Recommendations for Unbiased and Random Sampling for COVID-19 Testing in Hotspots to Determine an Accurate Fatality Rate

Yasir Suhail[1,2], Junaid Afzal[3], Kshitiz[1,2,*]

[1.] Department of Biomedical Engineering, University of Connecticut Health, Farmington, CT; USA
[2.] Center for Cancer Systems Biology @ Yale, West Haven, CT; USA
[3.] Department of Medicine, University of San Francisco, San Francisco, CA; USA
* Correspondence to be addressed to kshitiz@uchc.edu

ABSTRACT

The spread of COVID-19 across the world has been modeled with varying fatality rates among the infected population. The infection fatality rates suffer from systematic biases and large estimation variance due to reliance on data from individuals who have been tested. In addition, these estimates may mask the effect and numbers of individuals who have developed immunity to SARS-CoV-2 within a subpopulation. Recently developed serological tests could be utilized to determine the true infection rate and fatality rate. However, the current methods to determine fatality rates from world-wide tests are potentially biased and error prone. To counter the large errors and variations in the reported fatality rates, and ascertain accuracy, we recommend random and unbiased sampling of individuals in a population where the virus has spread widely. We also provide a method to measure the dynamics of acquired immunity with a novel virus where serological tests may not be available.

INTRODUCTION

The spread of SARS-CoV-2 across the world has had an unprecedented societal, medical, and economic impact. As it spreads to more than 200 countries, many countries have attempted to stem the spread of the viral infection by systemic societal policies ranging from partial to complete shut-down of social interactions. This remarkable response was initially kindled by the early reports from the World Health Organization (WHO) that stated a fatality rate of over 3.8%[1] for COVID-19 as it was first detected in Wuhan, China and spread across the world. Thereafter, there have been multiple mathematical studies modeling the kinetics of disease spreading with and without these social distancing interventions[2]. However, these are all dependent on model parameters estimated from very limited, and overwhelmingly biased and non-uniform sampling. There appears to be vast differences across countries in the rate of fatalities, when calculated against measured cases, or recovered cases. This raises significant concerns regarding the accuracy of the estimates of fatality and morbidity rates, with far reaching consequences on health and economic policy.

We were among the first to raise concerns regarding the accuracy of determined fatality rates, and posited that they may be irrelevant parameters if the underlying spread of immunity is not accounted for[3]. It is possible that fatality and morbidity rates, as well as the dynamics of disease spread are substantially higher or lower than the data predict across different countries. To date, attempts have been to interpret the publicly available data regarding the number of tests performed, confirmed infections, reported fatalities, and recovered cases. These numbers vary widely across different countries, resulting in large variations in suggested mortality rates. Verity et al. give estimates for the infection fatality rates based on testing of foreign nationals repatriated from China[4]. While this sample may not have been directly biased with symptom severity, it is still likely to be highly correlated to age, health, and placement within social and physical contact networks, and therefore indirectly correlated with infection status and susceptibility to fatality. In this commentary, we explain our major concerns regarding the accuracy of reported fatality rates and suggest a rapidly deployable method to accurately assess these parameters. We believe our proposal will be helpful in accurately assessing the impact of this virus, and therefore strengthen healthcare systems to be better prepared for future pandemic.

Estimation of Fatality Rates from World-wide Reported Cases is Inaccurate, and Incorrect

WHO initially estimated a fatality rate of over 3.8% from COVID-19 based on the number of reported cases and number of fatalities in China, and thereafter in other countries[1]. Other reports also based the fatality rates on these numbers, which were collected by public health agencies in the countries where virus had started spreading. These reports calculated fatality rates by dividing the number of deaths reported by the number of positive cases tested for[5-7]. However, there are wide differences across countries in the number of people who are tested, the availability of test kits, as well as the stratification of the population that got tested. Reverse transcription polymerase chain reaction (RT-PCR), the current method of choice to confirm the presence of COVID-19, only informs about the live status of the virus in the population. In several geographical locations, only cases with severe symptoms or associated illnesses are tested due to limited resources and/or to reduce the overwhelming of healthcare system. With this criteria, the presentation of current COVID-19 data in as case fatality rate (CFR) is difficult to interpret [5-7], masking the true extent of the disease spread. As CFR is a measure that is associated with the number of deaths in a total population afflicted with the disease, the presence of a significant sub-population with milder symptoms, such as 80% of cases in Wuhan[8], highlight the careful consideration of its use in estimation of the severity of crisis before a widespread testing is adopted[6,9].

A better estimate of asymptomatic and undiagnosed cases, using antibody-based testing, will inform about the past exposure of COVID-19, thereby ensuring the inclusion of the cohort that has developed immunity but is no more carrying the viral burden. COVID-19 is known to induce a detectable antibody response following few days of infection but in an emergent situation like COVID-19, with the lack of large-scale serological testing, it's difficult to estimate the asymptomatic cases in total population — FDA has recently approved qSARS-CoV-2 IgG/IgM, serological testing from Cellex Inc. Furthermore, large scale serological testing may still suffer from systemic biases in data collection across different populations and countries.

Here, we list the chief concerns associated with the currently deployed methods to measure rate of fatalities.

1. Underestimation of the infected population.

There is a strong need for better models in interpreting the fatality rates of COVID-19 like infectious diseases instead of current usage of CFR[5,7]. With the testing limitation, people with severe cases are screened heavily, while people with mild to no symptoms, or those who have developed immunity and are no longer carrying an active viral load, would not be detected by RT-PCR based testing. There is still no accurate estimate of the percentage of population which has developed immunity and are asymptomatic. The focus towards serological testing would be essential to estimate the (previously) infected population and to estimate the true disease spread.

2. Systematic biases in testing.

There are vast differences between the numbers of tests countries have performed, as well as differences in the stratification of patients to whom tests are provided. In many countries, tests are largely conducted for *symptomatic patients* which could bias the result in the direction of high death rates (e.g. India, Italy), while in countries where tests are more aggressively performed over a larger swathe of population, the data on death/active case may be more accurate (e.g. Iceland, Germany). Even here, people who have contracted the virus but do not carry an active load will be missed using current PCR based testing methods, heavily biasing the measurement. In addition, while more testing will give more accurate tests, it is still biased for the people who have been tested, an effect that is eliminated by random sampling.

3. Establishing the causality of COVID-19 in fatality, or morbidity of patients.

The earliest reports indicated that the average age of patients dying from COVID-19 is high, and even here, were correlated with high incidences and severity of co-morbidity[1,10]. In Italy, where the death rate is reported to be very high, significant co-morbidities were associated with COVID-19 based deaths[11]. It may not be meaningful to count each death associated with the viral infection as a coronavirus related death, if the virus is merely a correlative factor in the death. At the same time, it is likely that the viral load is indeed causal in fatalities in pre-disposed populations with significant co-morbidities. We believe that stratifying COVID-19 as primary, secondary, or tertiary instrument of fatality is necessary to determine the true mortality rate of this virus.

Randomized Unbiased Serological Sampling of Widely Infected Population is Necessary to Determine True Fatality Rates

Since the estimates of fatality rates calculated from collated data across different institutions and countries have large variance with some questioning the methodologies involved in reported figures[7,9], along with the other concerns mentioned above, we recommend a random sampling of a population which has suffered a large infection load. As serological tests are becoming available to detect the antibodies against SARS-CoV-2 specific antigen, it is possible to identify individuals who have developed immunity against the virus but may not necessarily test positive owing to reduced viral load. We believe it is necessary to test if the wider population in an area with high death numbers is a consequence of (i) a wider spread of the disease, or (ii) a larger death rate in a smaller subpopulation that has contracted the virus. In order to minimize the variance of the infection death rate, this random testing should be conducted among a population wherein the infection is understood to have spread widely.

We propose using census, tax ID, or driving licenses in a given area as an unbiased identifier of a sampling set, upon which the serological test and PCR (or pooled next generation sequencing-NGS based tests) could be conducted. We provide some calculations regarding the sample sizes that will be required to gain an accurate estimate of the community infection rates and infection fatality rates. A random selection of individuals will provide estimates without systematic biases, therefore the appropriate measure of accuracy need only be concerned with the variance of the estimates. In the following, we have chosen to frame this in terms of mostly confidence intervals and hypothesis testing.

Suppose the fraction of population that has contracted SARS-CoV-2 detectable by a serological test (the infection rate) is p. In addition, assume that within those with COVID-19, a fraction m have died or die within the study time frame. Therefore, in sampling a random sample of S samples, we expect to find Sp positive cases, and Spm deaths. In terms of the sampled numbers, if we find C positive cases out of a total S sample size and D deaths, the estimates of the infection rate will be $p = C/S$ and the estimate of mortality rate $m = D/C$. These are unbiased estimates, and their conservative, guaranteed confidence interval can be calculated from the Clopper-Pearson interval[12].

An initial calculation expectedly suggested that with low infection rates, attaining a 5% error of estimation for the infection rate would require a moderate sample size. In contrast, if the real infection rate is higher (closer to 50%), expectedly, a smaller sample set will be sufficient for an accurate estimate of infection rate (Figure 1). Therefore, in the present scenario, an example of an ideal location where such tests could be performed with a limited number of sample size (approximately 10,000) is New York City, where the deaths have rapidly climbed up in the last week of March 2020.

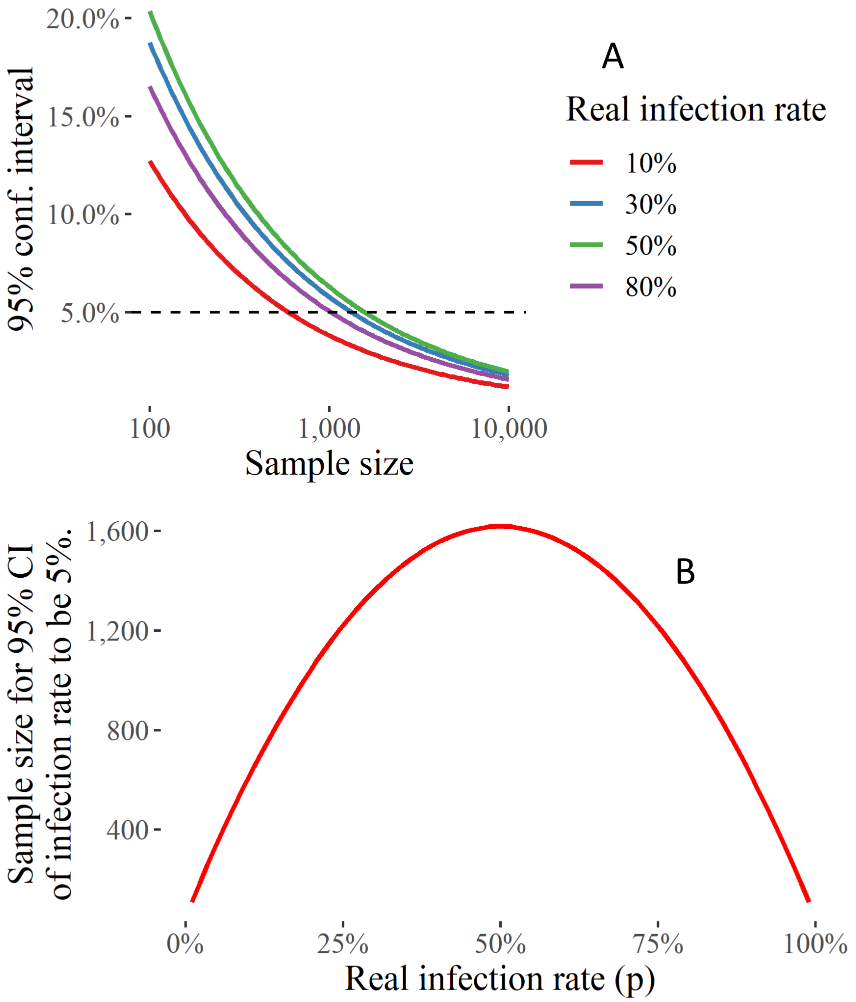

Figure 1: (A) The uncertainty (in terms of the 95% confidence interval) in the estimate of the fraction of population with COVID-19 (infection rate) with different sample sizes. (B) The sample size needed for infection rate 95% confidence interval to be 5%.

Estimating the mortality or infection fatality rate requires another probability to be multiplied to the estimate of infection rate within a sample population. A calculation of the sample size required for a 95% confidence interval indicates that even for a potentially highly infected population, like in NYC, it may require a very large sample size to accurately determine the true fatality rate (Figure 2). This may be one reason why countries have resorted to large sampling to obtain data for true fatality rate. However, biased and non-random sampling renders these data difficult to interpret to estimate the fatality rates.

We therefore propose to instead test the hypotheses that the true fatality rate is higher than a given value, which would be rejected if the upper limit of the 95% confidence interval is lower than the said value. Calculating these sample sizes with the statistically significant 95% confidence interval, we found that a relatively much smaller sample size would be sufficient to estimate if the true fatality rate is below or higher than a given percentage (Figure 3). Our calculations indicate that for a sample with 50% infection rate, a sample size of 1,000 may be sufficient to identify if fatalities are much lower than 1% , while for a sample with 25% infection rate, it may be below 10,000 — a logistically achievable size to determine a crucial parameter on which the world's health depends.

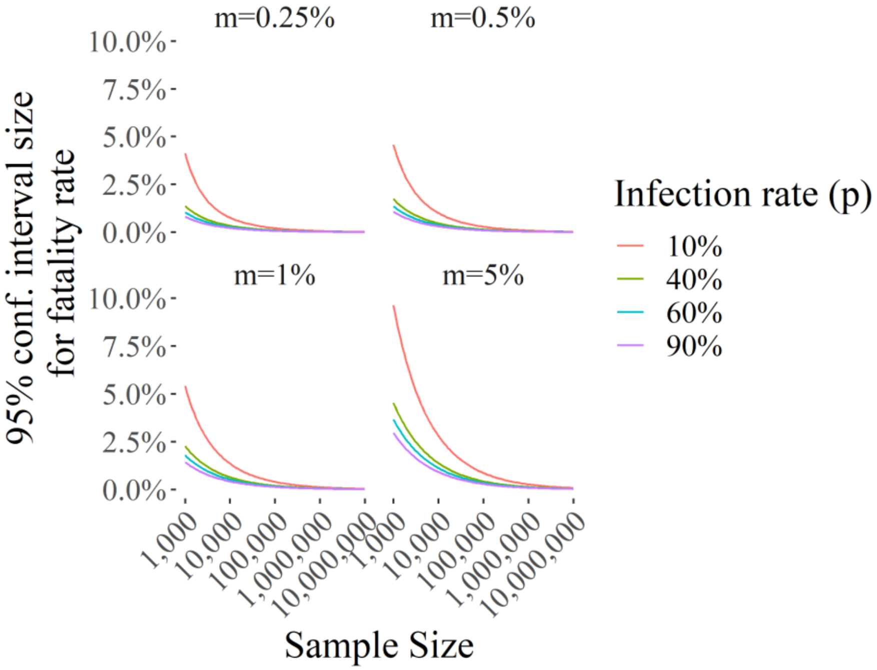

Figure 2: 95% Confidence Interval size for the death rate given a sample size (S, x-axis), infection rate (p, line colors), and the real death rate (m, subplot panel)

Continuous Sampling of a Selected Cohort can Provide Useful Dynamics on Acquirement of Immunity

The availability of serological test, if applied to a random and unbiased sampling could allow identification of a key subset of people who have developed immunity, but do not carry the infectious disease burden. However, it is usually not possible to have antibody tests available at the onset of a disease, and a rapidly spreading pandemic may make it difficult to gear policies based on accurate assessment of development of herd immunity. In contrast, recent development of genomic amplification or sequencing technologies has made it possible to prepare rapidly deployable tests to assess active infectious load. We therefore propose to use a continuous sampling of a representative unbiased cohort on a weekly basis to determine the initial onset of infection, the rate of its spread, development of herd immunity, and eventually the ensuing aftermath of the infection. Indeed, as we showed, for very small infection rate, a larger sample may be required. However, this concern is easily addressable for very small sample set by pooled sequencing (NGS), which can be used to determine rare onset, mutagenesis, and characterization of infections[13-15]. The dynamics of readout (of active viral load) in a fixed sample set will allow an accurate estimation of the development of immunity and its dynamics in a given population.

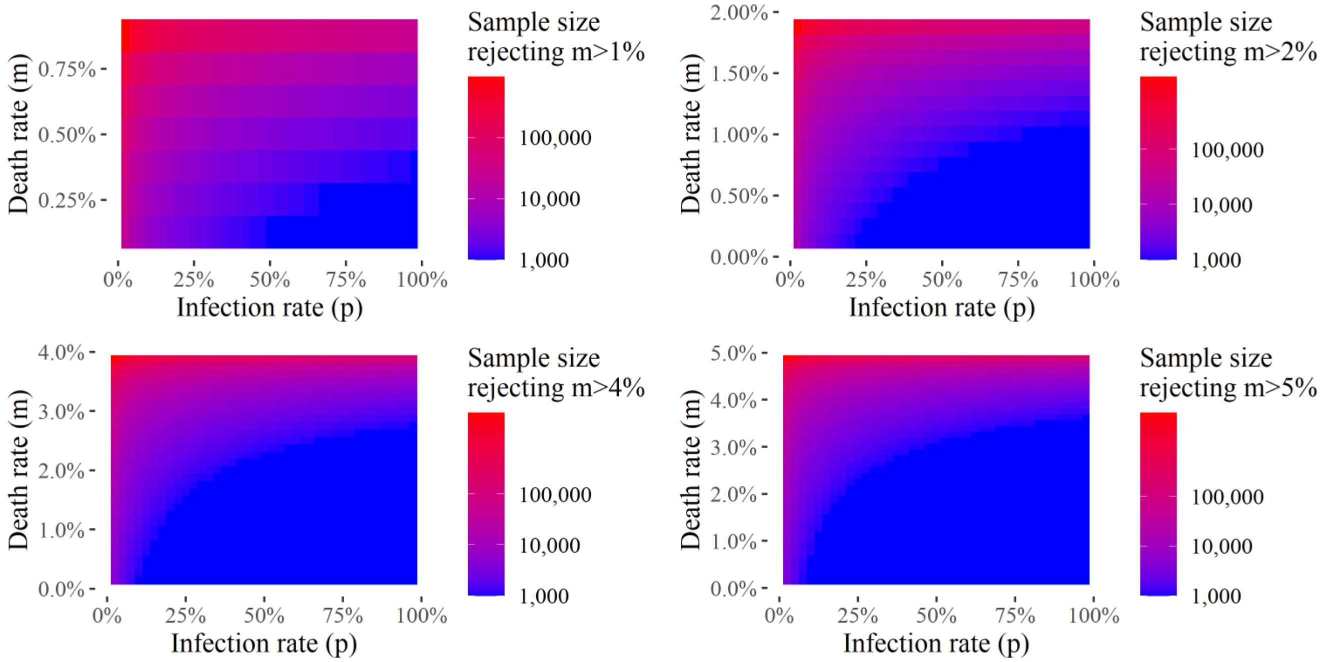

Figure 3: Sample sizes needed to reject hypotheses that infection death rate > than 1, 2, 4, or 5%.

CONCLUSION

The wide, and constantly updated, estimates in fatality rates associated with COVID-19 have resulted in questions being raised about the public health, societal, and economic policies adopted around the world. This is the most severe global health crisis to have inflicted humanity within this generation, although its true impact has still not been understood completely. Crucially, an accurate estimate of the true fatality rate is necessary to advocate for an apt public health policy, and we argue that current estimates have been lacking in this respect methodologically. Although much data is collected on the number of cases, the ensuing deaths, and those that have recovered, we believe that the interpretation of fatality and infection rates from non-uniform data across countries may be fraught with substantial inherent problems. Therefore, we recommend a limited, unbiased, random uniform sampling of population for fatality rate hypotheses testing using a combination of serological and genomic testing to determine the rate of viral infection, development of immunity, reduction in viral load, and resultant morbidity and fatality. We also propose a method to continually monitor a static sample set to estimate the onset, and dynamics of disease spread, acquired immunity, and ensuing morbidity and fatalities associated with an infectious spread. As a recent example, a large number of deaths in New York City could be explained either by a high fatality rate, or a rapid spread of the virus which has resulted in large number of people to develop immunity with a smaller percentage succumbing to the viral infection. In order to distinguish between the two widely varying scenarios, it is essential to accurately estimate the true rate of fatalities.

REFERENCES

1 WHO. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). (February, 2020).

2 Neil M Ferguson, D. L., Gemma Nedjati-Gilani,Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá,Gina Cuomo-Dannenburg,Amy Dighe, Ilaria Dorigatti, Han Fu, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley,Lucy C Okell, Sabine van Elsland, Hayley Thompson, Robert Verity, Erik Volz, Haowei Wang, Yuanrong Wang, Patrick GT Walker,Caroline Walters,PeterWinskill, CharlesWhittaker, ChristlADonnelly, Steven Riley, AzraCGhani. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf. *Imperial College COVID-19 Response Team* (2020).

3 Kshitiz. Should We have Locked the World Down for the COVID-19? https://medium.com/@kshitizkz/should-we-have-locked-the-world-down-for-the-covid-19-e0dc5191034c. (2020).

4 Verity, R. *et al.* Estimates of the severity of COVID-19 disease. *medRxiv*, doi:10.1101/2020.03.09.20033357 (2019).

5 P Spychalski, A. B.-S., J Kobiela. Estimating case fatality rates of COVID-19. *Lancet Infec Dis (2020) published online March 31. https://doi.org/10.1016/S1473-3099(20)30246-2*.

6 DD Kim, A. G. Estimating case fatality rates of COVID-19. *Lancet Infect Dis (2020) published online March 31. https://doi.org/10.1016/S1473-3099(20)30234-6*.

7 Lipsitch, M. Estimating case fatality rates of COVID-19. *Lancet Infect Dis (2020) published online March 31. https://doi.org/10.1016/S1473-3099(20)30245-0*.

8 Wu, Z. & McGoogan, J. M. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, doi:10.1001/jama.2020.2648 (2020).

9 Baud, D. *et al.* Real estimates of mortality following COVID-19 infection. *Lancet Infect Dis*, doi:10.1016/S1473-3099(20)30195-X (2020).

10 Team, C. C.-R. Preliminary Estimates of the Prevalence of Selected Underlying Health Conditions Among Patients with Coronavirus Disease 2019 — United States, February 12–March 28, 2020.

11 Group, C.-S. Characteristics of COVID-19 patients dying in Italy: report based on available data on March 20th, 2020. Rome, Italy: Instituto Superiore Di Sanita. (2020).

12 C. J. CLOPPER, P. THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL *Biometrika* **Volume 26**, Pages 404–413 (December 1934).

13 Greninger, A. L. *et al.* A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS One* **5**, e13381, doi:10.1371/journal.pone.0013381 (2010).

14 Sibley, C. D., Peirano, G. & Church, D. L. Molecular methods for pathogen and microbial community detection and characterization: current and potential application in diagnostic microbiology. *Infect Genet Evol* **12**, 505-521, doi:10.1016/j.meegid.2012.01.011 (2012).

15 Skums, P. *et al.* Computational framework for next-generation sequencing of heterogeneous viral populations using combinatorial pooling. *Bioinformatics* **31**, 682-690, doi:10.1093/bioinformatics/btu726 (2015).