Accounting for super-spreading gives the basic reproduction number R_0 of COVID-19 that is 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 e-mail: tlipnia@ippt.pan.pl

Transmission of infectious diseases is characterized by the basic reproduction number R_0, a metric used to assess the threat posed by an outbreak and inform proportionate preventive decision-making. Based on individual case reports from the initial stage of the coronavirus disease 2019 epidemic, R_0 is often estimated to range between 2 and 4. In this report, we show that a SEIR model that properly accounts for the distribution of the incubation period suggests that R_0 lie in the range 4.4–11.7. This estimate is based on the doubling time observed in the near-exponential phases of the epidemic spread in China, United States, and six European countries. To support our empirical estimation, we analyze stochastic trajectories of the SEIR model showing that in the presence of super-spreaders the calculations based on individual cases reported during the initial phase of the outbreak systematically overestimate the doubling time and thus underestimate the actual value of R_0.

Introduction

The basic reproduction number R_0 is a critical parameter characterizing the dynamics of the outbreak of an infectious disease. R_0 quantifies the expected number of secondary cases generated by an infectious individual in an entirely susceptible population. Although it is pertinent to the situation before imposition of protective measures, R_0 is influenced by natural conditions (such as seasonality) as well as socioeconomic factors (such as population density or ingrained societal norms and practices) (Delamater *et al.*, 2019). Importantly, R_0 suggests the extent of control measures that have to be implemented to stop the spread of the epidemic.

A preliminary estimate published by the World Health Organization (WHO) suggested that R_0 of coronavirus disease 2019 (COVID-19) lies in between 1.4 and 2.5 (WHO, 2020a). Later this estimate has been revised to 2–2.5 (WHO, 2020b), which is in agreement with numerous other studies that, based on official data from China, implied the range of 2–3 [see, *e.g.*, Liu *et al.* (2020) or Boldog *et al.* (2020) for a summary]. This range suggests an outbreak of a contagious disease that should be containable by imposition of relatively mild restrictions on social interactions, as initially implemented

in European countries such as Italy or Spain. Of note, there were early reports of R_0 higher than 4 by Cao *et al*. (2020) and Shen *et al*. (2020), in which, however, relatively long infectious periods were assumed.

Here we estimated R_0 of COVID-19 based on the doubling time observed in the (nearly) exponential phase of the increase of confirmed cases in China, United States, and six European countries. By considering the early phase in which the number of registered cases begins to exceed 100, we aimed at accounting for super-spreading events that are implausible when the number of infected individuals is very low. To provide additional motivation for this approach, we show that in the presence of super-spreaders the median trajectory starting from a single infection grows much slower than the average trajectory, which very likely leads to overestimation of the doubling time. We used a susceptible–exposed–infected–removed (SEIR) model with six subsequent exposed states to reproduce the shape of the reported distribution of the incubation period. This allowed us to determine a plausible range of R_0 at 4.4–11.7, considerably higher than most initial estimates.

Results

The SEIR model

We used a SEIR model with six *exposed* states, which is equivalent to the assumption that the incubation period is Erlang-distributed with the shape parameter 6. We assumed that the average incubation period $1/\sigma$ ranges between 5.28 days [mean value determined by Lauer *et al*. (2020), used by us as a lower bound] and 6.47 days [mean value determined by Backer *et al*. (2020), used by us as an upper bound]. The infectious period is either exponentially or Erlang-distributed with the shape parameter 2 [as in (Kucharski *et al*., 2020)], resulting in model variants with, respectively, a single or two subsequent *infectious* states. The average infectious period $1/\gamma$ is assumed to be 3 days [as in (Liu *et al*., 2020) and (Kucharski *et al*., 2020)]. See Methods for more details on the model.

Estimation of the basic reproduction number in the exponential growth phase

We estimated T_d within two- and three-week periods starting on the day in which the number of confirmed (in the SEIR model naming convention, *removed*) cases exceeded 100 in populations of China, United States, and six European countries (Fig. 1A). In nearly all countries (except the United States), values of T_d estimated based on three-week periods are higher than based on the two-week periods, which is likely a consequence of protective measures, such as social distancing, and isolation of positively tested individuals. Three-week T_d values range from 2.34 (United States) to 3.25 (China), while two-week T_d values lie in between 2.16 (Spain) and 2.88 (United Kingdom).

We estimated the range of R_0 as a function of the doubling time T_d using a formula provided in Methods. The lower bound has been obtained using the model variant with $1/\sigma = 5.28$ days and two *infectious* states, whereas the upper bound results from the model with $1/\sigma = 6.47$ days and one *infectious* state, Fig. 1B. After plugging in the overall longest $T_d = 3.25$ and the shortest $T_d = 2.16$ into, respectively, a variant of our model with the lowest $R_0(T_d)$ curve (parameters: $m = 6$, $n = 2$, $1/\sigma = 5.28$)

and a variant with the highest $R_0(T_d)$ curve (parameters: $m = 6$, $n = 1$, $1/\sigma = 6.47$), we arrived at the estimated R_0 range of 4.4–11.7. The two-week doubling time for China, 2.39, is remarkably consistent with the value of 2.4 reported by Sanche *et al*. (2020), who estimated that R_0 for China lies in the range 4.7 to 6.6, which is contained in our estimated range for China: 4.4–9.6. The models having one or two *exposed* states, often used to estimate the value of R_0, substantially underestimated R_0, whereas a SIR model with a fixed time delay equal to $1/\sigma$ overestimated R_0, *cf*. Fig. 1B and the articles by Wearing *et al*. (2005), Wallinga & Lipsitch (2007), and Kochańczyk *et al.* (2020, Fig. 4).

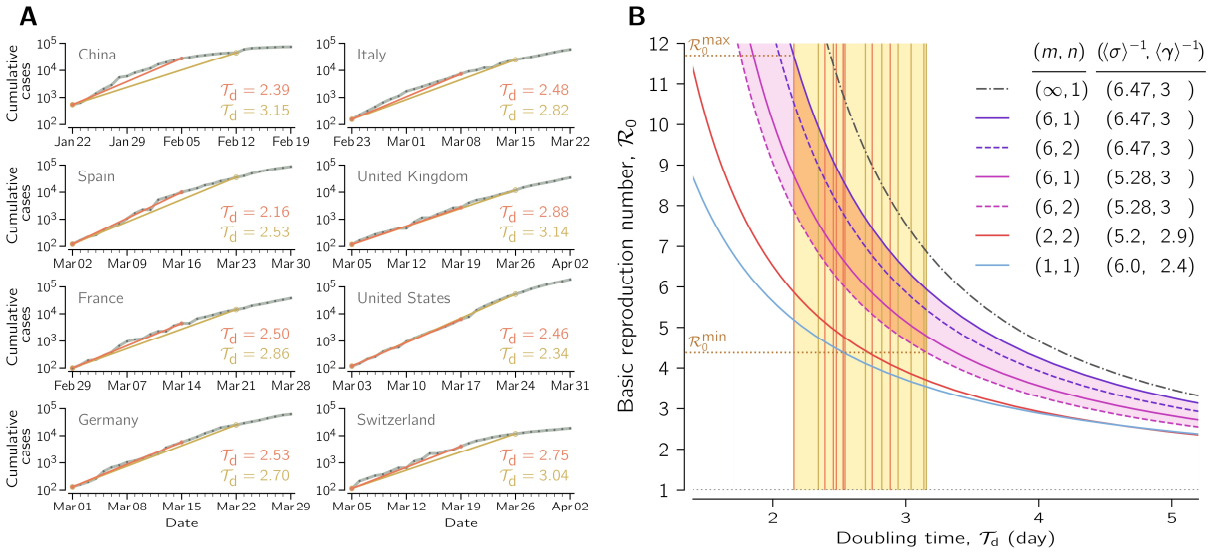

Figure 1. **Estimation of the doubling time and the resulting basic reproduction number R_0.** (**A**) Estimates of the doubling time T_d for China, United States, and six European countries using two- and three-week periods after the number of confirmed cases exceeded 100, according to data gathered and made available by Johns Hopkins University (Dong *et al.*, 2020). (**B**) The range of R_0 estimated using four variants of our SEIR model (pink and violet solid and dashed curves) for the range of T_d estimated in panel A (vertical lines in the yellow area). Red and blue solid curves correspond to $R_0(T_d)$ according to SEIR models structured and parametrized as in (Kucharski *et al*., 2020) ($n = 2$, $m = 2$) and by (Wu *et al*., 2020) ($n = 1$, $m = 1$). The highest $R_0(T_d)$ curve (black dash-dotted line) has been obtained assuming a Dirac δ-distributed incubation period and a corresponding delay-differential equations-based model formulation (see Methods).

Our calculations suggest the range of R_0 that is higher than initially estimated (Boldog *et al*., 2020; Liu *et al*., 2020). This is because (1) in contrast to most model-based calculations, we appropriately accounted for the distribution of the incubation period and (2) we estimated T_d based on the (nearly) exponential growth phases of the outbreak, which gives T_d shorter than estimates based on initial individual propagation events. The discrepancy in T_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 very early phase, in which the doubling time (growth rate) is estimated based on individual case reports, the effect of super-spreading events (such as football matches, carnival fests, demonstrations, masses, or hospital-acquired infections) is negligible due to a low probability of such events when the number of infected individuals is low. Sanche *et al.* (2020) inferred that the pre-exponential period in Wuhan has been dominated by simple transmission chains. 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 (Cereda *et al.*, 2020; Mercker *et al.*, 2020).

Stochastic simulations of the SEIR model dynamics with super-spreading

We analyzed the impact of super-spreading on estimation of T_d based on stochastic simulations of SEIR model dynamics. Simulations were performed in the perfectly mixed regime according to the Gillespie algorithm. We assumed that a predefined fixed proportion of individuals 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. We estimated 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 characterized by exponential growth, we assumed that the susceptible population remains constant.

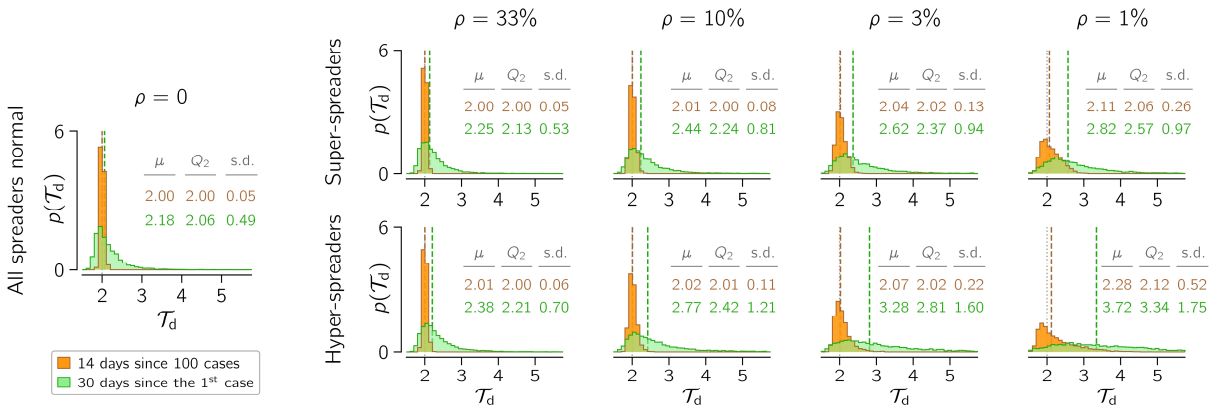

Figure 2. Estimation of the doubling time T_d based on stochastic simulations of the SEIR model with super- and hyper-spreaders. Histograms show probability density $p(T_d)$ estimated using the '14 days since 100 cases' method (orange) and the '30 days since the 1st case' method (green). In each column, 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 left-most panel with $\rho = 0$, to right), their infectiousness has been reduced to give the same deterministic $T_d = 2$ days (vertical dotted gray lines). Remaining model parameters: $(m, n) = (6, 1)$; $(1/\sigma, 1/\gamma) = (5.28$ days, 3 days). Each histogram results from 5,000 stochastic simulations, each starting from a single infected normal individual; trajectories resulting in outbreak failure were discarded. Each distribution is described in terms of its mean (μ), median (Q_2 and vertical dashed lines), and standard deviation (s.d.).

In Fig. 2 we show histograms of 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 T_d is calculated using the '14 days since 100 cases' method, its median value is slightly larger than T_d in the deterministic model (equal 2 days); however, when 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 T_d. In 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 T_d that is overestimated by 28% (67%). Both methods statistically overestimate T_d but when using the '30 days since the 1st case' method, overestimation is very significant. The effect is caused by low probability of appearance of super- or ultra-spreaders in the first weeks of the outbreak. We notice that T_d estimation for a given country based on available data is equivalent to the analysis of a single stochastic trajectory.

Conclusions

Based on epidemic data from China, United States, and six European countries, we have estimated that the basic reproduction number R_0 lies in the range 4.6–11.6 (4.4–9.6 for China), which is higher than most previous estimates (Boldog *et al.*, 2020; Liu *et al.*, 2020). There are two sources of the discrepancy in R_0 estimation. First, in agreement with data on the incubation period distribution, we used a model with six *exposed* states, which substantially increases $R_0(T_d)$ with respect to the models with one or two *exposed* states. Second, we estimated T_d based on the two- and three-week period exponential growth phase beginning on the day in which the number of registered cases exceeds 100. This approach, in contrast to estimation of R_0 based on individual case reports, allows to implicitly take into account super-spreading events that substantially shorten 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 25, 2020, these countries managed to terminate the exponential growth phase by means of country-wide quarantine. Current COVID-19 Community Mobility Reports (Google, 2020) 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 fivefold; additionally, massive testing reduced the infectious period, 1/γ. Consequently, we expect that the reproduction number $R = \beta/\gamma$ was reduced more than fivefold, which brought it to the values somewhat smaller than 1. This suggests that R_0 in these countries could have been larger than 5.

Methods

The SEIR model and its parametrization

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

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta\, I(t)\, S(t)/N,$$

$$\frac{\mathrm{d}E_1}{\mathrm{d}t} = \beta\, I(t)\, S(t)/N - m\sigma E_1(t),$$

$$\frac{\mathrm{d}E_i}{\mathrm{d}t} = m\sigma E_{i-1}(t) - m\sigma E_i(t), \quad 2 \le i \le m,$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = m\sigma E_m(t) - n\gamma I_1(t),$$

$$\frac{\mathrm{d}I_j}{\mathrm{d}t} = n\gamma I_{j-1}(t) - n\gamma I_j(t), \quad 2 \le j \le n,$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = n\gamma I_n(t),$$

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* states and n is the number of *infectious* states, there are $m + n + 2$ equations in the system.

To obtain $R_0(T_d)$ for Dirac δ-distributed latency (incubation) period, we use a SIR model described by the following system of delay-differential equations corresponding to the case of $m \to \infty$ and $n = 1$:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta\, I(t - 1/\sigma)\, S(t)/M,$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta\, I(t - 1/\sigma)\, S(t)/M - \gamma\, I(t - 1/\sigma),$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma\, I(t - 1/\sigma),$$

with $M = S(t) + I(t) + R(t)$ and $1/\sigma$ being the delay equal to the mean latency period.

We assume that the latency period is the same as the incubation period, estimates of which are available in the literature. In model variants with shorter incubation period, we assume that it is Erlang-distributed with the shape parameter 6 and the scale parameter equal 0.88, giving $1/\sigma = 5.28$, following exactly the estimates made by Lauer *et al*. (2020). In model variants with longer incubation period, we replaced the Γ distribution of Backer *et al*. (2020), having the shape parameter 6.1 and the scale parameter 1.06, resulting in $1/\sigma = 6.47$, with the Erlang distribution with the shape parameter 6 and the scale parameter 1.078, having nearly identical shape and the same $1/\sigma$ of 6.47.

Dependence of the basic reproduction number on the doubling time, $R_0(T_d)$

We determined $R_0(T_d)$ as:

$$R_0(T_\mathrm{d}) = \frac{\frac{\log 2}{T_\mathrm{d}}\left(\frac{\log 2}{T_\mathrm{d}\, m\, \sigma} + 1\right)^m}{\gamma\left(1 - \left(\frac{\log 2}{T_\mathrm{d}\, n\, \gamma} + 1\right)^{-n}\right)}$$

in accordance with Wallinga & Lipsitch (2007) and Wearing *et al*. (2005).

References

Backer, J.A., Klinkenberg, D., and Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. *Euro Surveill* **25**, 2000062.

Boldog, P., Tekeli, T., Vizi, Z., Dénes, A., Bartha, A.F., and Röst, G. (2020). Risk Assessment of Novel Coronavirus COVID-19 Outbreaks Outside China. *J Clin Med* **9**, 571.

Cao, Z., Zhang, Q., Lu, X., Pfeiffer, D., Jia, Z., Song, H., and Zeng, D.D. (2020). Estimating the effective reproduction number of the 2019-nCoV in China. MedRxiv 2020.01.27.20018952.

Cereda, D., Tirani, M., Rovida, F., Demicheli, V., Ajelli, M., *et al*. (2020). The early phase of the COVID-19 outbreak in Lombardy, Italy. ArXiv 2003.09320.

Delamater, P.L., Street, E.J., Leslie, T.F., Yang, Y.T., and Jacobsen, K.H. (2019). Complexity of the Basic Reproduction Number (R0). *Emerging Infect Dis* **25**, 1–4.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infect Dis S1473-3099*, 30120–30121.

Google (2020). COVID-19 Community Mobility Report. https://www.google.com/covid19/mobility

Kochańczyk, M., Grabowski, F., and Lipniacki, T. (2020). Dynamics of COVID-19 pandemic at constant and time-dependent contact rates. *Math Model Nat Phenom* **28**, 12.

Kucharski, A.J., Russell, T.W., Diamond, C., Liu, Y., Edmunds, J., *et al*. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect Dis S1473-3099*, 30144–4.

Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., *et al*. (2020). The incubation period of 2019-nCoV from publicly reported confirmed cases: estimation and application. MedRxiv 2020.02.02.20020016.

Liu, Y., Gayle, A.A., Wilder-Smith, A., and Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med* **27**, taaa021.

Mercker, M., Betzin, U., and Wilken, D. (2020). What influences COVID-19 infection rates: A statistical approach to identify promising factors applied to infection data from Germany. MedRxiv 2020.04.14.20064501.

Sanche, S., Lin, Y.-T., Xu, C., Romero-Severson, E., Hengartner, N., and Ke, R. (2020). The Novel Coronavirus, 2019-nCoV, is Highly Contagious and More Infectious Than Initially Estimated. MedRxiv 2020.02.07.20021154.

Shen, M., Peng, Z., Xiao, Y., and Zhang, L. (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. BioRxiv 2020.01.23.916726.

Wallinga, J., and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc Biol Sci* **274**, 599–604.

Wearing, H.J., Rohani, P., and Keeling, M.J. (2005). Appropriate models for the management of infectious diseases. *PLoS Med* **2**, e174.

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)

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)

Wu, J.T., Leung, K., Bushman, M., Kishore, N., Niehus, R., *et al*. (2020). Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat Med* **26**, 506–510.