Journal of
Clinical Medicine

MDPI

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 [1,3,*]

[1] Department of Medical and Surgical Sciences and Advanced Technologies "GF Ingrassia", University of Catania, 95123 Catania, Italy; andrea.maugeri@unict.it (A.M.); martina.barchitta@unict.it (M.B.)
[2] Department of Mathematics and Computer Science, University of Catania, 95123 Catania, Italy; battiato@dmi.unict.it
[3] Azienda Ospedaliero-Universitaria "Policlinico-Vittorio Emanuele", 95123 Catania, Italy
* Correspondence: agodia@unict.it

Received: 4 April 2020; Accepted: 28 April 2020; Published: 5 May 2020

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 the 23 January 2020 restrictions. To do this, we developed a Susceptible–Exposed–Infectious–Recovered–Dead (SEIRD) model that estimated unreported infections from the reported number of deaths. Our approach relied on the fact that observed deaths were less likely to be affected by ascertainment biases than reported infections. Interestingly, we estimated that the basic reproductive number (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%) of infections going unreported. Agreement between our estimates and those from previous studies proves that our approach is reliable for estimating the prevalence and incidence of undocumented SARS-CoV-2 infections. Once it has been tested on Chinese data, our model could be applied to 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, has caused 81,554 cases and 3312 deaths among the Chinese population as of 1 April 2020 [1]. Whilst the number of SARS-CoV-2 infections is decreasing in China, other countries are still facing the epidemic and global efforts to contain the virus are still ongoing [1]. However, given the uncertainty about the transmissibility and virulence of SARS-CoV-2, the effectiveness of strategies against the current epidemic should be assessed properly [2]. In this scenario, the proportion of unreported infections is particularly noteworthy due to its crucial role in modulating the spread of the virus [2]. Indeed, unrecognized cases—often patients who experience mild or no symptoms—could silently expose a far greater proportion of the population to SARS-CoV-2 [3]. Correspondingly, it has recently been estimated that the transmission rate of undocumented infections was about half of those

documented, and that undocumented infections could be the source of eight out of ten documented cases [2]. Several countries are implementing stringent testing strategies for severely ill patients or those who have come into contact with documented cases [4]. This could lead to losing track of mild or asymptomatic patients who, however, could be infectious [5]. Therefore, looking only at reported case data could lead to biased speculations and hasty conclusions. In contrast, observed deaths are less likely to be affected by ascertainment biases, with the exception of deaths in the early phase of the epidemic [5].

For these reasons, we hypothesized that we could estimate the unreported number of infections by working directly with reported deaths. We employed a Susceptible–Exposed–Infectious–Recovered–Dead (SEIRD) model to estimate the number of unreported infections 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 was exploited but care was also taken to separate 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, however, 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.

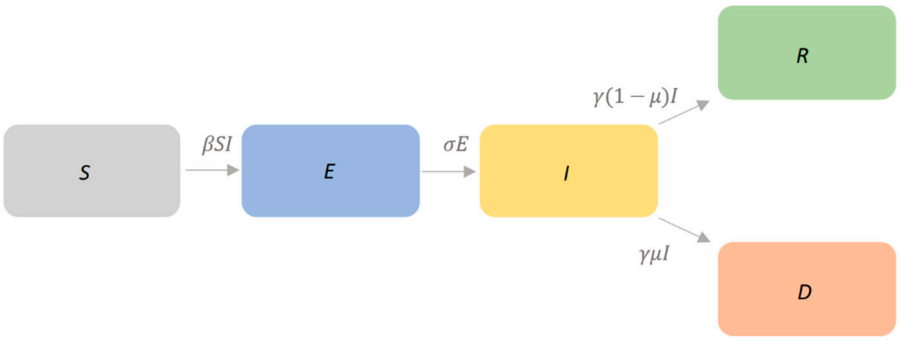

Figure 1. The employed Susceptible–Exposed–Infectious–Recovered–Dead (SEIRD) epidemic model for SARS-CoV-2. β, σ; γ, and μ denote the transmission rate, infection rate, removing rate, and probability of infectious individuals dying, respectively. S, E, I, R, and D denote susceptible, exposed, infectious, recovered, and dead individuals, respectively.

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 $D(t)$ are the numbers of susceptible, exposed (infected but not yet infectious), infectious, recovered, and dead individuals at the time (t);
- N is the total population as $N = S + E + I + R + D$. Note that the model relies on $\frac{S}{N}$ and hence is not affected by increasing N;
- β is the transmission rate, also known as the effective contact rate;
- σ is 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 is the removing rate and was assumed to be the inverse of the period between the onset of symptoms and recovery/death;
- μ is the probability of infectious individuals dying.

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

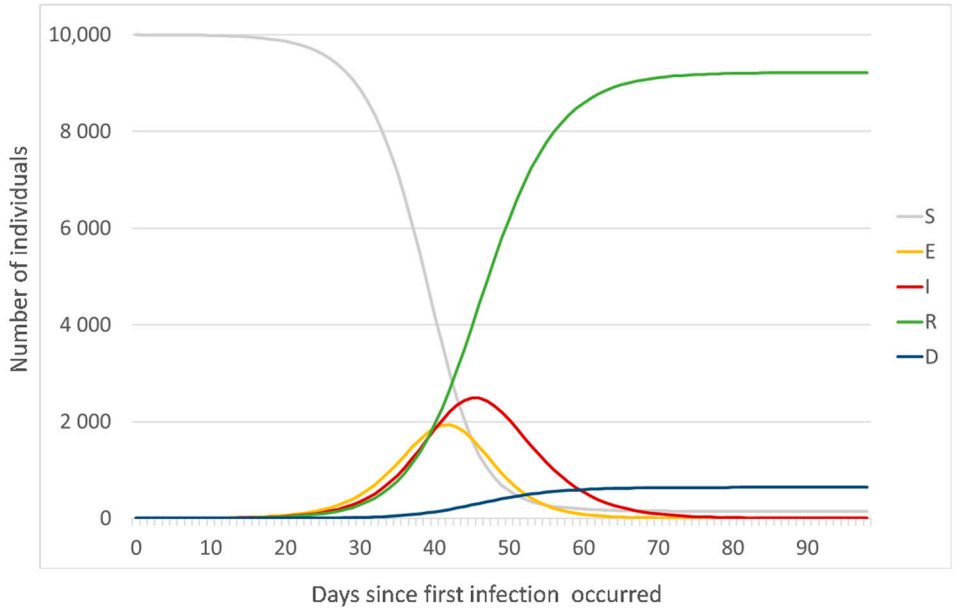

Figure 2. Generic representation of SEIRD states along the temporal axis. Estimates were obtained through a SEIRD model with β, σ, γ, and μ set as 0.8, 0.3, 0.2, and 0.2, respectively (for viewing purposes only).

In the current study, N was assumed to be 1 billion, R and D were initially set as 0, while the initial 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 trends, 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 had 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, respectively, 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 number of deaths, we applied a least squares optimization using an evolutionary algorithm (population size = 1×10^5, convergence = 1×10^{-6}, and mutation rate = 5×10^{-2}) and simulations (n = 1000) 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. The values of unreported new infections and total cases were obtained by subtracting the reported numbers from those estimated and are reported as a percentage. The basic reproductive number (R_0) was calculated from the SEIRD model as previously described [12]. We also performed sensitivity analyses to evaluate the impact of varying the infectious period and the initial number of infectious individuals on the estimation of unreported cases and infections.

3. Results

The cumulative number of cases and deaths by the day of the 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 ascertainment biases than documented cases.

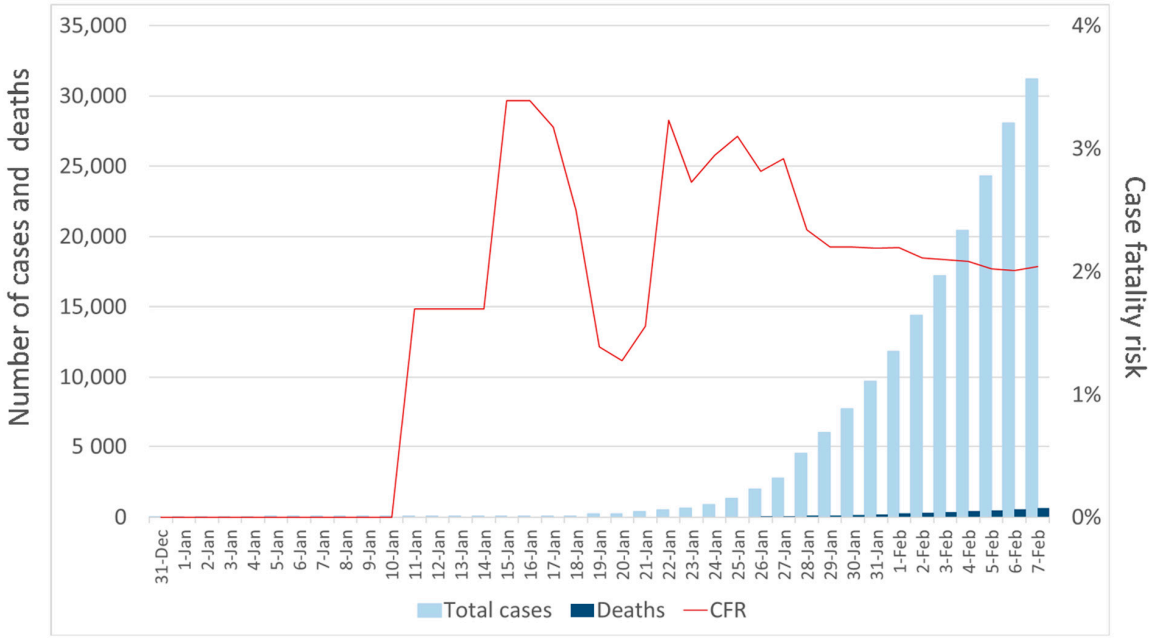

Figure 3. Number of reported cases and deaths in China from 31 December 2019 to 7 February 2020. The bars represent the cumulative number of reported coronavirus (SARS-CoV-2) cases and related deaths while the red line represents the case fatality risk (CFR).

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

Using the best-fitting parameters reported in Table 1, we estimated that the R_0 was 2.43 (95%CI = 2.42–2.44) with a total of 8724 (95%CI = 8478–8921) estimated cases on 23 January 2020. These estimates and their comparison with reported cases (Figure 5) revealed 8101 (95%CI = 7855–8298) unreported cases, which represented 92.9% (95%CI = 92.5%–93.1%) of estimated cases.

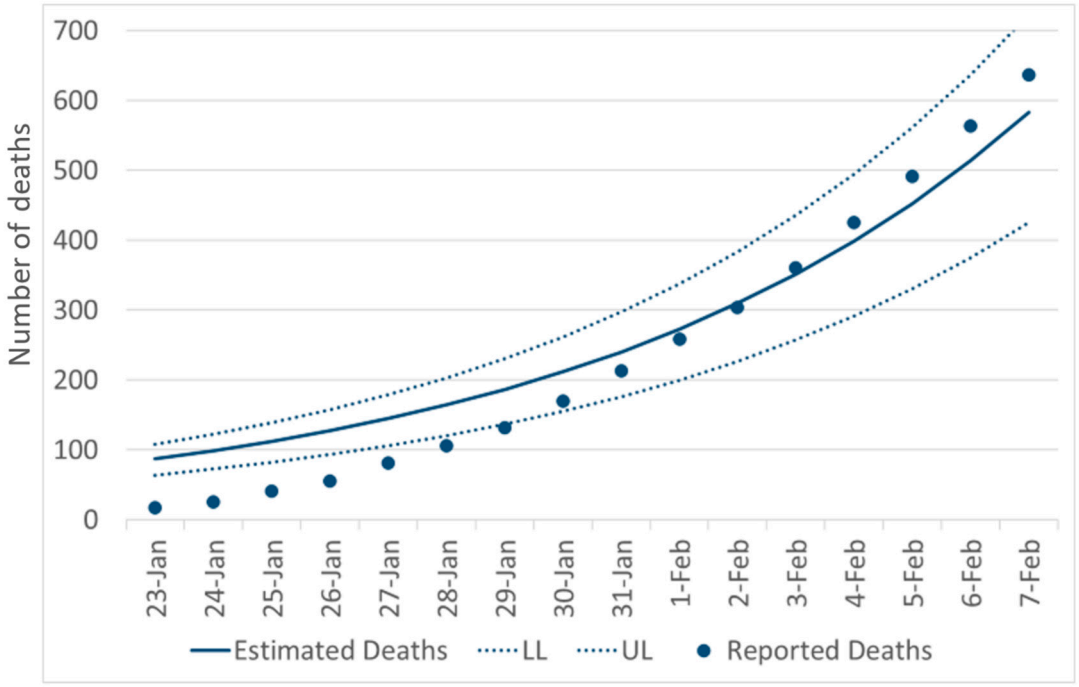

Figure 4. Fitting the SEIRD model to the reported number of deaths. The dots represent the daily cumulative number of reported deaths while the lines along the temporal axis represent the estimate and 95% confidence intervals (Upper and Low Level: UL, LL) through the SEIRD model.

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 [b]	Infection rate	0.19
γ [c]	Removing rate	0.28
μ [d]	Probability of dying	0.015 (95%CI = 0.011–0.018)

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

Accordingly, the estimated number of new infections from 31 December 2019 to 23 January 2020 was 8307 (95%CI = 8069–8498) (Figure 6). The proportion of unreported new infections by day ranged from 52.1% to 100%, which resulted in a total of 7684 (95%CI = 7446–7875) unreported new infections and a proportion of 91.8% (95%CI = 91.6%–92.1%).

Given that the removing rate 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 removing rate (Supplementary Figures S1–S4). Instead, the 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 to be 0.1 and 0.05, respectively. Similarly, we analyzed the condition where the initial number of infectious individuals was 100 times greater than the baseline scenario. Nevertheless, the estimates were not sensitive to changes, while the R_0 decreased to 1.60 (95%CI = 1.45–1.76) (Supplementary Figures S5–S6**)**.

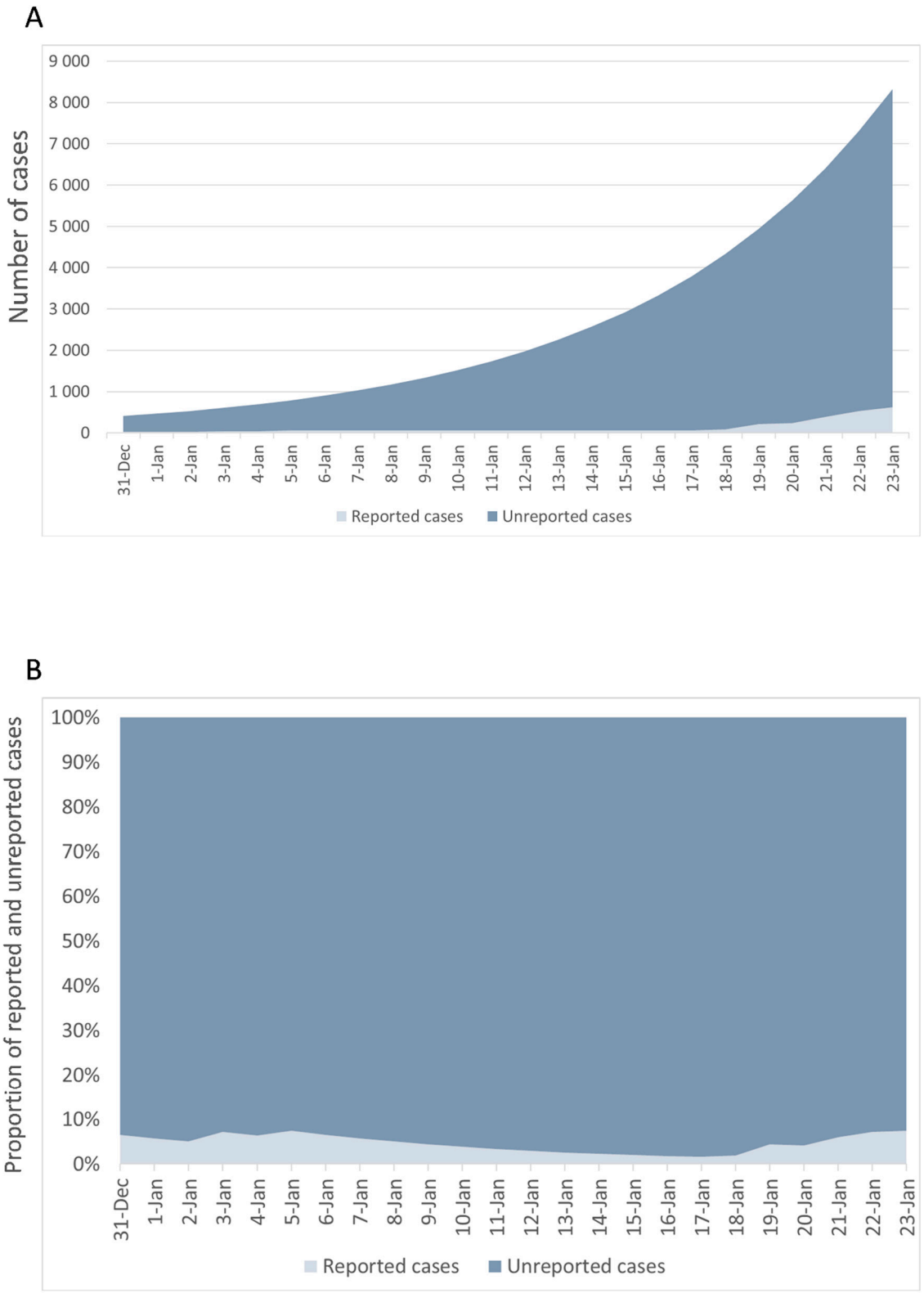

Figure 5. Estimated number of cases (**A**) and proportion of unreported events (**B**) from 31 December 2019 to 23 January 2020.

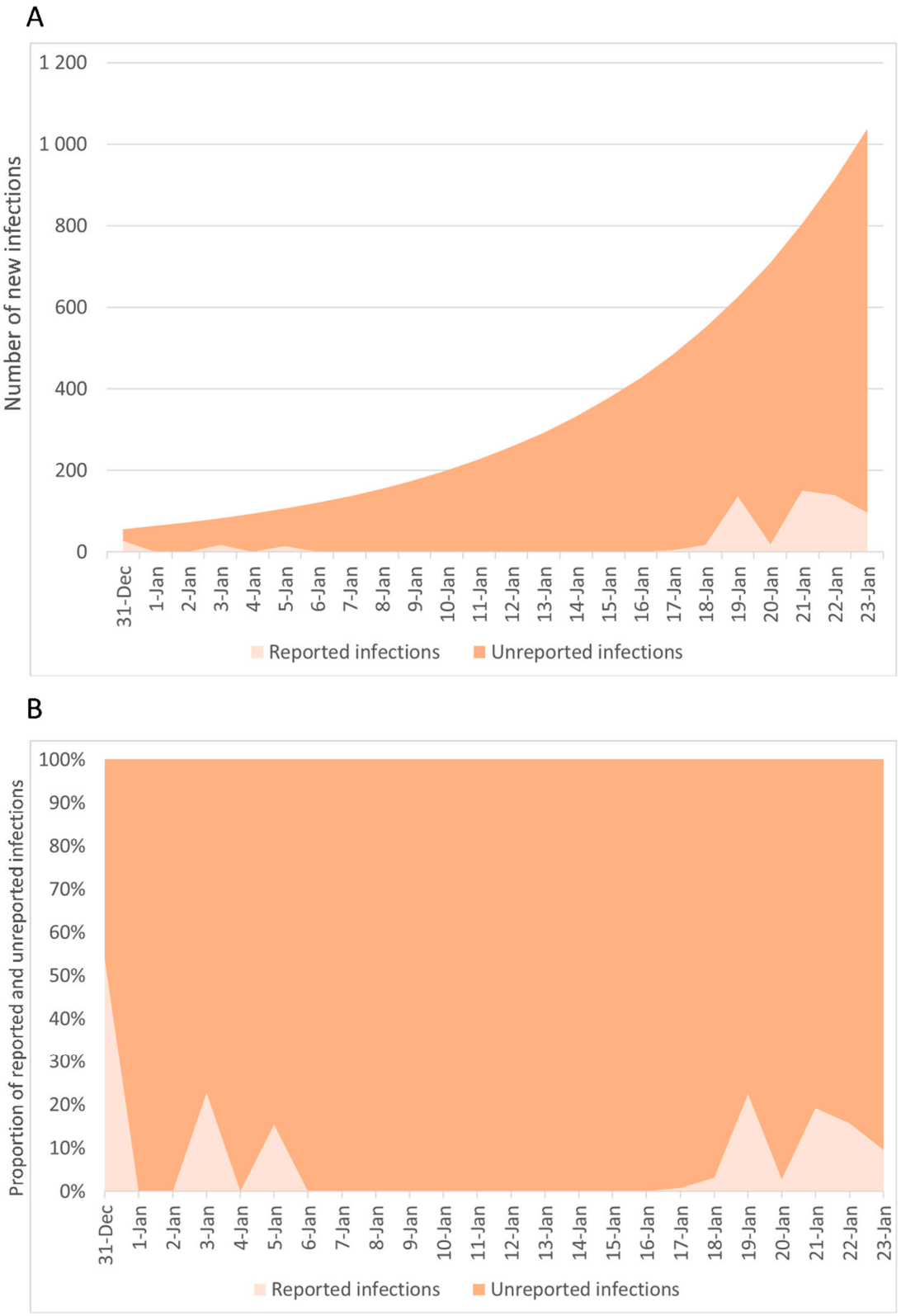

Figure 6. Estimated number of new infections (**A**) and proportion of unreported events (**B**) from 31 December 2019 to 23 January 2020.

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 reveal 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 [2,13]. For instance, Li and colleagues [2] 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 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, which was consistent with previous estimates [2,5,9,14,15] and which indicated a high capacity for sustained transmission at the beginning of the epidemic.

Our study has some limitations. First, our hypothesis was that data on deaths were less likely to be affected by under-reporting than data on infections, given that the proportion of deaths among mild or asymptomatic patients was supposed to be lower [4]. The number of deaths, however, was not exempt from ascertainment issues. Indeed, clear criteria for the definition of SARS-CoV-2-related deaths were not available [4], and thus it might be possible that some deaths were caused by pre-existing conditions rather than this infection. Nevertheless, our model did not rely on a causal relationship between SARS-CoV-2 infection and deaths but only on the probability of dying among infectious individuals (i.e., μ). This parameter, along with the removing rate (i.e., γ), regulated the transition of infectious individuals to death. We also recognized that our approach relied on several assumptions and that many parameters had to be fixed. However, we have provided reasonable grounds and relevant citations to previous studies and performed a sensitivity analysis for those parameters that required further investigations. Nevertheless, sensitivity analyses made using alternative γ values or increasing the initial number of infectious individuals gave similar estimates of unreported cases but different values of the R_0. Given this, we cannot rule out some degree of uncertainty from our estimates; however, they will be more reliable as more data become available.

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

Supplementary Materials: The following are available online at http://www.mdpi.com/2077-0383/9/5/1350/s1, Figure S1: Estimated number of cases (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020 using $\gamma = 0.1$; Figure S2: Estimated number of cases (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020 using $\gamma = 0.05$; Figure S3: Estimated number of new infections (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020 using $\gamma = 0.1$; Figure S4: Estimated number of new infections (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020 using $\gamma = 0.05$; Figure S5: Estimated number of cases (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020 using an initial infectious individuals number of 100; Figure S6: Estimated number of new infections (A) and proportion of unreported events (B) from 31 December 2019 to 23 January 2020 using an initial infectious individuals number of 100.

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 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 (PSN 2014–4.9.2).

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

References

1. World Health Organization. Novel Coronavirus 2019. Available online: https://www.who.int/emergencies/diseases/ (accessed on 1 April 2020).
2. 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**. [CrossRef] [PubMed]
3. 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. [CrossRef] [PubMed]
4. Onder, G.; Rezza, G.; Brusaferro, S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA* **2020**. [CrossRef] [PubMed]
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. *arXiv* **2020**, arXiv:2004.11342.
6. ECDC. Download Today's Data on the Geographic Distribution of COVID-19 Cases Worldwide. Available online: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide (accessed on 1 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**. [CrossRef]
8. Kucharski, A.; Russell, T.; Diamond, F.S.; Eggo, R. Analysis of early transmission dynamics of nCoV in Wuhan. *Lancet Infect. Dis.* **2020**, *20*, 553–558. [CrossRef]
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. [CrossRef]
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*, 571. [CrossRef] [PubMed]
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. [CrossRef] [PubMed]
12. van den Driessche, P. Reproduction numbers of infectious disease models. *Infect. Dis. Model.* **2017**, *2*, 288–303. [CrossRef] [PubMed]
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*, 388. [CrossRef] [PubMed]
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*. [CrossRef] [PubMed]
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*. [CrossRef] [PubMed]