Correcting under-reported COVID-19 case numbers

Alexander Lachmann[1]

[1]Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA

The COVID-19 virus has spread worldwide in a matter of a few month. Healthcare systems struggle to monitor and report current cases. Limited capabilities in testing result in difficult to guide policies and mitigate lack of preparation. Since severe cases, which more likely lead to fatal outcomes, are detected at a higher percentage than mild cases, the reported death rates are likely inflated in most countries. Such under-estimation can be attributed to under-sampling of infection cases and results in systematic death rate estimation biases.

The method proposed here utilizes a benchmark country (South Korea) and its reported death rates in combination with population demographics to correct the reported COVID-19 case numbers. By applying a correction, we predict that the number of cases is highly under-reported in most countries. In the case of China, it is estimated that more than 700.000 cases of COVID-19 actually occurred instead of the confirmed 80,932 cases as of 3/13/2020.

The full analysis workflow and data is available at:
https://www.kaggle.com/lachmann12/population-demographics-correction-covid-19

COVID-19 | COVID-19 | demographics | public health
Correspondence: ***alexander.lachmann@mssm.edu***

Introduction

Severe acute respiratory syndrome-related COVID-19 (COVID-19) is a novel virus with the initial outbreak most likely in China (1). It has reached pandemic status by the World Health Organization within less than four months of initial reports of the disease. The origin of the virus can be traced back to related strains predominantly found in bats (2). Individuals infected by the disease can experience a series of symptoms, including cough, chills, fever, and shortness of breath (3). From data currently available, fatal disease progression is higher than that of the common influenza strains and as such it resulted in more deaths than recent virus of Severe Acute Respiratory Syndrome (SARS) and Search Results Middle East Respiratory Syndrome (MERS) combined. (**?**). The infection rate of COVID-19 has been estimated between a R_0 of 2 and up to 6.49 (4) compared to influenza with about 1.3 (5). The severity of infection is highly correlated to the age of the infected individual. Younger parts of a population present a much lower risk than older populations. A current data release from the Center for Disease Control in South Korea shows that while there are no reported fatalities for individuals under 30 years of age, the death rate for individuals older than 80 is over 8% (6). Figure 1 shows eight countries with a significant number of reported COVID-19 cases. China, which has been the origin of the outbreak, registered the most cases with over 80,000. Through severe measures such as curfews, new infections have slowed significantly. Other countries that have been only recently affected are still in the exponential growth curve. Countries like Italy have only recently taken action to slow the spread of the virus. With a reported incubation time of about five days, it will take several days until the effects of a slowdown will be visible (7). Another country that is currently experiencing high numbers of reported COVID-19 cases is Iran, with more than 12,000 confirmed cases. Due to the limited information available, most parameters describing the dynamics of the disease spread have significant uncertainties around them. Healthcare systems in most countries are not capable of monitoring the exponential growth of a virus in this manner. South Korea, as of writing, has the most extensive capabilities of testing individuals with a capacity of around 20,000 tests a day. Hence, South Korea represents the best benchmark country in order to correct reported COVID-19 cases in other countries. The proposed method uses demographic information to identify the fraction of the vulnerable population. Countries such as China have a generally younger population reducing the overall risk of fatal outcomes and thereby should result in a lower death rate compared to South Korea. Countries, such as Italy with an older population compared to South Korea, should have a higher death rates. Estimating the true case count is relevant in identifying the correct measures to stop the disease from spreading.

Methods

A. Data. The case correction relies on two datasets. The first is the data published by the WHO, which is updated every day and contains case, recovery, and death numbers for countries reporting all known COVID-19 cases (8). The second dataset is a global demographic database maintained by the United Nations (9). This database contains the number of individuals per year of age for more than 200 countries. For the analysis, we extracted the data between 2007 and 2019. We always choose the most recent data entry for the countries if multiple exist. This file is hosted as a Kaggle dataset at: https://www.kaggle.com/lachmann12/world-population-demographics-by-age-2019.

B. Assumptions. This method makes a series of assumptions in order to adjust reported COVID-19 cases compared to the benchmark country (South Korea).

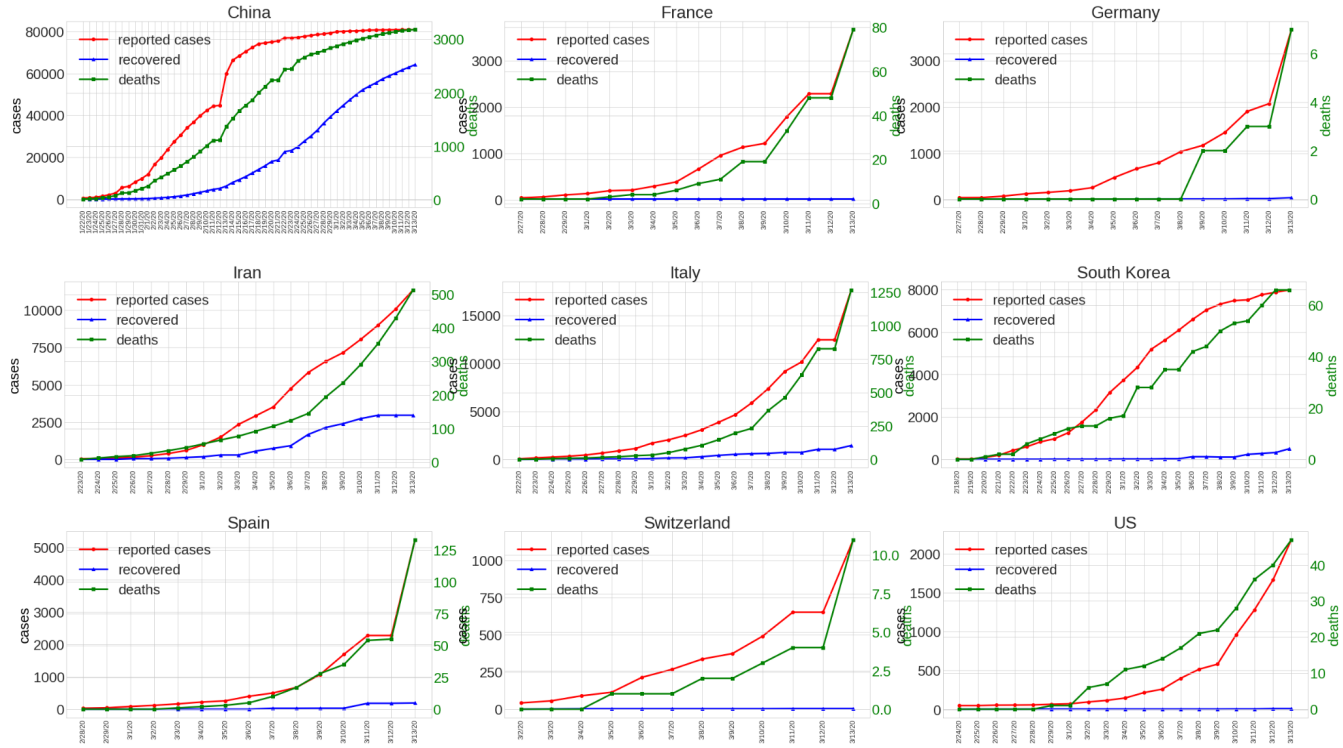

Fig. 1. Case progression for eight countries with highest number of COVID-19 cases with corresponding recoveries and deaths.

- **Deaths are confirmed equally** It is assumed that if a death occurs, caused by COVID-19, the case is confirmed. When there is under-reporting, the death rate would be lower than the true death rate.
- **The population is infected uniformly** We assume that the probability of infection is uniformly distributed across. The probability of an 80-year-old person to become infected is equal to the probability of a 30-year-old to become infected.
- **Treatment has minor influence on outcome** The provided healthcare in countries is comparable. For developed countries such as Italy and South Korea, it is assumed that the population has similar access to treatment. The death rates reported by age group are thus applicable in all countries.

C. Case Adjustment. Figure 2 shows the progression of death rate estimates for the US, Italy, China, and South Korea. It can be noted that South Korea shows the most consistent death rate estimates. Additionally, it also shows a significantly lower death rate compared to other countries, with the exception of Germany (not shown). The change of death rate over time within the same country is potentially caused by changes in the number of false-negative cases, meaning that many infections go unnoticed until they become fatal. In the case of Italy, there might not have been sufficient capacity to confirm infections. With a smaller fraction of potential cases tested, the estimated death rate will increase. In the case of Italy, the estimated rate increased from 2% to more than 6%.

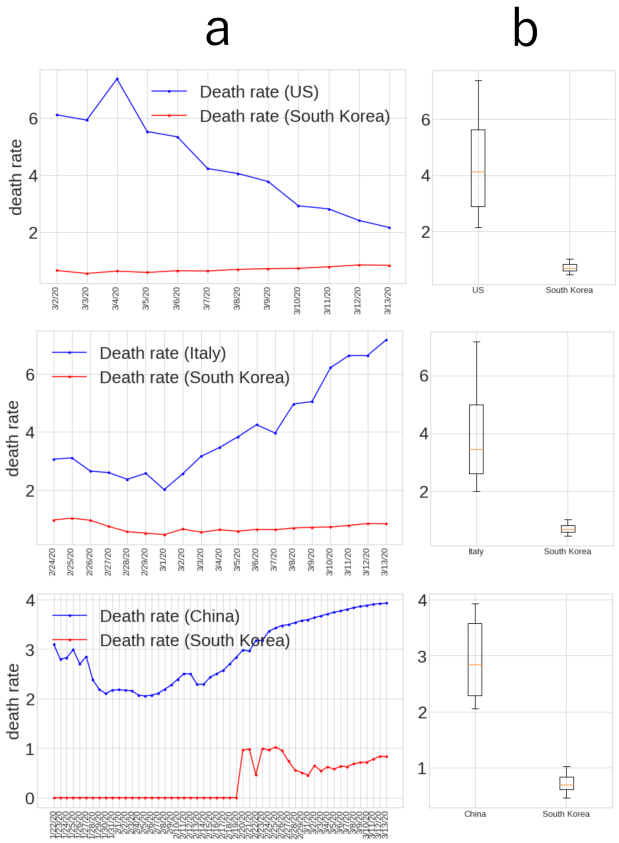

Fig. 2. Death rates for US, Italy, and China in comparison with Korea over time.

This method requires the comparison of two countries with sufficient confirmed cases and reported deaths. One country (target country) will be adjusted, given the information from the second country (benchmark country). In order to adjust for the difference in the population demographics of the target country, $\mathbb{T}$, and the benchmark country, $\mathbb{B}$, we compute a Vulnerability Factor ($V_{\mathbb{TB}}$).

$$V_{\mathbb{TB}} = \frac{\sum_{i=0}^{N} f_{\mathbb{T}i} r_i}{\sum_{i=0}^{n} f_{\mathbb{B}i} r_i}$$

, where $f_{\mathbb{T}i}$ is the fraction of the population with age i for target country $\mathbb{T}$, $f_{\mathbb{B}i}$ is the fraction of the population with age i for benchmark country $\mathbb{B}$, and r_i the death rate for age i. r_i is listed in Table 1.
If $V_{\mathbb{TB}} > 1$, then the population of $\mathbb{T}$ has a higher risk of fatal outcomes due to a larger percentage of the older population. It results in a higher death rate compared to $\mathbb{B}$. If $V_{\mathbb{TB}} < 1$, then $\mathbb{T}$ has a younger population and it should result in a lower death rate compared to $\mathbb{B}$.

Another correction factor is the fixed average death rate of the benchmark country, $D_{\mathbb{B}}$.

$$D_{\mathbb{B}} = \frac{\sum_{i=0}^{K} d_{\mathbb{B}i}}{K}$$

, where $d_{\mathbb{B}i}$ is the death rate of day i.
With both normalization factors we can now adjust the expected cases relative to $\mathbb{B}$. The methods applies the normalization to each time point. The original case number $o_{\mathbb{T}i}$ is adjusted for $\mathbb{T}$ and $\mathbb{B}$ at time point i with:

$$a_{\mathbb{TB}}(o_{\mathbb{T}i}) = \frac{o_{\mathbb{T}i}}{V_{\mathbb{TB}} D_{\mathbb{B}}}$$

Results

By applying the proposed correction, the number of adjusted cases is significantly higher for most countries. Figure 3a illustrates population age distributions. Figure 3b shows the expected number of fatal outcomes for a 100% infection rate. The vulnerability factor for the US compared to South Korea is 1.07. This means that the population is equally vulnerable to fatal outcomes of COVID-19 infections. Italy, in contrast, has a vulnerability factor of 1.57. This is due to a higher fraction of the population being at a higher risk of death. This would indicate the expected death rate would be 57% higher in Italy compared to South Korea. China, with a younger population relative to South Korea, has a vulnerability factor of 0.63. The expected death rate in China should be lower than in South Korea based on the population risk. After applying the case adjustment, we observe a significant increase in the number of COVID-19 infections. The discrepancy in reported death rates in combination with favorable population scores in the case of China suggests a large number of unreported COVID-19 infections. The adjustment suggests around 702,518 cases compared to 80,932 reported cases. This equates to an 868% higher case count than previously reported. The corrections for Italy and the US are similar, but not as extreme. Italy has an adjusted number of cases of 112,182 cases and the US potentially 6,085 cases. Table 2 shows the adjusted number of cases for a selected number of countries. Iran is the country with the most substantial adjustment of 1,363%, reaching 154,853 cases.

Summary

This study suggests that the current reporting of COVID-19 cases is significantly underestimating the true scale of the pandemic. The lack of testing makes the estimation of the true death rate difficult and causes a significant misinformation. This study tries to leverage the information derived from a well-tested sub-population (South Korea). With testing capacities of 20,000 tests daily, it has the largest and most accurate coverage compared to all other countries as of writing. The low false-negative rate in detecting COVID-19 infections leads to the lowest death rate compared to all other countries (0.84) with major case count. By applying the parameters, estimated from this benchmark country, the proposed method can adjust global COVID-19 case numbers. This method is limited in its ability to predict the exact number of cases accurately. The method relies on the assumption that deaths by COVID-19 are detected and reported reliably. False-negative rates can have a distorting effect on the case adjustment. This is especially true if the benchmark country does not adequately report deaths from COVID-19. Germany, as an example, only reports eight deaths from with 3,675 reported cases. This could be due to the very recent increase in actual cases leaving not enough time for fatal disease progression. Over time, when more data is available, death rates will most likely increase in Germany. Additionally, the assumption of a globally similar death rate is untested. Improvements in this method could look at the case number of other viral diseases to see if there are significant differences between countries. This method explains the observed fluctuations in death rate over time by country. It is unlikely that the death rate in the same country can fluctuate by multiple percent points over a period of a few days. This method suggests that due to the fast exponential growth of true case counts, most modern healthcare systems are not able to track the changes adequately. In addition, the method suggests that computational tools can be used to impute missing information based on regions where testing and tracking is more advanced. It also highlights the importance of publicly accessible real time data and the relevance of combining global healthcare efforts.

ACKNOWLEDGEMENTS
I want to thank Dr Avi Ma'ayan and Federico Giorgi for feedback on the original manuscript and Alon Bar Tal for insightful discussion as well as the Kaggle community. Special thanks to the seamless accessibility of up-to-date COVID-19 case statistics published on GitHub by Johns Hopkins and the World Health Organization.

Bibliography

1. Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, pages 1–4, 2020.

		Confirmed Cases n (%)		Deaths n (%)		Fatality Rate %
Total		**7,979**	**(100.0)**	**67**	**(100.0)**	**0.84**
Age Groups	**80 and above**	253	(3.2)	21	(31.3)	8.30
	70-79	506	(6.3)	24	(35.8)	4.74
	60-69	985	(12.3)	14	(20.9)	1.42
	50-59	1,523	(19.1)	6	(9.0)	0.39
	40-49	1,117	(14.0)	1	(1.5)	0.09
	30-39	823	(10.2)	1	(1.5)	0.12
	20-29	2.274	(28.5)	0	(0.0)	0
	10-19	421	(5.3)	0	(0.0)	0
	0-9	77	(1.0)	0	(0.0)	0

Table 1. Confirmed cases of COVID-19 in the Republic of Korea and the corresponding deaths and fatality rates stratified by age groups as of 3/11/2020.

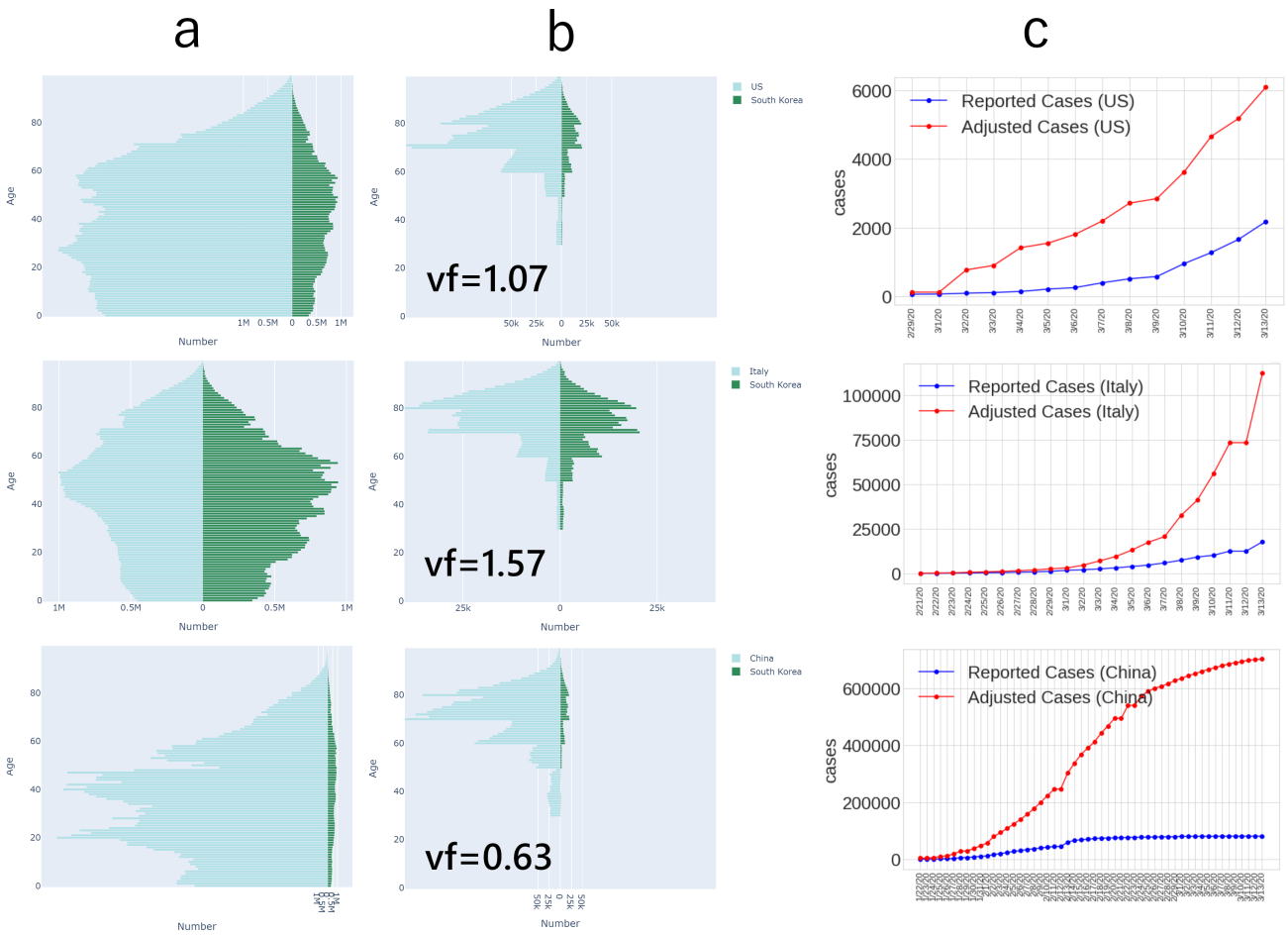

Fig. 3. a) Population demographic comparison between US, Italy, and China compared to South Korea. **b)** Population demographic with fatal outcome of COVID-19 at 100% infection rate based on Table 1. Vulnerability Factor $V_{\mathbb{TE}}$ relative to South Korean population (vf). **c)** Reported case numbers and adjusted case numbers.

	Reported Cases	Adjusted Cases	% adjustment
China	80,932	702,518	868
France	3,667	7,940	217
Iran	11,364	154,853	1,363
Italy	17,660	112,182	635
Spain	5,232	13,556	259
US	2,179	6,085	279

Table 2. Reported and adjusted cases compared to South Korea death rates and population demographics.

2. Chih-Cheng Lai, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, and Po-Ren Hsueh. Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and corona virus disease-2019 (covid-19): the epidemic and the challenges. *International journal of antimicrobial agents*, page 105924, 2020.
3. Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, Lei Huang, Chao Zhang, Shuhong Liu, Peng Zhao, Hongxia Liu, Li Zhu, et al. Pathological findings of covid-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, 2020.
4. Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of travel medicine*, 2020.
5. Joseph T Wu, Kathy Leung, Ranawaka APM Perera, Daniel KW Chu, Cheuk Kwong Lee, Ivan FN Hung, Che Kit Lin, Su-Vui Lo, Yu-Lung Lau, Gabriel M Leung, et al. Inferring influenza infection attack rate from seroprevalence data. *PLoS pathogens*, 10(4), 2014.
6. Republic of Korea Center of Disease Control. *Updates on COVID-19 in Republic of Korea*, 3/13/2020.
7. Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*.
8. World Health Organization et al. Laboratory testing for coronavirus disease 2019 (covid-19) in suspected human cases: interim guidance, 2 march 2020. Technical report, World Health Organization, 2020.
9. Gene Shackman, Xun Wang, and Ya-Lin Liu. Brief review of world demographic trends-trends in age distributions. *Available at SSRN 2180600*, 2012.