medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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 4.0 International license .

On the true numbers of COVID-19 infections: behind the available data

M.E. QADMIRY[1], E.H. TAHRI[2], and Y. HASSOUNI[3]

[1,2]Laboratory of Matter and Radiations Physics, Faculty of Sciences, University of Mohamed I, Av. Mohamed VI, BP 717, Oujda, Morocco.
[3]ESMaR, Faculty of Sciences, Mohammed V University, Av. Ibn Battouta, B.P. 1014, Agdal, Rabat, Morocco

May 26, 2020

In December-2019 China reported several cases of a novel coronavirus later called COVID-19. In this work, we will use a probabilistic method for approximating the true daily numbers of infected. Based on two distribution functions to describe the spontaneous recovered cases on the one hand and the detected cases on the other hand. The impact of the underlying variables of these functions is discussed. The detected rate is predicted to be between 5.3% and 12%, which means that there would be about 68 million infected until now (25-May 2020), rather than the officially declared number of 5.37 million worldwide cases.

1 Introduction

Since the outbreak of the novel coronavirus COVID-19 and its spread around the world, several published works were concerned with the analysis of the available data to extract information that characterizes the daily evolution of this new pandemic. Among the concerned information, there are the incubation period[1,2], the reproduction number[3], mortality rate[4], and the asymptomatic proportion as in [5], where was used the data of the Diamond Princess cruise ship; which means that the study is done on a closed population, a fact that is worthy to mention.

Our aim in this work is to approach the true number of infected cases, and to develop an analytical method in order to allow the simulation of different scenarios that can occur if we modify the underlying variables of two special

[1] medkadmiri@hotmail.fr
[2] hassanfa@yahoo.com
[3] y-hassou@fsr.ac.ma

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

probabilistic functions. These probabilistic functions control the numbers of the infected as well as the numbers of the detected cases. To predict the evolution of the true number of cases; which escapes from the official counting, we take into account the reproduction number.

The importance of this investigation lies in the fact that the important part of the infection comes from people who are not in hospitals or not isolated, i.e., that the virus transmission occurs in general before symptoms appear [6].

Then, we think that the numerical simulation achieved here can be helpful to estimate the consequences of different strategies adopted at the governmental level, such as social distancing, strict-confinement for a short period, mass testing, and other measures that aim to stop the evolution of the pandemic.

2 Method

In this section, we will introduce the distribution functions that describe the spontaneous outcome cases and cases detected. In addition to the reproduction number responsible for exponential growth.

2.1 Spontaneous outcome cases

When there is no medical intervention in favor of the infected, the biological defending system has to withstand alone as a natural solution. Therefore, eventually the cases end by either recovering disability or death. Thus, for a closed population initially infected, the number of infected keeps steadily decreasing since the illness onset until the end of infection, concerning many factors including age, health status, viral load, immunity, etc.

Considering these facts, we will suggest a suitable distribution function to lower the amount of the infected population through time. It is denoted as $P_s(t)$ -(s)-subscript as survival- and its analytical expression is chosen so as to be compatible with the empirical observation taken from the data available in [7] based on 1987 control patients, so it is given as follows:

$$P_s(t) = e^{-\frac{t}{7-\frac{t}{8}}}\theta(30-t), \qquad (1)$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

where $\theta(t)$ is the Heaviside step function equals to 1 if $t \leq 30$ and to 0 otherwise, it ensures here that the function vanishes if the time exceeds the period of 30-days since the illness onset. This is because there is no remaining infected after this period, which after all, we will see that it does not have a significant impact on the results. Let us note also that we do not use the polynomial regression to describe $P_s(t)$, because it demands a greet polynomial degree.

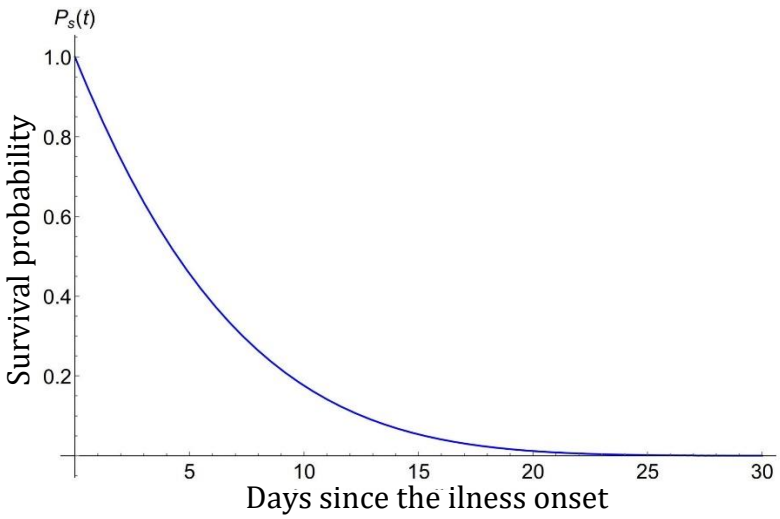

Figure 1: The behavior of $P_s(t)$ function

In Fig: (1), we show the behavior of the $P_s(t)$ function. The real curve can be extracted from the data concerning the control patients under the drug treatments experiment as for example in [7,8], here we use the data of the first one as we mention above. The comparison is shown in Fig: (2), where we take the initial number of infected to be 1987. Thus, we plot the expected survival cases -at every day- given by the product 1987 × $P_s(t)$, with the real survival cases following the data. The standard error equals to 19, which means that we have a good approximation.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

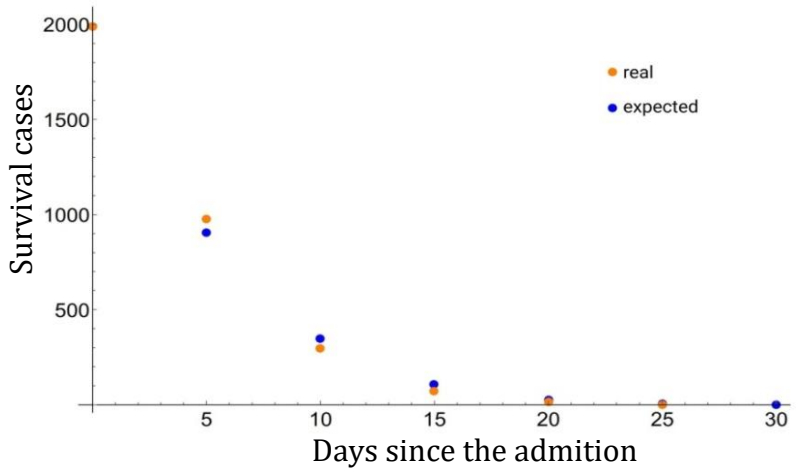

Figure 2: Real (orange) and expected (blue) survival cases using the available data and the $P_s(t)$ function, respectively

2.2 Detected cases

There is no doubt that the cases counting is not perfect, because there are many factors that may affect the results, such as test accuracy; viral load; medical protocol; testing strategy... Nevertheless, we can estimate another cumulative distribution function (CDF), which describes the probability to detect the infected cases among the entire population. This function is well-known used for several precedent epidemics (see [9]), e.g., SARS coronavirus; Rhinovirus; Influenza B...

The logistic distribution is useful to recognize this CDF, it is characterized by two parameters: the first one, denoted by M, represents the maximum absolute accuracy of testing and the second one is denoted by δ_0 and represents the incubation period of COVID-19, which is estimated to be between 5.5 and 9.5 days (see [5]), e.g. it has fixed to 5.8 days in [10] and 5.1 days in [2]. Let us denote this distribution function by $P_d(t)$ -(d)-subscript as detected, and "t" as time representing days since infection- and write its explicit expression in the following manner:

$$P_d(t) = \frac{M}{1+e^{-(t-\delta_0)}} \tag{2}$$

In what follows, we will choose the value of the incubation period as δ_0 = 6.4-days following the result of the reference [11] (where it was chosen for the accuracy of the data used in the analysis). Concerning the testing

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

accuracy, we take the maximum equal to M = 90%, which is a widely varied quantity regarding the testing strategy and the asymptomatic proportion estimated in [5] to be 17.9%.

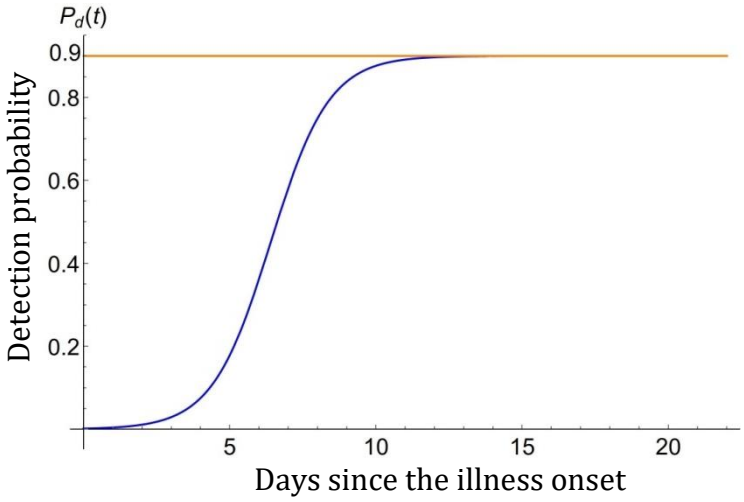

Figure 3: Graphic illustrating the behavior of the $P_d(t)$ CDF by considering M=90% and δ_0 = 6.4 days

In Fig:(3) we show the evolution of $P_d(t)$ CDF with respect to the above conditions, from which we get the probability to detect 2.9% after 2 days; 58% after 7 days; 90% after 12 days.

2.3 Evolution of reproduction number

The reproduction number is an important index to evaluate the spread of an epidemic. It is defined as the number of infected caused by an infectious person during the infectious period, denoted by R_0.

What interests us in this work is the infection-producing contacts per one day, which is described by a function of time denoted by $r(t)$, obtained simply as the ratio of the reproduction number R_0 to the infectious period τ (neglecting the latent period). This quantity is sensitive to the confinement measures and always lower with respect to time. We suppose that $r(t)$ follows an exponential evolution, which we could justify by the fact that at early times it is much easier to reduce the infected-healthy contacts than at a late time. Therefore, we can write it as:

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

$$r(t) = r_0 e^{-\frac{t}{d}}, \tag{3}$$

where $r_0 = R_0/\tau$ is the initial value calculated from the value of the reproduction number R_0, which itself is estimated according to several considerations as 3.28-days in [12] or 6.47-days in [13]. Here, we will use the last one as a precaution to prevent bad decisions, thus $r_0 = 0.32$. The parameter d is the number that characterizes the isolation strategy, which is -by definition- the number of days required to reduce the initial value r_0 by $1/e \simeq 0.37$. In our case, we must choose it as $d = 341$-days to get the simulation of detected cases compatible with the worldwide data of confirmed cases.

3 Results

3.1 Semi-closed infected population

Let us first start with a simulation to see what happens if we have a semi-closed infected population, where it is only possible to lose the infected without any gain. Therefore, let us fix the initial population to be $N_0 = 100$ infected and extract every detected case following the $P_d(t)$ CDF. To do so, every time $t_n = n$-days, we recalculate the numbers of infected and detected cases, where $t_0 = 0$ is the moment of illness onset for the entire population. The number of infected $N(t)$ at t_n is then given by:

$$N(t_n) = P(t_n) N_0, \tag{4}$$

where we introduced a new distribution function $P(t)$ written as follows:

$$P(t_n) \coloneqq P_s(t_n) \prod_{i=1}^{n} \left(1 - P_d(t_{i-1})\right). \tag{5}$$

The number of detected cases $\tilde{N}(t)$ among the infected cases is given by:

$$\tilde{N}(t_n) = \tilde{P}(t_n) N_0, \tag{6}$$

where $\tilde{P}(t)$ is a CDF function defined by:

$$\tilde{P}(t_n) \coloneqq P(t_n) P_d(t_n). \tag{7}$$

The behavior of these two numbers N and $\tilde{N}$ with respect to time -since the illness onset- is given in Fig: (4). It is noteworthy that the detected cases

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

reach its maximum value near the incubation period and decrease afterward. The accumulation of detected cases is less than the initial number of infected N_0 by a rate of 63% (the majority of infected are not detected).

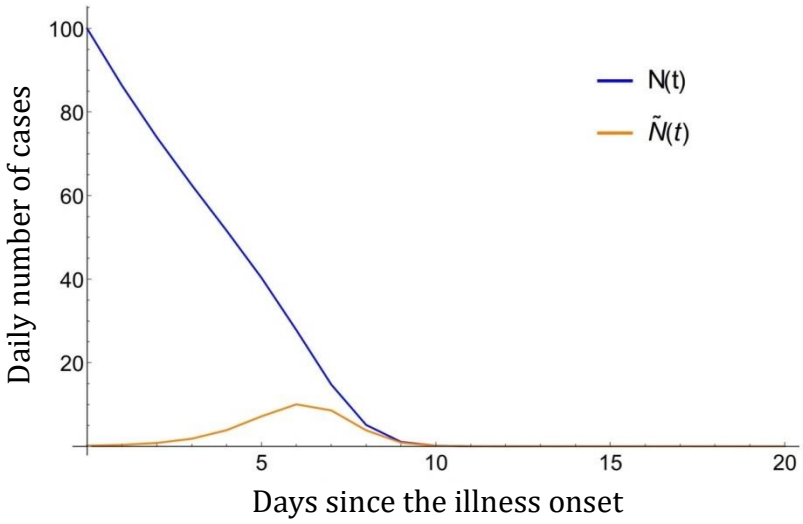

Figure 4: The graphs show the behavior of the number of infected cases $N(t)$ (blue) and detected cases $\tilde{N}(t)$ (orange) in function of time

In the following section, we shall consider each "new cases" as a semi-closed population. Remark that after only 10-days there is no infected stay free, they are either isolated or recovered. For the last reason, we know that the period from the illness onset to death does not play a significant role in the detection mechanism.

3.2 Open population with an initial infected

Now, to apply the method to the worldwide status, the infected population must be considered as an open population where it is possible to gain and lose the infected. In essence, every day we must recalculate its remain cases as the following:

$$\textit{Remain cases} = \textit{New cases} + \textit{Not recovered spontaneously} - \textit{Detected},$$

where we suppose that the new cases are given through the exponential growth [14]. Explicitly, if we suppose that the latent period [15] is zero so as

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

to be able to multiply by $r(t)$ since the first day of infection, we can then write it as:

$$New\ cases(t_n) = r(t_{n-1}) * (Remains\ cases(t_{n-1})).$$

In general, we can write the true number of infected cases by a recurrence relation as:

$$N(t_n) = r(t_{n-1})N(t_{n-1}) + \sum_{i=2}^{31} P(t_{i-1})r(t_{n-i})N(t_{n-i}) + P(t_n)N_{0,} \quad (8)$$

where $N(t_{n-i}) = 0$ if $n < i$. The first term represents the new cases, the second term (formed by the sum of 30-terms) represents the cases that have not recovered spontaneously from the previously generated cases, and the last term represents the cases that have not recovered spontaneously from the initial number of infected (which of course is equal to zero if t_n>30-days).

Likewise, the number of detected cases (confirmed) can be expressed as a recurrence relation as follows:

$$\tilde{N}(t_n) = \sum_{i=2}^{31} \tilde{P}(t_{i-1})r(t_{n-i})N(t_{n-i}) + \tilde{P}(t_n)N_0, \quad (9)$$

where the first terms under the summation symbol represent the detected cases from the previously generated cases and the second term represents the detected cases from the initial number of infected (which is equal to zero if t_n>30-days).

Thus, using the above equations (8) and (9) by taking an initial number of infected cases as $N_0 = 270$. Note that there were 27 cases at 15-Dec 2019 according to [16], and a detected rate claimed to be 10%. Thus, we can realize the simulation showing the evolution of the true infected cases and that of the detected one. The result is given by the two curves in Fig: (5)

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

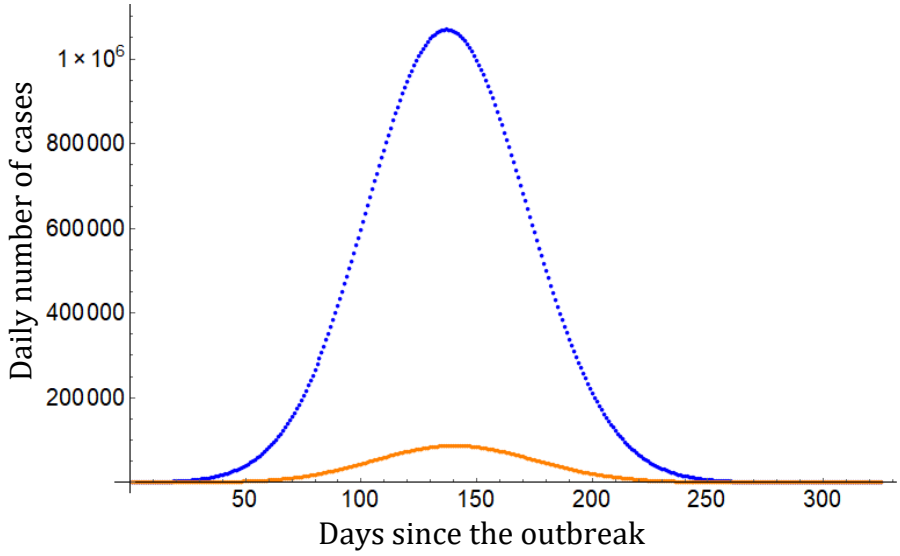

Figure 5: The evolution of the true infected cases $N(t)$ (blue) and detected cases $\tilde{N}(t)$ (orange) in function of time

By looking at the two curves, we observe a great difference between them: the confirmed cases appear very smaller than the true infected cases. With a detection rate increasing across time (Fig: (6)). Beginning with an average of 5.3% in the first 50-days and ending by an average of 12% in the final 50-days before the disappearance of the pandemic. These results are well-matched with the results of the studies achieved in [17], [18]. In Fig: (6) we can see the daily evolution of the detection rate

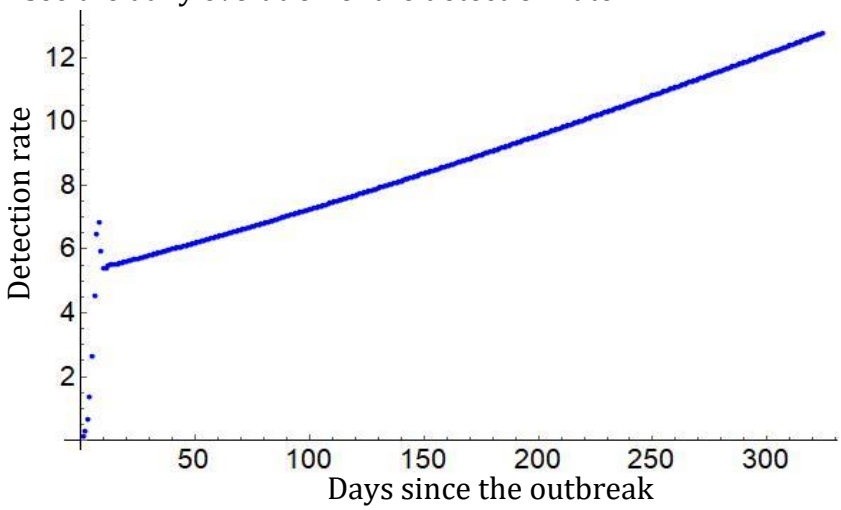

Figure 6: Detection rate evolution during the period of the pandemic

Finally, as we cited before, the curve of detected cases should be compatible with the confirmed cases as it is known from the available

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

worldwide data [19]. This is what we do In Fig: (7), indeed, we compare the theoretical curve (dashed) with the real curve (continued) up to the present time.

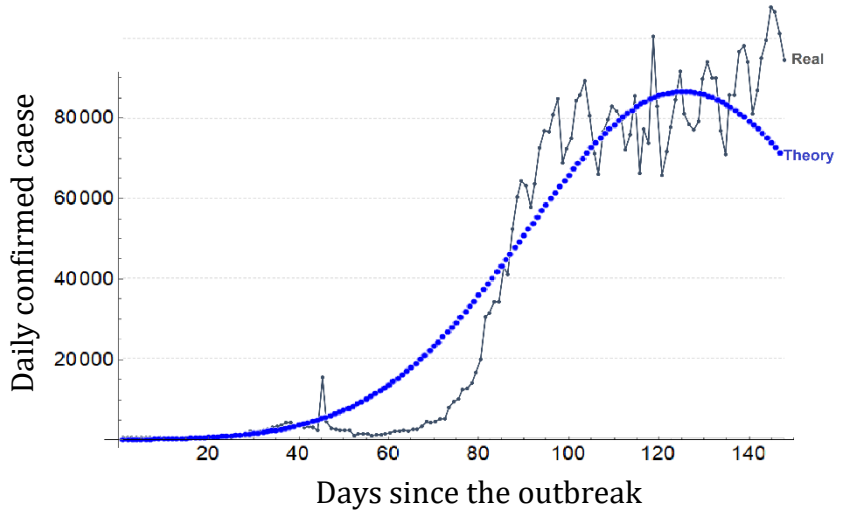

Figure 7: The graphs represent real (continued) and theoretical (dashed) curves of daily-confirmed cases.

Furthermore, we can deduce a numerical comparison between the simulated and the real accumulated -confirmed- cases $\sum \tilde{N}(t)$ which are 5.32M and 5.37M respectively (up to 15-May 2020 [19] In the other side, the true accumulated cases $\sum N(t)$ give about 68.6M cases rather than 5.37M.

4 Conclusion

Across this work, we developed a new method to simulate the true evolution of the COVID-19 pandemic. This method consists of three concepts modeled by three analytic functions: the spontaneous outcome, the detected cases, and the well-known exponential growth. The results show in particular that the simulated curve of the true infected cases gives a large number in comparison with the confirmed one. The method is presented in a way that can be extended for use to other pandemics and/or to a given country. Furthermore, this simulation may help in predicting the impact of various measures taken to tackle the epidemic.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

References

[1] Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S. M., ... Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. Journal of clinical medicine, 9(2), 538. doi:http://dx.doi.org/10.3390/jcm9020538

[2] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., ... &Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med. doi:10.7326/M20-0504

[3] Liu, Y., Gayle, A. A., Wilder-Smith, A., &Rockl¨ov, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of travel medicine, 27(2). doi: https://doi.org/10.1093/jtm/taaa021

[4] Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., & Favre, G. (2020). Real estimates of mortality following COVID-19 infection. The Lancet infectious diseases. doi: https://doi.org/10.1016/ S1473-3099(20)30195-X

[5] Mizumoto, K., Kagaya, K., Zarebski, A., &Chowell, G. (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance, 25(10), 2000180. doi: https://doi.org/10.2807/1560-7917. ES.2020.25.10.2000180

[6] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., ... &Flasche, S. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health, 8(4), e488e496. doi: https://doi.org/10.1016/S2214-109X(20)30074-7

[7] Paranjpe, I., Fuster, V., Lala, A., Russak, A., Glicksberg, B. S., Levin, M. A., Charney, A. W., Narula, J., Fayad, Z. A., Bagiella, E., Zhao, S., & Nadkarni, G. N. (2020). Association of Treatment Dose Anticoagulation with In-Hospital Survival Among Hospitalized Patients with COVID-19. Journal of the American College of Cardiology. doi:https://dx.doi.org/10.1016% 2Fj.jacc.2020.05.001

[8] Gautret, P., Lagier, J. C., Parola, P., Meddeb, L., Mailhe, M., Doudier, B., ... &Honor´e, S. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. International journal of antimicrobial agents, 105949. doi: https://doi.org/10.1016/j.ijantimicag.2020.105949

[9] Lessler, J., Reich, N. G., Brookmeyer, R., Perl, T. M., Nelson, K. E., & Cummings, D. A. (2009). Incubation periods of acute respiratory viral infections: a systematic review. The Lancet infectious diseases, 9(5), 291-300. doi: https://doi.org/10.1016/S1473-3099(09)70069-6

medRxiv preprint doi: https://doi.org/10.1101/2020.05.26.20114074; this version posted May 28, 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 4.0 International license .

[10] Backer, J. A., Klinkenberg, D., &Wallinga, J. (2020). The incubation period of 2019-nCoV infections among travellers from Wuhan, China. published online. doi: http://dx.doi.org/10.1101/2020.01.27.20018986

[11] Backer, J. A., Klinkenberg, D., &Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. Eurosurveillance, 25(5), 2000062. doi: https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062

[12] Liu, Y., Gayle, A. A., Wilder-Smith, A., &Rockl¨ov, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of travel medicine, 27(2). doi: https://doi.org/10.1093/jtm/taaa021

[13] Tang, B., Wang, X., Li, Q., Bragazzi, N. L., Tang, S., Xiao, Y., & Wu, J. (2020). Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. Journal of Clinical Medicine, 9(2), 462. doi: https://doi.org/10.3390/jcm9020462

[14] Shim, E., Tariq, A., Choi, W., Lee, Y., &Chowell, G. (2020). Transmission potential and severity of COVID-19 in South Korea. International Journal of Infectious Diseases, 93, 339-344. doi:https://doi.org/10.1016/j.ijid. 2020.03.031

[15] Latent Period Definition: https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2g-communicable-disease/definitions-disease-control

[16] Website of Livescience: https://www.livescience.com/ first-case-coronavirus-found.html

[17] Bommer, C., & Vollmer, S. (2020). AverageThe following experimental design was used: ... detection rate of SARS-CoV-2 infections is estimated around six percent, https://www.uni-goettingen.de/en/606540.html

[18] Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science, 368(6490), 489-493. doi: https://doi.org/10.1126/science.abb3221

[19] website: Our world in data https://ourworldindata.org/grapher/ daily-cases-covid-19