 doi:10.20944/preprints202004.0052.v1

Article

Estimation of Unreported Novel Coronavirus (SARS-CoV-2) Infections from Reported Deaths: a Susceptible Exposed Infectious Recovered Dead Model

Andrea Maugeri [1], Martina Barchitta [1], Sebastiano Battiato [2] and Antonella Agodi [1,*]

[1] Department of Medical and Surgical Sciences and Advanced Technologies "GF Ingrassia", University of Catania, Catania, Italy
[2] Department of Mathematics and Computer Science, University of Catania, Catania, Italy
* Correspondence: agodia@unict.it

Abstract: In the midst of the novel coronavirus (SARS-CoV-2) epidemic, examining reported case data could lead to biased speculations and conclusions. Indeed, estimation of unreported infections is crucial for a better understanding of the current emergency in China and in other countries. In this study, we aimed to estimate the unreported number of infections in China prior to 23 March 2020 restrictions. To do that, we developed a Susceptible-Exposed-Infectious-Recovered-Dead (SEIRD) model which estimated unreported cases and infections from the reported number of deaths. Our approach relied on the fact that observed deaths were less likely to be affected by reporting biases than reported infections. Interestingly, we estimated that R_0 was 2.43 (95%CI= 2.42 – 2.44) at the beginning of the epidemic, and that 92.9% (95%CI= 92.5% - 93.1%) of total cases were not reported. Similarly, the proportion of unreported new infections by day ranged from 52.1% to 100%, with a total of 91.8% (95%CI= 91.6% - 92.1%) unreported infections. Agreement between our estimates and those from previous studies proved that our approach was reliable to estimate prevalence and incidence of undocumented SARS-CoV2 infections. Once tested on Chinese data, our model could be applied on other countries with different surveillance and testing policies.

Keywords: Novel Coronavirus, COVID-19, Epidemic Model, Epidemiology

1. Introduction

The novel coronavirus (SARS-CoV-2) outbreak, which spread in Wuhan (Hubei Province, China) at the end of 2019, caused 81,554 cases and 3,312 deaths among the Chinese population as of 1 April 2020 [1]. Whilst China is winning its own battle against SARS-CoV-2, other countries are still facing the epidemic and global efforts to contain the virus are still ongoing. However, given the uncertainty on transmissibility and virulence of SARS-CoV-2, the effectiveness of strategies against the current epidemic remains quite unknown. In this scenario, the proportion of unreported infections is particularly noteworthy, since its crucial role in modulating the spread of the virus. Indeed, unrecognized cases – often patients who experience mild or no symptoms – could silently expose a far greater proportion of the population to the SARS-CoV-2 [2]. In line, it has recently been estimated that the transmission rate of undocumented infections was about half of those documented, and that they could be the source for eight out of ten documented cases [3]. Several countries, however, are implementing stringent testing strategies of severely ill patients or those who came into contact with documented cases [4]. This could lead to loosing a fraction of mild or asymptomatic patients which, however, could be infectious. Therefore, looking only at reported case data, biased speculations and hasty conclusions cannot be completely excluded. In contrast, observed deaths are less likely to be affected by reporting biases, with the exception of deaths in the early phase of the epidemic [5].

For these reasons, we hypothesized to estimate the unreported number of infections working directly on reported deaths. We employed a Susceptible-Exposed-Infectious-Recovered-Dead

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0052.v1

(SEIRD) model to estimate the number of unreported infections and cases of SARS-CoV-2 in China prior to 23 January 2020, the date on which China imposed a lockdown in Wuhan and other cities of Hubei province in an effort to quarantine the epicenter of the SARS-CoV-2 outbreak.

2. Materials and Methods

We used available public data on the daily number of cases and deaths in China released by the European Centre for Disease Prevention and Control [6]. All cases were laboratory confirmed following the case definition by the National Health Commission of China [6]. In line with previous studies [7-11], a susceptible-exposed-infectious-removed (SEIR) model has been exploited but having care of separating the removed state into two classes: recovered cases (R) and deaths (D). Indeed, in the traditional SEIR model, the removed state ideally includes both recovered and dead patients. In our study, instead, we aimed to estimate the number of deaths through the SEIRD model, and to fit the model itself to the reported number of deaths. A visual summary of our model is displayed in **Figure 1**.

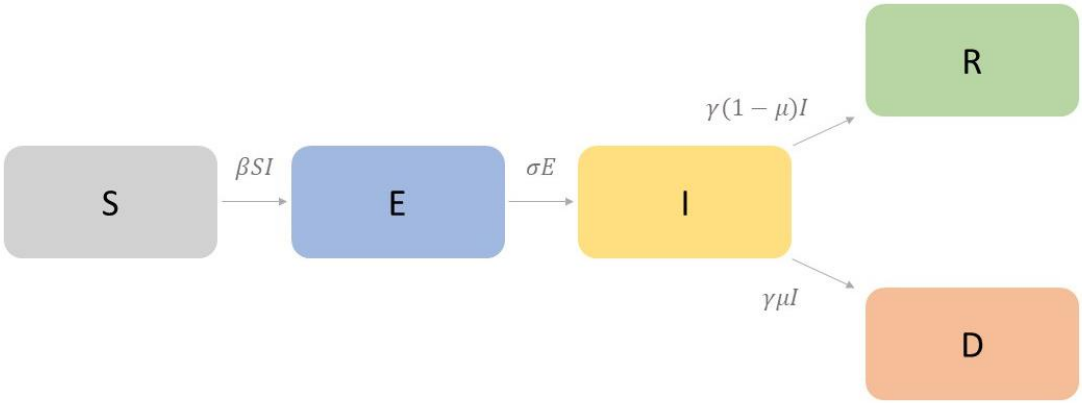

Figure 1. The employed SEIRD epidemic model for SARS-CoV-2.

In particular, the model was defined by the following ordinary differential equations:

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

$$\frac{dE(t)}{dt} = \frac{\beta S(t) I(t)}{N} - \sigma E(t)$$

$$\frac{dI(t)}{dt} = \sigma E(t) - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma(1-\mu) I(t)$$

$$\frac{dD(t)}{dt} = \gamma\mu I(t)$$

Where:

- *S(t)*, *E(t)*, *I(t)*, *R(t)*, and *D(t)* were the number of susceptible, exposed (infected but not yet be infectious), infectious, recovered and dead individuals at time (t);
- *N* was the total population as *N= S+E+I+R+D*;
- β was the transmission rate also known as the effective contact rate;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0052.v1

- σ was the infection rate and was assumed to be the inverse of the incubation period (i.e. the period from infection to the onset of symptoms);
- γ was the removing rate and was assumed to be the inverse of the period between the onset of symptoms and recovering/death;
- μ was the probability of dying among infectious individuals.

Figure 2 depicts, as an example, the number of individuals in each state since the first infection occurred in a population of 10,000 individuals. Estimates were obtained through the SEIRD model with β, σ, γ, and μ set as 0.8, 0.3, 0.2, and 0.2, respectively.

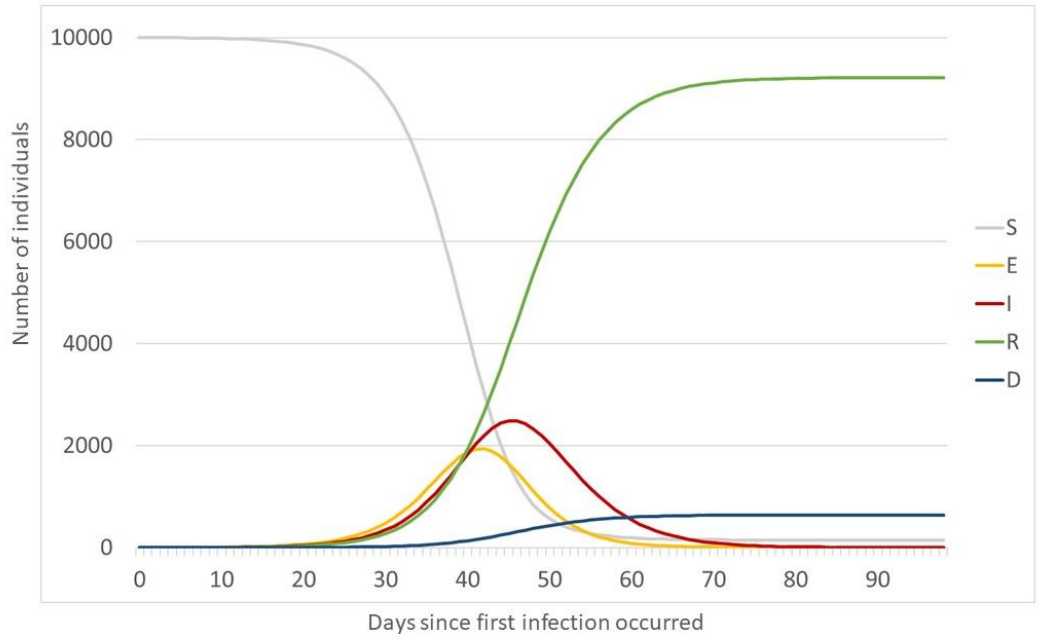

Figure 2. Representation of SEIRD states along the temporal axis. Estimates were obtained through a SEIRD model with β, σ, γ, and μ set as 0.8, 0.3, 0.2, and 0.2, respectively.

In the current study, N was assumed to be 1 billion, R and D were initially set as 0, with the initial assumed number of infectious individuals was set to 1. In the early phase of the epidemic, it was not possible to completely exclude a small fraction of undocumented deaths. Moreover, given the lag of 2-3 weeks between transmission changes and their impact on mortality trend, we were very confident in using data within 2 weeks after the travel restrictions. For these reasons, we fitted our model to the reported number of deaths from 23 January (i.e. the day after China has cumulatively observed 10 deaths) to 7 February. In the baseline scenario, we assumed σ and γ as 1/5.2 days and 1/3.5 days according to previous studies [2,3]. The initial ranges of the unknown model parameters were $0.1 \leq \beta \leq 1$ and $0.001 \leq \mu \leq 0.200$, respectively.

To estimate unknown parameters with their 95% confidence interval (95%CI), which best explained the reported numbers of deaths, we applied a least squares optimization using an evolutionary algorithm (population size=1x105, convergence=1x10-6, and mutation rate=5x10-2) and simulations (n=1,000) on randomly generated samples from the cumulative distribution function of reported deaths. Estimated infections and total cases from 31 December to 23 January were obtained from the best-fitting SEIRD model. Values of unreported new infections and total cases were obtained by subtracting the reported numbers from those estimated, and then reported as percentage. The basic reproductive number (R_0) was calculated from the SEIRD model as previously described [12].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0052.v1

We also performed sensitivity analyses to evaluate the impact of varying the infectious period on the estimation of unreported cases and infections.

3. Results

The cumulative number of cases and deaths by day of report from 31 December 2019 to 7 February 2020 are shown in **Figure 3**. Looking at the case fatality risk (i.e., the number of deaths in persons who tested positive for SARS-CoV-2 divided by number of SARS-CoV-2 cases), we noted high fluctuations that could be attributed to the proportion of unreported cases or deaths. However, as previously discussed, observed deaths were less prone to be affected by reporting biases than documented cases.

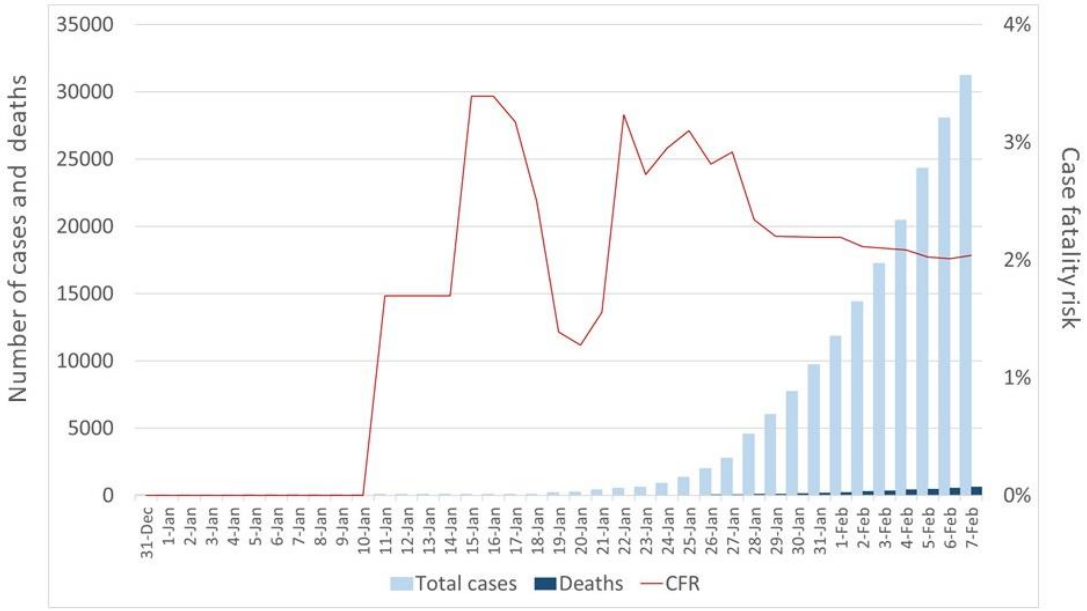

Figure 3. Number of reported cases and deaths in China from 31 December 2019 to 7 February 2020. The bars represent the cumulative number of reported SARS-CoV-2 cases and related deaths while the red line represents the case fatality risk.

Accordingly, we first fitted our SEIRD model to reported deaths, as shown in **Figure 4**, which suggested an overall good fit between estimated and reported deaths (Correlation Coefficient R^2 = 0.987). The slight overprediction in the early phase of our modelling was likely due to a still existing proportion of undocumented deaths among SARS-CoV-2 cases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0052.v1

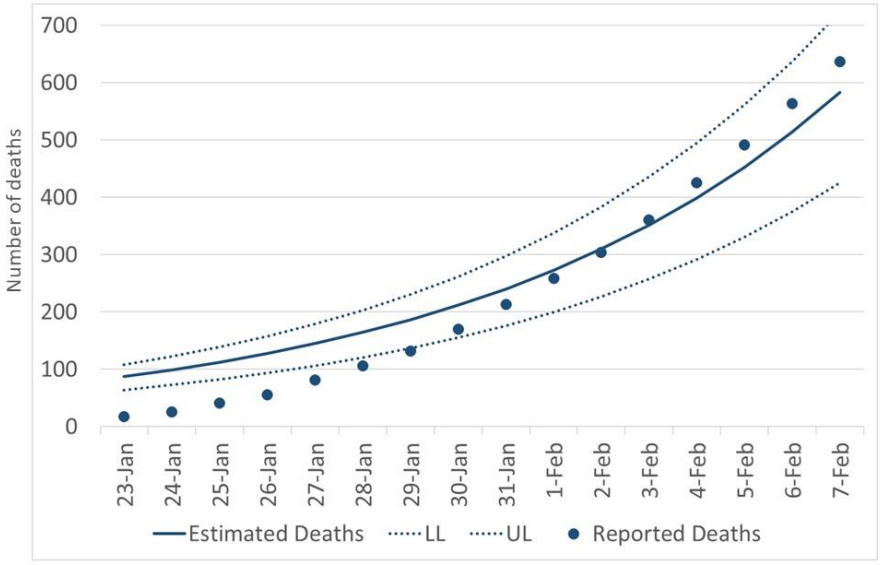

Figure 4. Fitting the SEIRD model to the reported number of deaths. The dots represent the daily cumulative number of reported deaths while the lines represent the estimate and 95% confidence intervals through the SEIRD model, along the temporal axis.

Using the best-fitting parameters reported in **Table 1**, we estimated that R_0 was 2.43 (95%CI= 2.42 – 2.44) with a total of 8,724 (95%CI= 8,478-8,921) estimated cases on 23 January 2020. These estimates and their comparison with reported cases (**Figure 5**) revealed 8,101 (95%CI= 7,855 - 8298) unreported cases, which represented 92.9% (95%CI= 92.5% - 93.1%) of estimated cases.

Table 1. Initial conditions, assumptions and best-fitting parameters in the baseline scenario

SEIRD Parameters	Definition	Assumed or Estimated Parameters
β^a	Transmission rate	0.73 (95%CI= 0.72 – 0.74)
σ	Infection rate	0.19
γ	Removing rate	0.28
μ^b	Probability of dying	0.015 (95%CI= 0.011 – 0.018)

[a] Estimated through the model with a potential range $0.1 \leq \beta \leq 1.0$
[b] Assumed to be 1/5.2 days according to Li and colleagues [2]
[c] Assumed to be 1/3.5 days according to Li and colleagues [3]
[d] Estimated through the model with a potential range $0.01 \leq \mu \leq 0.20$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0052.v1

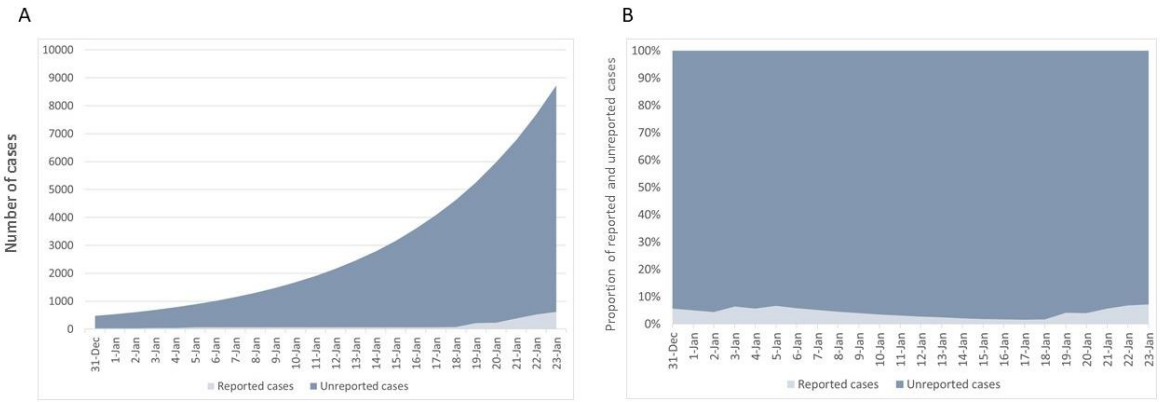

Figure 5. Estimated number of cases (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020.

Accordingly, the estimated number of new infections from 31 December 2019 to 23 January 2020 was 8,307 (95%CI= 8,069 – 8,498) (**Figure 6**). The proportion of unreported new infections by day ranged from 52.1% to 100%, which resulted in a total of 7,684 (95%CI= 7,446 – 7,875) unreported new infections and a proportion of 91.8% (95%CI= 91.6% - 92.1%).

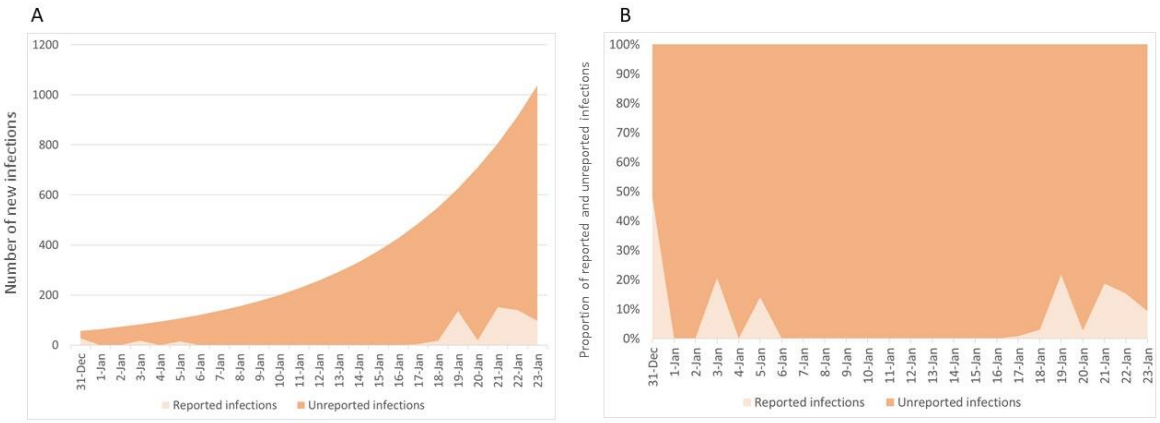

Figure 6. Estimated number of new infections (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020.

Given that infectious period was one of the most debated epidemic parameters – with previous estimates ranging from 3 to 20 days – we performed a sensitivity analysis, where we fitted the SEIRD model with different γ values. However, neither estimated values nor unreported proportions were sensitive to changes in the infectious period (**Supplementary Material**). Instead, R_0 would increase to 4.07 (95%CI= 3.91 - 4.17) or 6.50 (95%CI= 6.45 - 6.55) if we assumed γ as 0.1 and 0.05, respectively.

4. Discussion

In this study, we estimated the unreported number of SARS-CoV-2 cases in China prior to the 23 January 2020 lockdown. Our estimates revealed a very high proportion of unreported new infections every day, which resulted in 92.9% unreported cases. This finding was almost aligned with other recent estimates of unreported infections for the same time period [3,13]. For instance, Li and colleagues [3] reported that 86% of all infections were undocumented prior to travel restrictions, and that the transmission rate of undocumented infections was approximately 50% of documented infections. Yet, we obtained similar estimates by using a modified SEIR model, which took into account of dead individuals in the removed state. To the best of our knowledge, our study was the first that applied a SEIRD model to estimate the number of infections from observed deaths. Only a

few research groups are investigating the SARS-CoV-2 epidemic curve by calculating backwards from the deaths observed over time [5]. Our findings were also corroborated by the estimated R_0 – approximately 2.4 – that was consistent with previous estimates [3,5,9,14,15] and that indicated a high capacity for sustained transmission at the beginning of the epidemic.

We recognized that our approach relied on several assumptions and that many parameters had to be fixed. However, for these assumptions, we have provided reasonable grounds and relevant citations to previous studies. Moreover, we performed a sensitivity analysis for that parameters that required further investigations. Yet, sensitivity analyses made using alternative γ values (i.e. the inverse of infectious period) gave similar estimates of unreported cases but higher values of R_0. Given that, we cannot rule out some degree of uncertainty of our estimates, however, they will be more reliable as more knowledge become available.

In conclusion, our estimates are important for a better understanding of SARS-CoV-2 epidemic in China and in other countries. Our approach, based on the observed deaths, has proven to be reliable to estimate prevalence and incidence of undocumented SARS-CoV2 infections. Thus, our model could be applied on other countries with different surveillance and testing policies, and partially explains - for instance - differences in epidemic transmission and case fatality risk worldwide.

Author Contributions: Conceptualization, A.M. and A.A.; methodology, A.M. and S.B.; software, A.M.; formal analysis, A.M. and M.B.; data curation, A.M. and S.B.; writing—original draft preparation, A.M. and M.B.; writing—review and editing, all the Authors; visualization, A.M.; supervision, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Assessorato della Salute, Regione Siciliana - Progetti Obiettivo di Piano Sanitario Nazionale.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization, W.H. N ovel coronavirus 2019. Availabe online: https://www.who.int/emergencies/diseases/ (accessed on 2 April 2020)
2. Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K.S.M.; Lau, E.H.Y.; Wong, J.Y., et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* **2020**, *382*, 1199-1207, doi:10.1056/NEJMoa2001316.
3. Li, R.; Pei, S.; Chen, B.; Song, Y.; Zhang, T.; Yang, W.; Shaman, J. Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (SARS-CoV2). *Science* **2020**, doi:10.1126/science.abb3221.
4. Onder, G.; Rezza, G.; Brusaferro, S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA* **2020**, doi:10.1001/jama.2020.4683.
5. Imperial College COVID-19 Response Team. *Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries*; 2020.
6. European Center for Disease Prevention and Control. Download today's data on the geographic distribution of COVID-19 cases worldwide. Availabe online: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide (accessed on on 2 April 2020)
7. Read, J.; Bridgen, J.; Cummings, D.; Ho, A.; Jewell, C. Novel Coronavirus 2019-nCoV: Early Estimation of Epidemiological Parameters and Epidemic Predictions. medRciv: 2020.
8. Kucharski, A.; Russell, T.; Diamond, C.; Group, C.n.; Funk, S.; Eggo, R. Analysis of Early Transmission Dynamics of nCoV in Wuhan. 2020. medRciv: 2020.

9. Wu, J.T.; Leung, K.; Leung, G.M. Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-nCoV Outbreak Originating in Wuhan, China: a Modelling Study. *Lancet* **2020**, *395*, 689-697, doi:10.1016/S0140-6736(20)30260-9.
10. Boldog, P.; Tekeli, T.; Vizi, Z.; Dénes, A.; Bartha, F.A.; Röst, G. Risk Assessment of Novel Coronavirus COVID-19 Outbreaks Outside China. *J Clin Med* **2020**, *9*, doi:10.3390/jcm9020571.
11. Wang, H.; Wang, Z.; Dong, Y.; Chang, R.; Xu, C.; Yu, X.; Zhang, S.; Tsamlag, L.; Shang, M.; Huang, J., et al. Phase-adjusted Estimation of the Number of Coronavirus Disease 2019 Cases in Wuhan, China. *Cell Discov* **2020**, *6*, 10, doi:10.1038/s41421-020-0148-0.
12. van den Driessche, P. Reproduction Numbers of Infectious Disease Models. *Infect Dis Model* **2017**, *2*, 288-303, doi:10.1016/j.idm.2017.06.002.
13. Zhao, S.; Musa, S.S.; Lin, Q.; Ran, J.; Yang, G.; Wang, W.; Lou, Y.; Yang, L.; Gao, D.; He, D., et al. Estimating the Unreported Number of Novel Coronavirus (2019-nCoV) Cases in China in the First Half of January 2020: A Data-Driven Modelling Analysis of the Early Outbreak. *J Clin Med* **2020**, *9*, doi:10.3390/jcm9020388.
14. Du, Z.; Wang, L.; Cauchemez, S.; Xu, X.; Wang, X.; Cowling, B.J.; Meyers, L.A. Risk for Transportation of 2019 Novel Coronavirus Disease from Wuhan to Other Cities in China. *Emerg Infect Dis* **2020**, *26*, doi:10.3201/eid2605.200146.
15. Riou, J.; Althaus, C.L. Pattern of Early Human-to-human Transmission of Wuhan 2019 Novel Coronavirus (2019-nCoV), December 2019 to January 2020. *Euro Surveill* **2020**, *25*, doi:10.2807/1560-7917.ES.2020.25.4.2000058.