medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Super-spreading events initiated the exponential growth phase of COVID-19 with $\mathcal{R}_0$ higher than initially estimated

Marek Kochańczyk[1], Frederic Grabowski[2], and Tomasz Lipniacki[1,*]

[1]Department of Biosystems and Soft Matter, Institute of Fundamental Technological Research, Polish Academy of Sciences, 02-106 Warsaw, Poland

[2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 02-097 Warsaw, Poland

*Corresponding author email: tlipnia@ippt.pan.pl

Abstract

The basic reproduction number $\mathcal{R}_0$ of the coronavirus disease 2019 has been estimated to range between 2 and 4. Here we used a SEIR model that properly accounts for the distribution of the latent period and, based on empirical estimates of the doubling time in the near-exponential phases of epidemic progression in several locations, we estimated that $\mathcal{R}_0$ lies in the range 4.7–11.4. We explained this discrepancy by performing stochastic simulations of model dynamics in a population with a small proportion of super-spreaders. The simulations revealed two-phase dynamics, in which an initial phase of relatively slow epidemic progression diverts to a faster phase upon appearance of infectious super-spreaders. Early estimates obtained for this initial phase may suggest lower $\mathcal{R}_0$.

Introduction

The basic reproduction number $\mathcal{R}_0$ is a critical parameter characterising the dynamics of an outbreak of an infectious disease. By definition, $\mathcal{R}_0$ quantifies the expected number of secondary cases generated by an infectious individual in an entirely susceptible population. $\mathcal{R}_0$ may be influenced by natural conditions (such as seasonality) as well as socioeconomic factors (such as population density or ingrained societal norms and practices) [1]. Accurate estimation of $\mathcal{R}_0$ is of crucial importance because it informs the extent of control measures that should be implemented to terminate the spread of an epidemic. Also, $\mathcal{R}_0$ determines the immune proportion f of population that is required to achieve herd immunity, $f = 1 - 1/\mathcal{R}_0$.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

A preliminary estimate published by the World Health Organization (WHO) suggested that $\mathcal{R}_0$ of coronavirus disease 2019 (COVID-19) lies in between 1.4 and 2.5 [2]. Later this estimate has been revised to 2–2.5 [3], which is broadly in agreement with numerous other studies that, based on official data from China, implied the range of 2–4 (see, *e.g.*, Liu *et al*. [4] or Boldog *et al*. [5] for a summary). This range suggests an outbreak of a contagious disease that should be containable by imposition of moderate restrictions on social interactions. Unfortunately, moderate restrictions that were implemented in, *e.g.*, Italy or Spain turned out to be insufficient to prevent a surge of daily new cases and, consequently, nationwide quarantines had to be introduced.

We estimated the range of $\mathcal{R}_0$ of COVID-19 based on the doubling times observed in the exponential phases of the epidemic in China, Italy, Spain, France, United Kingdom, Germany, Switzerland, and New York State. For each of these locations, we used trajectories of both cumulative confirmed cases and deaths [6]. Since our stochastic simulations suggested that the epidemic may have two-phase dynamics — slow (and susceptible to extinction) before any super-spreading events occur and fast and steadily expanding after the occurrence of super-spreading events — to capture the second phase of the trajectories, we analysed them after a fixed threshold of cases or deaths has been exceeded, in two-week intervals. Both the stochastic simulations and $\mathcal{R}_0$ estimates were obtained within a susceptible–exposed–infected–removed (SEIR) model that correctly reproduces the shape of the latent period distribution and yields a plausible mean generation time. We concluded that the range of $\mathcal{R}_0$ is 4.7–11.4, which is considerably higher than most early estimates. We conjecture that these early estimates were obtained for the first phase of the epidemic in which super-spreading events were absent.

Results

The SEIR model

We used a SEIR model (see Methods for model equations and justification of parameter values) in which:

- we assumed that the latent period is the same as the incubation period and is Erlang-distributed with the shape parameter $m = 6$ and the mean of 5.28 days = $1/\sigma$ [7];
- we assumed that the infectious period is Erlang-distributed with the shape parameter $n = 1$ (exponentially distributed) or $n = 2$, and the mean of 2.9 days = $1/\gamma$ [8, 9];

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

- the infection rate coefficient β was determined from σ, γ, m, n, and doubling time $\mathcal{T}_\mathrm{d}$, which in turn was estimated based on the epidemic data as described in the next subsection, ultimately allowing us to estimate $\mathcal{R}_\mathrm{o} = \beta/\gamma$ as $\mathcal{R}_\mathrm{o}(\mathcal{T}_\mathrm{d})$.

The use of the Erlang distributions directly translates to the inclusion of multiple consecutive substates in the SEIR model, meaning that we assumed m 'exposed' substates and n 'infectious' substates (Erlang distribution is a distribution of a sum of independent exponentially distributed variables of the same mean).

Estimation of $\mathcal{R}_\mathrm{o}$ in the exponential growth phase

First, we estimated the doubling time $\mathcal{T}_\mathrm{d}$ within two-week periods beginning on the day in which the number of confirmed (in the SEIR model naming convention, 'removed', see Methods) cases exceeded 100 or the number of deaths exceeded 10 in China, six European countries, and New York State (figure 1*a* and figure 1*b*). Values of $\mathcal{T}_\mathrm{d}$ that we obtained lie in between $\mathcal{T}_\mathrm{d}^\mathrm{min} = 1.86$ (based on cases in New York State) and $\mathcal{T}_\mathrm{d}^\mathrm{max} = 2.96$ (based on deaths in Switzerland).

Then, we estimated the range of $\mathcal{R}_\mathrm{o}$ as a function of the doubling time $\mathcal{T}_\mathrm{d}$ using a formula that takes into account the mean latent and infectious period, $1/\sigma$ and $1/\gamma$, respectively, as well as the shape parameters m and n, see equation (8) in Methods. The lower bound has been obtained using the model variant with $n = 2$ (two 'infectious' substates), whereas the upper bound results from the model with $n = 1$ (one 'infectious' substate), figure 1*c*. After plugging $\mathcal{T}_\mathrm{d}^\mathrm{min}$ and $\mathcal{T}_\mathrm{d}^\mathrm{max}$ in, respectively, the variant of our model with the lower $\mathcal{R}_\mathrm{o}(\mathcal{T}_\mathrm{d})$ curve ($n = 2$) and the variant with the higher $\mathcal{R}_\mathrm{o}(\mathcal{T}_\mathrm{d})$ curve ($n = 1$), we arrived at the estimated $\mathcal{R}_\mathrm{o}$ range of 4.7–11.4. The cases-based doubling time for China, 2.34, is consistent with the value of 2.4 reported by Sanche *et al.* [10], who estimated that $\mathcal{R}_\mathrm{o}$ for China lies in the range 4.7 to 6.6, that overlaps with our estimated range for China: 5.6–7.3. The models having one or two 'exposed' substates, often used to estimate the value of $\mathcal{R}_\mathrm{o}$, substantially underestimated $\mathcal{R}_\mathrm{o}$, *cf.* figure 1*c* and the articles by Wearing *et al.* [11], Wallinga & Lipsitch [12], and Kochańczyk *et al.* [13].

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

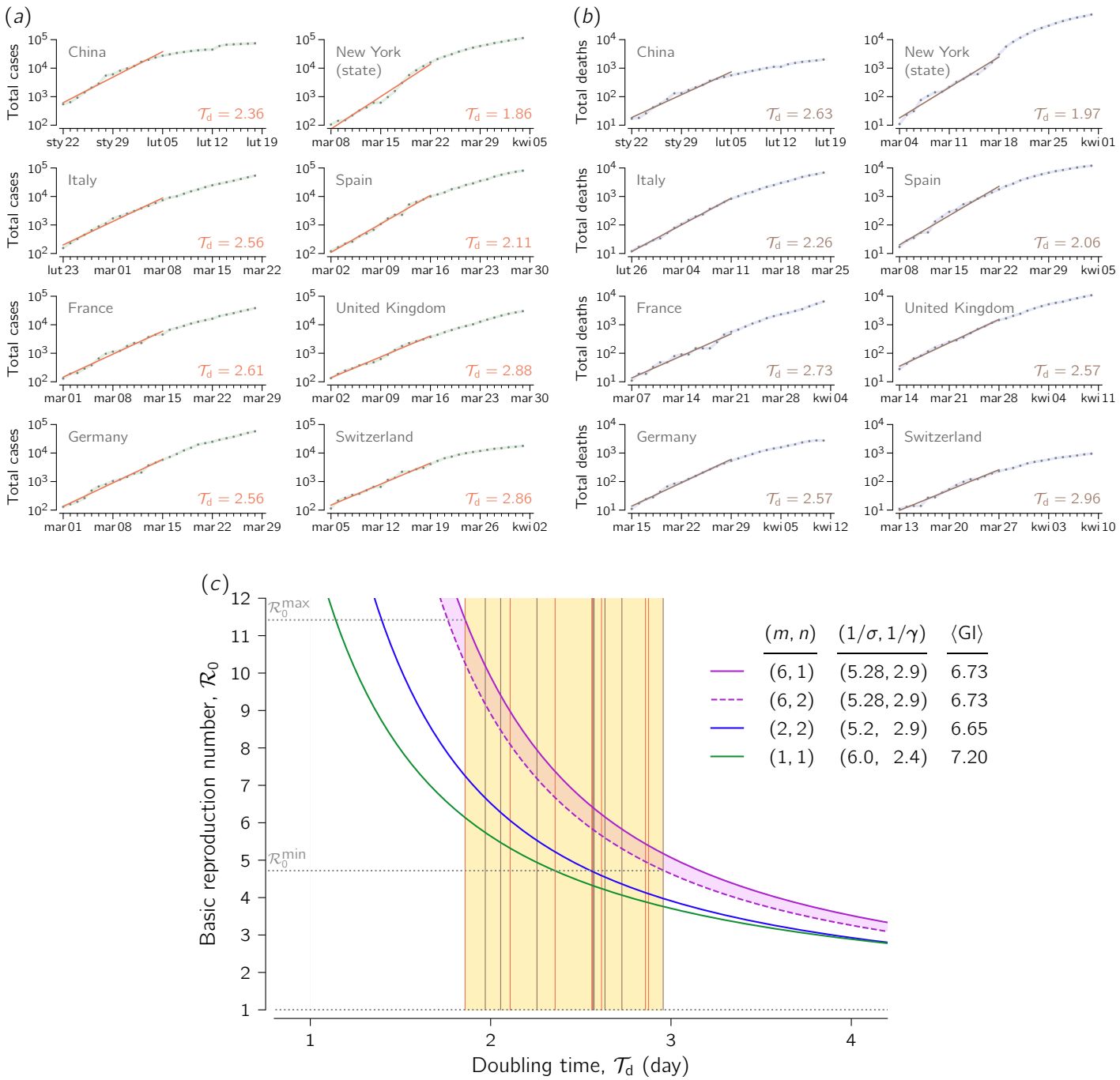

Figure 1: Estimation of the doubling time and the resulting basic reproduction number $\mathcal{R}_0$. (*a*, *b*) Estimates of the doubling time $\mathcal{T}_d$ for China, six European countries, and New York State using two-week periods beginning (*a*) when the number of confirmed cases exceeds 100 or (*b*) when the number of deaths exceeded 10, according to data gathered and made available by Johns Hopkins University [6]. (*c*) The range of $\mathcal{R}_0$ estimated using two variants of our SEIR model (violet solid and dashed curves) for the range of $\mathcal{T}_d$ estimated in panels *a* and *b*. Vertical lines in the yellow area are $\mathcal{T}_d$ estimates based on the cumulative number cases (orange, from panel *a*) or the cumulative number of deaths (brown, from panel *b*). Blue and green solid curves correspond to $\mathcal{R}_0(\mathcal{T}_d)$ according to SEIR models structured and parametrised as in the study of Kucharski *et al.* [9] ($m = 2$, $n = 2$) and Wu *et al.* [14] ($m = 1$, $n = 1$).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

There are two main reasons why our estimates of the basic reproduction number are higher compared to other published estimates:

1. Our SEIR model comprises 6 'exposed' substates to account for the latent period distribution. As shown in figure 1c, broader latent period distributions, exponential or Erlang with $m = 2$, result in lower $\mathcal{R}_0$ estimates (at the same remaining model parameters). We demonstrated sensitivity of $\mathcal{R}_0$ with respect to the mean latent period, $1/\sigma$, in electronic supplementary material (figure S1).

2. We estimated the doubling time, $\mathcal{T}_d$, from the growth of the number of cumulative cases and cumulative deaths in the two-week-long exponential phases of the epidemic in six locations, obtaining $\mathcal{T}_d$ ranging from 1.86 to 2.96. These values are much lower than the values reported in the early influential studies of Wu *et al*. [14, 15] and Li *et al*. [16]: 5.2 days, 6.4 days, and 7.4 days, correspondingly. In these studies the basic reproduction number has been estimated to lie in between 1.94 and 2.68. A summary in Table 1 shows that the lower $\mathcal{R}_0$ estimates follow from much longer estimates of T_d.

T_d	$1/\sigma$	$1/\gamma$	$\langle SI \rangle$ or $\langle GI \rangle$	$\mathcal{R}_0$	Reference
?	5.2	2.9	6.65[a]	2.35 (1.15–4.77)	Kucharski *et al*. [9]
5.2 (4.6–6.1)	6.5	?	7.0 (5.8–8.1)	1.94 (1.83–2.06)	Wu *et al*. [15]
6.4 [5.8–7.1]	6	2.4[b]	8.4	2.68 (2.47–2.86)	Wu *et al*. [14]
7.4	5.2 (4.1–7.0)	?	7.5 (5.3–19)	2.2 (1.4–3.9)	Li *et al*. [16]

Table 1: Relation between $\mathcal{T}_d$, model parameters (mean latent period or mean incubation period, $1/\sigma$, mean period of infectiousness, $1/\gamma$, and consequent mean generation interval, $\langle GI \rangle$), mean searial period, $\langle SI \rangle$, and $\mathcal{R}_0$. The unit of all values, except for $\mathcal{R}_0$, is day. Confidence intervals are given in oval brackets; a credible interval is given in square brackets. [a]The $\langle GI \rangle$ value is not given in the article but calculated from the assumed values of $1/\sigma$ and $1/\gamma$ as $\langle GI \rangle = 1/\sigma + \frac{1}{2}/\gamma$ [17]. [b]The value $1/\gamma$ was obtained by the authors as $\langle SI \rangle - 1/\sigma$, which is inconsistent with the assumption that the infection occurs in a random time during the period of infectiousness.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Impact of super-spreading on $\mathcal{T}_\mathrm{d}$ estimation

The discrepancy in $\mathcal{T}_\mathrm{d}$ estimation may be potentially attributed to the fact that not all 'removed' individuals are registered. In the case when the ratio of registered to 'removed' individuals is increasing over time, the true increase of the 'removed' cases may be overestimated. We do not rule out this possibility, although we consider it implausible as the expansion of testing capacity in considered countries has been slower than the progression of the outbreak. We rather attribute the discrepancy to the fact that in the early phase, in which the doubling time (growth rate) is estimated based on individual case reports, the consequences of potential super-spreading events (such as football matches, carnival fests, demonstrations, masses, or hospital-acquired infections) are negligible due to a low probability of such events when the number of infected individuals is low. In a given region or country, occurrence of first super-spreading events triggers transition to the faster-exponential growth, in which subsequent super-spreading events become statistically significant and may become decisive drivers of the epidemic spread [18]. Based on case reports in China, Sanche *et al*. [10] inferred that the initial epidemic period in Wuhan has been dominated by simple transmission chains. Phylogenetic analyses by Worobey *et al*. [19] revealed that first cases recorded in USA and Europe did not initiate sustained SARS-CoV-2 transmission networks. In turn, super-spreading events were very likely the main drivers of the epidemic spread in, *e.g*., Italy and Germany, where, in the early exponential phase, spatial heterogeneity of registered cases has been evident [20, 21]. In Italy, Spain, and France, this explosive phase was followed by a phase of slower growth, during which mass gatherings were forbidden, but quarantine (that finally brought the effective reproduction number below 1) has not been yet introduced.

Motivated by these considerations, we analysed the impact of super-spreading on estimation of $\mathcal{T}_\mathrm{d}$ based on stochastic simulations of SEIR model dynamics (see electronic supplementary material, listing S1). Simulations were performed in the perfectly mixed regime according to the Gillespie algorithm [22]. We assumed that a predefined fixed proportion of individuals (equal 33%, 10%, 3% or 1%) has higher infectiousness and as such is responsible for on average either half of infections ('super-spreaders') or two-third of infections ('hyper-spreaders'). To reproduce these fractions in systems with different assigned proportions of super- or hyper-spreaders, their infectiousness is assumed to be inversely proportional to their ratio in the simulated population. In figure 2 we show dynamics of the epidemic spread in the presence of 1% of hyper-spreaders to demonstrate that the phase of

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

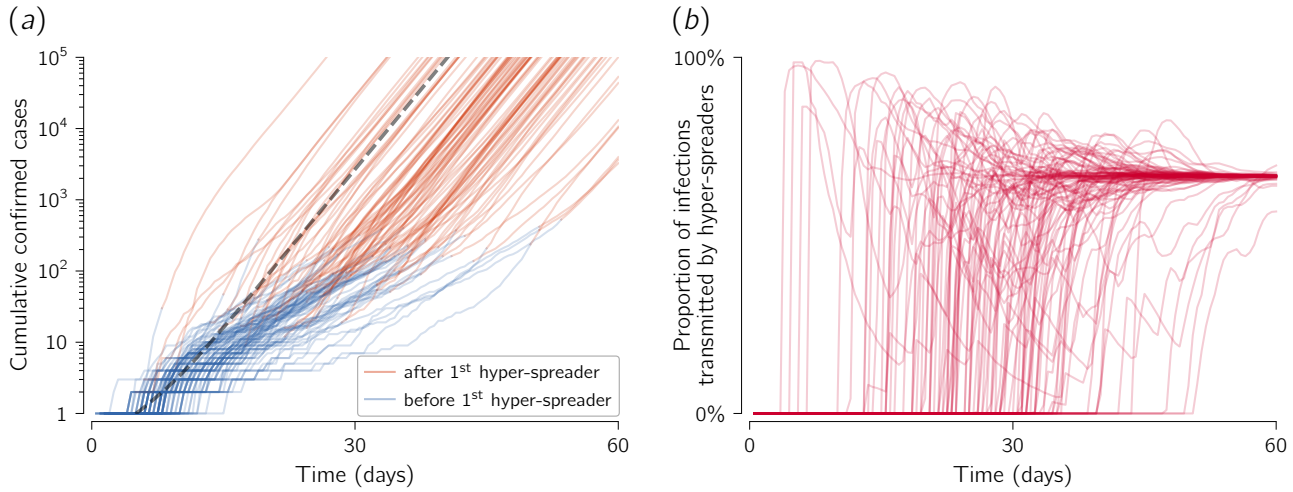

Figure 2: Stochastic epidemic spread in the presence of 1% of hyper-spreaders. (*a*) Trajectories of confirmed cases (cumulative R in terms of SEIR compartments) resulting from 100 independent stochastic simulations. When the first hyper-spreading event occurs, the colour of the line is changed from blue to brown. Dashed grey line shows a deterministic trajectory. (*b*) Proportion of infections transmitted by hyper-spreaders among all transmission events over time. Stochastic trajectories stabilise at 66.7%. Trajectories shown in both panels results from the same set of simulations; simulations resulting in outbreak failure were discarded. Model parameters used for simulations in both panels: $(m, n) = (6, 1)$, $(1/\sigma, 1/\gamma) = (5.28 \text{ days}, 2.9 \text{ days})$. Infection rate coefficient of hyper-spreaders was set $\beta_h = 198 \times \beta_n$ (where β_n is the infection rate coefficient for normal spreaders), which assures that in the deterministic limit 66.7% of infections are transmitted by hyper-spreaders. In turn β_n was set such that the average infection rate coefficient $\beta = 2.97 \times \beta_n$ gives $\mathcal{T}_d = 2$ days (see equation (7) in Methods).

slower growth is transformed into the faster-exponential growth phase upon the occurrence of hyper-spreading events.

We estimated $\mathcal{T}_d$ in two ways: based on one month of growth of the number of new cases since the first registered case ('30 days since the 1st case') and based on growth of new cases in the two-week period after the number of registered cases exceeds 100 ('14 days since 100 cases'). As we are interested in the initial phase characterised by exponential growth, we assumed that the susceptible population remains constant. In figure 3 we show histograms of $\mathcal{T}_d$ calculated using either the '14 days since 100 cases' method or the '30 days since the 1st case' method. One may observe that the histograms calculated using the '30 days since the 1st case' method are broader than those calculated using the '14 days since 100 cases' method, and the width of all histograms increases with increasing infectiousness (which is set inversely proportional to ρ). When $\mathcal{T}_d$ is calculated using the '14 days since 100 cases' method, its median value is slightly larger than $\mathcal{T}_d$ in the deterministic model (equal

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

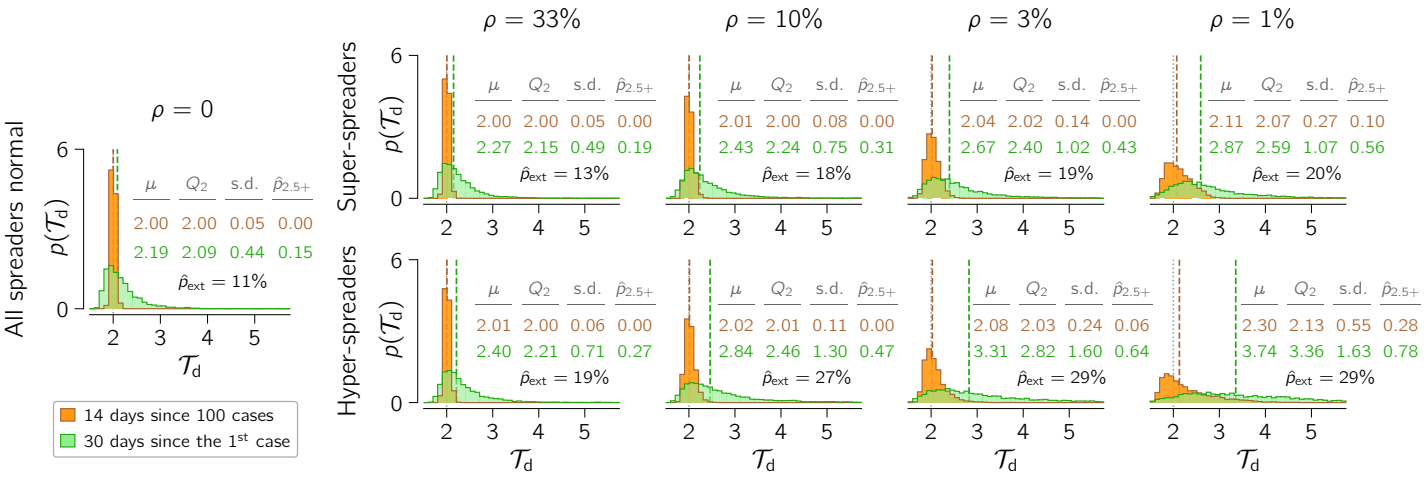

Figure 3: Estimation of the doubling time $\mathcal{T}_d$ based on stochastic simulations of the SEIR model with super- and hyper-spreaders. Histograms show probability density $p(\mathcal{T}_d)$ estimated using the '14 days since 100 cases' method (orange) and the '30 days since the 1st case' method (green). In each column, ρ denotes a fixed proportion of super-spreaders (top row) or hyper-spreaders (bottom row) in the population. For decreasing proportions of super- and hyper-spreaders (from left, except the shared leftmost panel with $\rho = 0$, to right), their infection rate coefficient β has been reduced to give the same deterministic $\mathcal{T}_d = 2$ days (vertical dotted grey lines). Remaining model parameters: $(m, n) = (6, 1)$; $(1/\sigma, 1/\gamma) = (5.28 \text{ days}, 2.9 \text{ days})$. Each histogram results from 5,000 stochastic simulations starting from a single infected normal individual; trajectories resulting in outbreak failure were discarded; fraction of trajectories that resulted in epidemic extinction for given conditions is given as $\hat{p}_{ext}$. Each distribution is described in terms of its mean (μ), median (Q_2 and vertical dashed lines), standard deviation (s.d.), and the fraction of probability mass for $\mathcal{T}_d > 2.5$ days ($\hat{p}_{2.5+}$).

2 days); however, when $\mathcal{T}_d$ is calculated using the '30 days since the 1st case' method, then for high infectiousness of super- and hyper-spreaders (correspondingly, for low ρ) its median value becomes much larger than the deterministic $\mathcal{T}_d$. Using the '30 days since the 1st case' method for the case of the lowest considered ρ = 1%, when super-spreaders (hyper-spreaders) have their infectiousness about 100 times (200 times) higher than the infectiousness of normal individuals, one obtains median $\mathcal{T}_d$ larger than $\mathcal{T}_d$ obtained in the deterministic model by 29% (67%), while for '14 days since 100 cases' the $\mathcal{T}_d$ overestimation is negligible, 3% (6%). This difference is caused by low probability of appearance of super- or hyper-spreaders in the first weeks of the outbreak.

Finally, we notice that $\mathcal{T}_d$ estimation for a given country based on available data is equivalent to the analysis of a single stochastic trajectory and that at a very initial stage the epidemic can cease. Probability of extinction is larger when a small fraction of super-spreaders is responsible for a large fraction of cases. In figure 3 we provided extinction probability, $\hat{p}_{ext}$, which in the extreme case of 1% of hyper-spreaders reaches 29%, whereas without hyper-spreaders (or super-spreaders) is 11%.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The examples shown in figure 2 and figure 3 are focused on the case in which the $\mathcal{T}_d = 2$ days, which is close to $\mathcal{T}_d$ estimated for Spain and New York State. After removing super-spreaders (assumed to be responsible for 50% of transmissions) the doubling time would be equal 3.05 days, whereas after removing hyper-spreaders (responsible for 66.7% of transmissions) the doubling time would be equal to 4.24 days. The doubling times in the range 5.2–7.4, obtained by analysing early onsets of the epidemic (Wu *et al*. [14, 15] and Li *et al*. [16]), exceed our model prediction obtained after removing 66.7% of transmissions by hyper-spreaders, suggesting that the fraction of transmissions for which hyper-spreaders are responsible can be even larger. Endo *et al*. estimated that 80% of secondary transmissions could have been caused by 10% of infectious individuals [18].

Conclusions

Based on epidemic data from China, United States, and six European countries, we have estimated that the basic reproduction number $\mathcal{R}_0$ lies in the range 4.7–11.4 (5.6–7.3 for China), which is higher than most previous estimates [5, 8, 4]. There are two sources of the discrepancy in $\mathcal{R}_0$ estimation. First, in agreement with data on the incubation period distribution (assumed to be the same as the latent period distribution), we used a model with six 'exposed' states, which substantially increases $\mathcal{R}_0(\mathcal{T}_d)$ with respect to the models with one or two 'exposed' states. Second, we estimated $\mathcal{T}_d$ based on the two-week period of the exponential growth phase beginning on the day in which the number of cumulative registered cases exceeds 100, or when the number of cumulative registered fatalities exceeds 10. Importantly, values of $\mathcal{T}_d$ estimated from the growth of registered cases and from the growth of the registered fatalities led to similar $\mathcal{R}_0$ estimates. This approach, in contrast to estimation of $\mathcal{R}_0$ based on individual case reports, allows to implicitly take into account super-spreading events that substantially shorten $\mathcal{T}_d$. Spatial heterogeneity of the epidemic spread observed in many European countries, including Italy, Spain, and Germany, can be associated with larger or smaller super-spreading events that initiated outbreaks is particular regions of these countries.

Our estimates are consistent with current epidemic data in Italy, Spain, and France. As of April 24, 2020, these countries managed to terminate the exponential growth phase by means of country-wide quarantine. Current COVID-19 Community Mobility Reports [23] show about 80% reduction of mobility in retail and recreation, transit stations, and workplaces in these countries. Together with increased social distancing, this reduction possibly lowered the infection rate β at least five-fold; additionally, massive testing reduced the infectious period, $1/\gamma$. Consequently, we suspect that the

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

reproduction number $R = \beta/\gamma$ was reduced more than five-fold, which brought it to the values somewhat smaller than 1. This suggests that $\mathcal{R}_0$ in these countries could have been larger than 5.

Methods

SEIR model equations and parametrisation

The dynamics of our SEIR model is governed by the following system of ordinary differential equations:

$$\frac{dS}{dt} = -\beta I(t) S(t)/N \tag{1}$$

$$\frac{dE_1}{dt} = \beta I(t) S(t)/N - m\sigma E_1(t), \tag{2}$$

$$\frac{dE_i}{dt} = m\sigma E_{i-1}(t) - m\sigma E_i(t), \quad 2 \leq i \leq m, \tag{3}$$

$$\frac{dI_1}{dt} = m\sigma E_m(t) - n\gamma I_1(t), \tag{4}$$

$$\frac{dI_j}{dt} = n\gamma I_{j-1}(t) - n\gamma I_j(t), \quad 2 \leq j \leq n, \tag{5}$$

$$\frac{dR}{dt} = n\gamma I_n(t), \tag{6}$$

where $N = S(t) + E_1(t) + \ldots + E_m(t) + I_1(t) + \ldots + I_n(t) + R(t)$ is the constant population size, and $I(t) = I_1(t) + \ldots + I_n(t)$ is the size of infectious subpopulation. As m is the number of 'exposed' substates and n is the number of 'infectious' substates, there are $m + n + 2$ equations in the system. In the early phase of the epidemic, $1 - S(t)/N \ll 1$ and with constant coefficients β, σ, and γ the growth of R (as well as E_i and I_j) is exponential.

An important property of a given SEIR model parametrisation is its implied distribution of generation interval (GI), the period between subsequent infections events in a transmission chain. While the expected GI is easily computable from model parameters as $\langle \mathrm{GI} \rangle = \sigma^{-1} + \frac{1}{2}\gamma^{-1}$ (the mean period of infectiousness is halved to reflect the assumption that the infection occurs in a random time during the period of infectiousness [17]), it can be hardly estimated based on even detailed epidemiological data. It should be noted that in some sources the formula $\langle \mathrm{GI} \rangle = \sigma^{-1} + \gamma^{-1}$ is used (see, *e.g.*, Ref. [24]), in which it is assumed that the infection occurs at the end of the period of infectiousness, not at a random point of this period. GI may be related to the serial interval, SI, the period between the occurrence of symptoms in the infector and the infectee. Although GI and SI may have different distributions, their means are expected to be equal and thus may be directly compared. Our

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

parametrisation implies $\langle \mathrm{GI} \rangle = 6.73$ days, which is consistent with $\langle \mathrm{GI} \rangle$ of the model by Ferguson *et al*. (6.5 days) [25] and the estimates of $\langle \mathrm{SI} \rangle$ by Wu *et al*. (7.0 days) [15], Ma *et al*. (6.8 days) [26], or Bi *et al*. (6.3 days) [27].

The short period of effective infectiousness reflects the assumption that the individuals with confirmed infection are quickly isolated or self-isolated and then cannot infect other susceptible individuals. This enabled us to identify the reported increase of confirmed cases with the transfer of the individuals from the (last substate of the) 'infectious' compartment to the 'removed' compartment of the SEIR model. In addition to the currently diseased individuals that remain isolated, the 'removed' compartment contains the recovered (and assumed to be resistant) and deceased individuals.

We assume the same $1/\gamma = 2.9$ days in all locations and times, being however aware that the mean infectious period may shorten over time due to the implementation of protective health care practices, increased diagnostic capacity, and contact tracing [27]. In turn, the mean latent period, $1/\sigma$, may be considered an intrinsic property of the disease. As the distribution of the latent period is not known, as a simplification, in our model the distribution of the latent period (time since infection during which an infected individual cannot infect) is assumed to be the same as the distribution of the incubation period (time since infection during which an infected individual has not yet developed symptoms). We demonstrated the influence of $1/\sigma$ on the estimation of $\mathcal{R}_0$ in electronic supplementary material (figure S1).

Estimation of the doubling time and the basic reproduction number

Growth rates used for estimation of respective doubling times, $\mathcal{T}_\mathrm{d}$, were determined by linear regression of the logarithm of the cumulative confirmed cases and cumulative deaths in the exponential phase of the epidemic separately in each of eight considered location. We discarded initial parts of trajectories with less than 100 confirmed cases (or 10 registered fatalities) and used two-week-long periods to strike a balance between: (*i*) analysis of epidemic progression when stochastic effects associated with individual transmission events, including super-spreading, are relatively small (see stochastic simulation trajectories in figure 2*a*) and (*ii*) analysis of the exponential phase of epidemic progression, which is relatively short due to imposition of restrictions. We expect that the trajectories of deaths may be less affected by under-reporting; nevertheless, doubling times obtained from growth rates of cumulative cases and cumulative deaths turn out to be quite consistent.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

In the context of our SEIR model, the doubling time $\mathcal{T}_\mathrm{d}$ and parameters β, σ, γ, n, m satisfy the relation

$$\beta(\mathcal{T}_\mathrm{d};\, \sigma, \gamma, m, n) = \frac{\dfrac{\log 2}{\mathcal{T}_\mathrm{d}} \left(\dfrac{\log 2}{\mathcal{T}_\mathrm{d}\, m\, \sigma} + 1 \right)^{m}}{1 - \left(\dfrac{\log 2}{\mathcal{T}_\mathrm{d}\, n\, \gamma} + 1 \right)^{-n}} \tag{7}$$

that enables calculation of the basic reproduction number using the doubling time $\mathcal{T}_\mathrm{d}$ estimated directly from the epidemic data as

$$\mathcal{R}_\mathrm{o}(\mathcal{T}_\mathrm{d}) = \frac{\beta(\mathcal{T}_\mathrm{d};\, \sigma, \gamma, m, n)}{\gamma} \tag{8}$$

in accordance with Wearing *et al*. [11] and Wallinga & Lipsitch [12].

Data accessibility. All data used in this theoretical study is referenced.

Authors' contributions. M.K. conceived study, performed model and data analysis, prepared figures and wrote manuscript; F.G. conceived study, performed model analysis and prepared figures; T.L. conceived study, wrote manuscript. All authors gave final approval for publication.

Competing interests. The authors declare no competing interests.

Funding. This study was supported by the National Science Centre (Poland) grant number 2018/29/B/NZ2/00668.

References

[1] Delamater P, Street E, Leslie T, Yang YT, Jacobsen K. 2019 Complexity of the Basic Reproduction Number (R_0). *Emerging Infectious Disease* **25**, 1. (doi:10.3201/eid2501.171901).

[2] WHO. 2020a Statement on the meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus 2019 (n-CoV) on 23 January 2020. https://www.who.int/news-room/detail/23-01-2020-statement-on-the-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov). Accessed: 2020-06-26.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[3] WHO. 2020b Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19). Accessed: 2020-06-26.

[4] Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. 2020 The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine* **27**, taaa021. (doi:10.1093/jtm/taaa021).

[5] Boldog P, Tekeli T, Vizi Z, Dénes A, Bartha FA, Röst G. 2020 Risk Assessment of Novel Coronavirus COVID-19 Outbreaks Outside China. *Journal of Clinical Medicine* **9**, 571. (doi:10.3390/jcm9020571).

[6] Dong E, Du H, Gardner L. 2020 An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, 533–534. (doi:10.1016/S1473-3099(20)30120-1).

[7] Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith H, Azman AS, Reich NG, Lessler J. 2020 The incubation period of 2019-nCoV from publicly reported confirmed cases: Estimation and application. *medRxiv*. (doi:10.1101/2020.02.02.20020016).

[8] Liu T, Hu J, Xiao J, He G, Kang M, Rong Z, Lin L, Zhong H, Huang Q, Deng A, Zeng W, Tan X, Zeng S, Zhu Z, Li J, Gong D, Wan D, Chen S, Guo L, Li Y, Sun L, Liang W, Song T, He J, Ma W. 2020 Time-varying transmission dynamics of Novel Coronavirus Pneumonia in China. *bioRxiv*. (doi: 10.1101/2020.01.25.919787).

[9] Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM. 2020 Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *The Lancet Infectious Diseases* **20**, 553–558. (doi:10.1016/S1473-3099(20)30144-4).

[10] Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. 2020 High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerging Infectious Diseases* **26**, 1470–1477. (doi:10.3201/eid2607.200282).

[11] Wearing HJ, Rohani P, Keeling MJ. 2005 Appropriate Models for the Management of Infectious Diseases. *PLOS Medicine* **2**, 0020174. (doi:10.1371/journal.pmed.0020174).

[12] Wallinga J, Lipsitch M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B* **274**, 599–604. (doi:10.1098/rspb.2006.3754).

[13] Kochańczyk, Marek, Grabowski, Frederic, Lipniacki, Tomasz. 2020 Dynamics of COVID-19 pandemic at constant and time-dependent contact rates. *Mathematical Modeling of Natural Phenomena* **15**, 28. (doi:10.1051/mmnp/2020011).

[14] Wu JT, Leung K, Leung GM. 2020a Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* **395**, 689–697. (doi:10.1016/s0140-6736(20)30260-9).

[15] Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, Cowling BJ, Lipsitch M, Leung GM. 2020b Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine* **26**, 506–510. (doi:10.1038/s41591-020-0822-7).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[16] Li Q, Guan X, Wu P, Wang X, Zhou L et al.. 2020 Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *New England Journal of Medicine* **382**, 1199–1207. (doi:10.1056/NEJMoa2001316).

[17] Nelson KE, Williams K. 2014 *Infectious Disease Epidemiology: Theory and Practice*. Jones & Bartlett Learning, Burlington, MA. 3rd edition.

[18] Endo A, Abbott S, Kucharski A, Funk S. 2020 Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 3; peer review: 2 approved]. *Wellcome Open Research* **5**, 67. (doi:10.12688/wellcomeopenres.15842.3).

[19] Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, Rambaut A, Suchard MA, Wertheim JO, Lemey P. 2020 The emergence of SARS-CoV-2 in Europe and the US. *bioRxiv*. (doi:10.1101/2020.05.21.109322).

[20] Cereda D, Tirani M, Rovida F, Demicheli V, Ajelli M, Poletti P, Trentini F, Guzzetta G, Marziano V, Barone A, Magoni M, Deandrea S, Diurno G, Lombardo M, Faccini M, Pan A, Bruno R, Pariani E, Grasselli G, Piatti A, Gramegna M, Baldanti F, Melegaro A, Merler S. 2020 The early phase of the COVID-19 outbreak in Lombardy, Italy. (arXiv:2003.09320).

[21] Mercker M, Betzin U, Wilken D. 2020 What influences COVID-19 infection rates: A statistical approach to identify promising factors applied to infection data from Germany. *medRxiv*. (doi:10.1101/2020.04.14.20064501).

[22] Harris LA, Hogg JS, Tapia JJ, Sekar JAP, Gupta S, Korsunsky I, Arora A, Barua D, Sheehan RP, Faeder JR. 2016 BioNetGen 2.2: advances in rule-based modeling. *Bioinformatics* **32**, 3366–3368. (doi:10.1093/bioinformatics/btw469).

[23] Google LLC. 2020 Google COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility. Accessed: 2020-06-26.

[24] Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, Fisman D, Murray M. 2003 Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science* **300**, 1966–1970. (doi:10.1126/science.1086616).

[25] Ferguson N, Laydon D, Nedjati Gilani G, Imai N, Ainslie K, Baguelin M. 2020 Report 9. https://spiral.imperial.ac.uk:8443/handle/10044/1/77482. Accessed: 2020-03-26.

[26] Ma S, Zhang J, Zeng M, Yun Q, Guo W, Zheng Y, Zhao S, Wang MH, Yang Z. 2020 Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries. *medRxiv*. (doi:10.1101/2020.03.21.20040329).

[27] Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, Liu X, Wei L, Truelove SA, Zhang T, Gao W, Cheng C, Tang X, Wu X, Wu Y, Sun B, Huang S, Sun Y, Zhang J, Ma T, Lessler J, Feng T. 2020 Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet Infectious Diseases* **20**, 911–919. (doi:10.1016/S1473-3099(20)30287-5).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

ELECTRONIC SUPPLEMENTARY MATERIAL

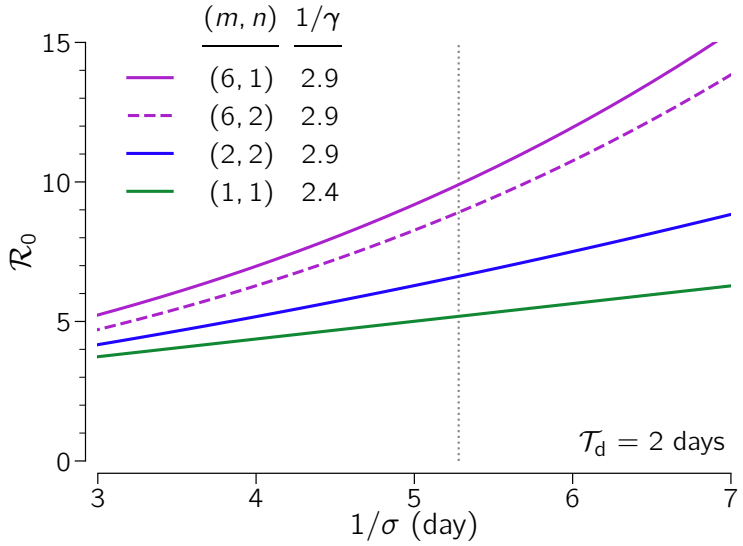

Figure S1: Basic reproduction number, $\mathcal{R}_0$, *vs.* mean latent period, $1/\sigma$. The SEIR model parameters are given in the legend; default $1/\sigma = 5.28$ days is marked with a dotted vertical line. The assumed doubling time $\mathcal{T}_d$ is 2 days.

Listing S1: BioNetGen language (BNGL)-encoded SEIR-type model of epi demic spread in the presence of super-spreaders. Parameters of the model are set so that the proportion of superspreaders in the population is set according to their infectiousness (relative w.r. to "normal" individuals) so that both groups, superspreaders and "normal" individuals, have identical total effective infectiousness. Also, contact rate is normalized so that the number of infecter "normal" individuals does not depend on the current relative infectiousness of superspreaders. The file contains both the model definition and, at the bottom, simulation protocol that runs a simulation with or without superspreaders. By changing `method=>...`, one can run either a deterministic simulation (`method=>ode`), using numerical integration of a resulting system of ODES, or an exact stochastic simulation (`method=>ssa`) of a corresponding Markov chain, performed according the Gillespie algorithm. To simulate model dynamics, please download and install BioNetGen (from http://bionetgen.org). For the sake of convenience, you may use BioNetGen within RuleBender (http://www.rulebender.org). The model was developed in BioNetGen version 2.4.0 (in hope that it will be also compatible with future BioNetGen releases).

```
begin model

  begin parameters
    super_fraction      1/(100 - 1)
    super_strength      1/super_fraction
  # ^--- To simulate hyperspreaders, multiply super_strength by 2.
```

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

```
  normal_prob        1/(1 + super_fraction)               # => 99%
  super_prob         super_fraction/(1 + super_fraction)  # =>  1%
  infected_0_normal  1  # |_ initial condition
  infected_0_super   0  # |
  z         (1 + super_fraction*super_strength)/(1 + super_fraction)

  beta    3.41478/z     # contact rate [1/day]  # => Td == 2 days
  gamma   1/2.9         # removal rate [1/day]
  tau     5.28          # average incubation period  [day]
  k       6             # no. exposed states
# ^--- This value scales rates and does not affect model structure.
end parameters

begin seed species
  E1_super()     0
  E1_normal()    0
  E2_super()     0
  E2_normal()    0
  E3_super()     0
  E3_normal()    0
  E4_super()     0
  E4_normal()    0
  E5_super()     0
  E5_normal()    0
  E6_super()     0
  E6_normal()    0
  I_super()      infected_0_super
  I_normal()     infected_0_normal
  R_super()      0
  R_normal()     0
  event_normal() 0
  event_super()  0
end seed species

begin observables
  # -- individuals in SEIR compartments
  Molecules E1_super  E1_super()
  Molecules E1_normal E1_normal()
  Molecules E2_super  E2_super()
  Molecules E2_normal E2_normal()
  Molecules E3_super  E3_super()
  Molecules E3_normal E3_normal()
  Molecules E4_super  E4_super()
  Molecules E4_normal E4_normal()
  Molecules E5_super  E5_super()
  Molecules E5_normal E5_normal()
  Molecules E6_super  E6_super()
  Molecules E6_normal E6_normal()
  Molecules I_super   I_super()
  Molecules I_normal  I_normal()
```

medRxiv preprint doi: https://doi.org/10.1101/2020.04.26.20080788; this version posted August 4, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

```
    Molecules R_super   R_super()
    Molecules R_normal  R_normal()

    # -- transmission event counters
    Molecules event_normal event_normal()
    Molecules event_super  event_super()
  end observables

  begin reaction rules
    I_normal() -> I_normal() + E1_normal() + event_normal() normal_prob*beta
    I_normal() -> I_normal() + E1_super()  + event_normal() super_prob*beta
    I_super()  -> I_super()  + E1_normal() + event_super() normal_prob*beta*super_strength
    I_super()  -> I_super()  + E1_super()  + event_super() super_prob*beta*super_strength
    E1_super()  -> E2_super()   k/tau
    E1_normal() -> E2_normal()  k/tau
    E2_super()  -> E3_super()   k/tau
    E2_normal() -> E3_normal()  k/tau
    E3_super()  -> E4_super()   k/tau
    E3_normal() -> E4_normal()  k/tau
    E4_super()  -> E5_super()   k/tau
    E4_normal() -> E5_normal()  k/tau
    E5_super()  -> E6_super()   k/tau
    E5_normal() -> E6_normal()  k/tau
    E6_super()  -> I_super()    k/tau
    E6_normal() -> I_normal()   k/tau
    I_super()   -> R_super()    gamma
    I_normal()  -> R_normal()   gamma
  end reaction rules

end model


generate_network({overwrite=>1});
simulate({method=>"ssa", suffix=>"ssa", t_end=>60, n_steps=>600, \
         stop_if=>"R_normal() + R_super() > 5000000"})
```