the country (only severe cases were tested).

251 Care must be taken when comparing the estimates from our discrete stochastic model to that of  
252 earlier models. For instance, the non-Markovian deterministic model by Sofonea *et al.* [21], which  
253 estimates the date of onset to be slightly later (January 20) includes host age structure. Regarding  
254 the more classical deterministic and Markovian SEAIRHD model, its ability to fit the data is limited  
255 (Fig. S2), except when only considering the exponential phase before the lockdown. This poor inference  
256 of underlying epidemiological dynamics is largely due to the absence of memory in the underlying  
257 processes, as stressed by earlier studies [21? ]. When incorporating memory on the hospitalization to  
258 death delay, we obtain a much better fit, and the time to 100 daily deaths is then comparable to that of  
259 the model without superspreading events.

260 We also estimated the mean number of days of full intensity lock-down required to achieve extinc-  
261 tion with a 95% confidence. With our stochastic model, we find that in average 190 (IC 95%: 183-199)  
262 days of lock-down are necessary to reach extinction in a homogeneous scenario, starting the lock-down  
263 mid-March. Accounting for superspreading events decreases the median estimate value by 20 days.  
264 Initiating the lock-down one month earlier strongly affects these estimates: a 30 days anticipated start  
265 reduces the mean number of days spent in full intensity lock-down by 96 days, i.e. a 49% reduction.  
266 This confirms that early interventions have a disproportionate impact.

267 Finally, we investigated the risk of an epidemic rebound upon lock-down lifting. In this scenario,

268 super-spreading has a striking impact as expected in limiting this risk.

269 There are several limitations to this work. First, the serial interval  $\omega$  and the time from infection  
270 to death  $\theta$ , are largely unknown in France, as well as in many countries. Most of the known serial  
271 interval estimates rely on contact tracing data from Asia [14, 16], which could be slightly different from  
272 the distribution in France, due to different contact structure, or different non-pharmaceutical measures  
273 applied. Obviously, the serial interval distribution has a strong impact on the dynamics. We do show  
274 however in Figure S3 that the variance of this interval does not have a strong impact on the results.

275 Another important limitation about the estimation of the date of origin of the epidemic comes from  
276 the assumption that only one initial infected person caused the epidemic. Clearly, most epidemics  
277 outside China were seeded by multiple importation events. The problem is that there is an identifiabil-  
278 ity issue because it is impossible to estimate both the number of initial infected cases and the time to  
279 threshold of 100 deaths with incidence data only. However, some estimates made in the UK from phy-  
280 logenetic data as well as the combination of prevalence and travel data show that the estimated number  
281 of importation events is less than 5 per day before the end of February, when the virus was beginning  
282 to circulate at higher levels throughout Europe [19]. Assuming that the dynamic was similar in France,  
283 we could verify that the dynamic was only sensitive to the importation events within the first days after  
284 the beginning of the epidemic wave. While these events may have enabled the epidemic to escape the  
285 stochastic phase faster, they would not have strongly affected the estimated date of beginning of the  
286 wave (Figure S5). In a quite extreme scenario of 5 importations per day during 30 days, the median  
287 day of the epidemic beginning was estimated to be 16 days later (*i.e.* Feb 2 for France).

288 Another limitation comes from the lack of data regarding individual heterogeneity in COVID-19  
289 epidemics. Such heterogeneity was supported by early limited data [9, 14] but recent additional ev-  
290 idence from Chinese transmission chains further supports this result [22], although with a higher  $k$   
291 parameter value than the one used here (0.30 versus 0.16 here), meaning a less heterogeneous transmis-  
292 sion. Therefore, our assessment of superspreading events impact seems conservative.

293 These results have several implications. First, they can help reconcile the fact that cases may be  
294 detected long before the emergence of the transmission chains related to an epidemic wave. This is  
295 particularly important for an audience not familiar with stochasticity. Second, the estimate of the time  
296 required to ensure that the epidemic is gone is directly informative to public health officials. In the case  
297 of France for instance, one can directly see that enforcing a strict lock-down until epidemic extinction  
298 is practically unfeasible. This may not be the case if measures are taken early enough in the epidemic.  
299 Furthermore, our work also illustrates the risk of epidemic rebound as a function of the duration of the  
300 lock-down. Overall, this work calls for further studies, especially to assess the importance of super-

301 spreading events in the global spread of SARS-CoV-2.

## 302 **Acknowledgements**

303 Thomas Beneteau is supported by a doctoral grant from the Ligue Contre le Cancer. We thank the  
304 CNRS, the IRD, and the IRD itrop HPC (South Green Platform) at IRD montpellier for providing HPC  
305 resources that have contributed to the research results reported within this paper. We also thank the  
306 ETE modelling team for discussion.

## 307 **References**

- 308 [1] Althouse, B. M., Wenger, E. A., Miller, J. C., Scarpino, S. V., Allard, A., Hébert-Dufresne, L. & Hu,  
309 H., 2020 Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for  
310 interventions and control. *PLOS Biology* **18**, e3000897. (doi: 10.1371/journal.pbio.3000897).
- 311 [2] Anderson, R. M. & May, R. M., 1991 *Infectious Diseases of Humans. Dynamics and Control*. Oxford:  
312 Oxford University Press.
- 313 [3] Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M.-L., Nalla,  
314 A., Pepper, G., Reinhardt, A., Xie, H., Shrestha, L., Nguyen, T. N., Adler, A., Brandstetter, E., Cho,  
315 S., Giroux, D., Han, P. D., Fay, K., Frazar, C. D., Ilcisin, M., Lacombe, K., Lee, J., Kiavand, A.,  
316 Richardson, M., Sibley, T. R., Truong, M., Wolf, C. R., Nickerson, D. A., Rieder, M. J., Englund,  
317 J. A., Investigators†, T. S. F. S., Hadfield, J., Hodcroft, E. B., Huddleston, J., Moncla, L. H., Müller,  
318 N. F., Neher, R. A., Deng, X., Gu, W., Federman, S., Chiu, C., Duchin, J. S., Gautom, R., Melly,  
319 G., Hiatt, B., Dykema, P., Lindquist, S., Queen, K., Tao, Y., Uehara, A., Tong, S., MacCannell, D.,  
320 Armstrong, G. L., Baird, G. S., Chu, H. Y., Shendure, J. & Jerome, K. R., 2020 Cryptic transmission  
321 of SARS-CoV-2 in Washington state. *Science* (doi: 10.1126/science.abc0523).
- 322 [4] Britton, T. & Scalia Tomba, G., 2019 Estimation in emerging epidemics: biases and remedies. *Jour-*  
323 *nal of The Royal Society Interface* **16**, 20180670. (doi: 10.1098/rsif.2018.0670).
- 324 [5] Danesh, G., Elie, B., Michalakis, Y., Sofonea, M. T., Bal, A., Behillil, S., Destras, G., Boutolleau, D.,  
325 Burrel, S., Marcelin, A.-G., Plantier, J.-C., Thibault, V., Simon-Lorriere, E., der Werf, S. v., Lina, B.,  
326 Josset, L., Enouf, V. & Alizon, S., 2020 Early phylodynamics analysis of the COVID-19 epidemic in  
327 France. *medRxiv* **2020.06.03.20119925**, ver. 3 peer-reviewed and recommended by *PCI in Evolu-*  
328 *tionary Biology*. (doi: 10.24072/pci.evolbiol.100107).

- 329 [6] Danesh, G., Saulnier, E., Gascuel, O., Choisy, M. & Alizon, S., 2020 Simulating trajectories and  
330 phylogenies from population dynamics models with TiPS. (doi: 10.1101/2020.11.09.373795).
- 331 [7] Djaafara, B. A., Imai, N., Hamblion, E., Impouma, B., Donnelly, C. A. & Cori, A., 2020 A quantita-  
332 tive framework to define the end of an outbreak: Application to ebola virus disease. *Am J Epidemiol*  
333 p. kwaa212. (doi: 10.1093/aje/kwaa212).
- 334 [8] Eichner, M. & Dietz, K., 1996 Eradication of poliomyelitis: when can one be sure that polio virus  
335 transmission has been terminated? *American journal of epidemiology* **143**, 816–822.
- 336 [9] Endo, A., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working  
337 Group, Abbott, S., Kucharski, A. J. & Funk, S., 2020 Estimating the overdispersion in COVID-19  
338 transmission using outbreak sizes outside China. *Wellcome Open Res* **5**, 67. (doi: 10.12688/well-  
339 comeopenres.15842.1).
- 340 [10] Hartfield, M. & Alizon, S., 2013 Introducing the outbreak threshold in epidemiology. *PLoS Pathog.*  
341 **6**, e1003277. (doi: 10.1371/journal.ppat.1003277).
- 342 [11] 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,  
343 X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling,  
344 B. J., Li, F. & Leung, G. M., 2020 Temporal dynamics in viral shedding and transmissibility of  
345 COVID-19. *Nat Med* pp. 1–4. Publisher: Nature Publishing Group, (doi: 10.1038/s41591-020-0869-  
346 5).
- 347 [12] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D.,  
348 Kucharski, A. J., Edmunds, W. J., Sun, F., Flasche, S., Quilty, B. J., Davies, N., Liu, Y., Clifford,  
349 S., Klepac, P., Jit, M., Diamond, C., Gibbs, H., Zandvoort, K. v., Funk, S. & Eggo, R. M., 2020 Fea-  
350 sibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global*  
351 *Health* **0**. (doi: 10.1016/S2214-109X(20)30074-7).
- 352 [13] Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., Yuan, B.,  
353 Kinoshita, R. & Nishiura, H., 2020 Incubation Period and Other Epidemiological Characteristics  
354 of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly  
355 Available Case Data. *Journal of Clinical Medicine* **9**, 538. (doi: 10.3390/jcm9020538).
- 356 [14] Liu, Y., Eggo, R. M. & Kucharski, A. J., 2020 Secondary attack rate and superspreading events for  
357 SARS-CoV-2. *The Lancet* **395**, e47. (doi: 10.1016/S0140-6736(20)30462-1).



- 358 [15] Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M., 2005 Superspreading and the effect  
359 of individual variation on disease emergence. *Nature* **438**, 355–9. (doi: 10.1038/nature04153).
- 360 [16] Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R., 2020 Serial interval of novel coron-  
361 avirus (COVID-19) infections. *International Journal of Infectious Diseases* **93**, 284–286. (doi:  
362 10.1016/j.ijid.2020.02.060).
- 363 [17] Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R., 2020 Serial interval of novel coronavirus  
364 (COVID-19) infections. *International Journal of Infectious Diseases* **0**. (doi: 10.1016/j.ijid.2020.02.060).
- 365 [18] Nishiura, H., Miyamatsu, Y. & Mizumoto, K., 2016 Objective Determination of End of MERS Out-  
366 break, South Korea, 2015. *Emerging Infectious Diseases* **22**, 146–148. (doi: 10.3201/eid2201.151383).
- 367 [19] Plessis, L. d., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ash-  
368 worth, J., Colquhoun, R., Connor, T. R., Faria, N. R., Jackson, B., Loman, N. J., O’Toole, , Nicholls,  
369 S. M., Parag, K. V., Scher, E., Vasylyeva, T. I., Volz, E. M., Watts, A., Bogoch, I. I., Khan, K., Con-  
370 sortium†, C.-. G. U. C.-U., Aanensen, D. M., Kraemer, M. U. G., Rambaut, A. & Pybus, O. G., 2021  
371 Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* Publisher:  
372 American Association for the Advancement of Science Section: Research Article, (doi: 10.1126/sci-  
373 ence.abf2946).
- 374 [20] R Core Team, 2020 *R: A Language and Environment for Statistical Computing*. R Foundation for  
375 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- 376 [21] Sofonea, M. T., Reyné, B., Elie, B., Djidjou-Demasse, R., Selinger, C., Michalakis, Y. & Alizon, S.,  
377 2020 Epidemiological monitoring and control perspectives: application of a parsimonious mod-  
378 elling framework to the COVID-19 dynamics in France. *medRxiv* p. 2020.05.22.20110593. (doi:  
379 10.1101/2020.05.22.20110593).
- 380 [22] Sun, K., Wang, W., Gao, L., Wang, Y., Luo, K., Ren, L., Zhan, Z., Chen, X., Zhao, S., Huang, Y., Sun,  
381 Q., Liu, Z., Litvinova, M., Vespignani, A., Ajelli, M., Viboud, C. & Yu, H., 2020 Transmission het-  
382 erogeneities, kinetics, and controllability of SARS-CoV-2. *Science* Publisher: American Association  
383 for the Advancement of Science Section: Research Article, (doi: 10.1126/science.abe2424).
- 384 [23] Thompson, R. N., Morgan, O. W. & Jalava, K., 2019 Rigorous surveillance is necessary for high  
385 confidence in end-of-outbreak declarations for Ebola and other infectious diseases. *Philosophical*  
386 *Transactions of the Royal Society B: Biological Sciences* **374**, 20180431. (doi: 10.1098/rstb.2018.0431).

387 [24] Trapman, P., Ball, F., Dhersin, J.-S., Tran, V. C., Wallinga, J. & Britton, T., 2016 Inferring  $R_0$  in  
388 emerging epidemics—the effect of common population structure is small. *J. R. Soc. Interface* **13**,  
389 20160288. (doi: 10.1098/rsif.2016.0288).