medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license .

Estimation of the infection fatality rate and the total number of SARS-CoV-2 infections

Carlos Hernandez-Suarez[a,*], Paolo Verme[b], Efren Murillo-Zamora[c]

[a] *Facultad de Ciencias, Universidad de Colima, Bernal Diaz del Castillo 340, Colima, Colima, 28040, MEXICO*
[b] *World Bank, 1818 H St, NW Washington, DC 20433, USA*
[c] *Departamento de Epidemiología, Unidad de Medicina Familiar No. 19, IMSS, Av. Javier Mina 301, 28000, Colima, Colima. MEXICO*

Abstract

We introduce a simple methodology to estimate the infection fatality rate (IFR) and from here the total number of infected with SARS-CoV-2. The virus has shown to be highly infectious and thus we based our method under the assumption that all members of a household with at least one confirmed case of COVID-19 should be infected, therefore we estimate the IFR using the number of secondary fatalities in households. The simplicity of the methodology allows for large sample sizes, since it requires minimal laboratory testing capabilities.

Keywords: COVID-19, SARS-CoV-2, IFR, Asymptomatic, Immunity, Total Infections

1. Introduction

It is known that the immune response to SARS-CoV-2 may range from fully asymptomatic to exhibit mild or even severe responses that may cause death. Estimates of the probability of presenting a particular response are useful for prevention and attention purposes or even for building appropriate mathematical models that may provide some projections at the population level, specially to analyze the evolution of the immune population with the purpose

*Corresponding author
Email addresses: carlosmh@mac.com (Carlos Hernandez-Suarez), pverme@worldbank.org (Paolo Verme), efren.murilloza@imss.gob.mx (Efren Murillo-Zamora)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.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/10.1101/2020.04.23.20077446; this version posted May 30, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license .

of economic recovery. These estimates are particularly important to estimate the total number of infections by expanding the fraction of observed in some category, for the instance the number of hospitalized persons or the number of deaths.

Let p be the probability that an individual will die given that it is infected with SARS-CoV-2, that is, p is the infection fatality rate (IFR). If $\hat{p}$ is and estimate of p then we can build an estimate of the number of infections per every death as $1/\hat{p}$. If the total number of deaths M is known, one can estimate the total number of infections with $M/\hat{p}$.

There are current estimates of the probability of showing a specific reaction to infection, for instance, being asymptomatic, presenting mild or severe symptoms [1, 2, 3, 4], but their statistical properties are unknown. A possible design that would allow to estimate p is random screening for infection or antibodies, and categorizing the response of infected or already immune individuals. Some of these studies have been recently released for Iceland [**?**] and there are ongoing studies in other countries.

Here we suggest a simple study design based on the number of deaths observed in households with at least one confirmed case of COVID-19.

Methodology

Let's define an *effective contact* or *contact* for short as any act between an infectious and a susceptible individual that would result in the infection of the susceptible [5]. Let's suppose that have n individuals that we know had a *contact*. Then and estimate of the IFR is $\hat{p} = x/n$ where x is then number of observed deaths among the n individuals.

From here, the importance of finding individuals that we know had a *contact*. But these individuals may easy to find: several studies have suggested that household transmission as well as familial transmission is very high [6, 7, 8, 9, 10, 11, 12] or even in offices for relative short interactions [13]. Therefore, if we are willing to concede that all the members of a household with a diagnosed

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

individual had a *contact* with the initial infected in the household, the fraction of deaths among the remaining members of the household is an estimate of p. It is possible to pool data from several households to obtain a better estimate. In what follows, we formalize this estimate.

Suppose that we have a confirmed case of COVID-19. This confirmed case can lead us to household j with n_j members in total. Define the individual that led us to a household as the *index case* (not necessarily the first case in a household). Assume that:

(i) The remaining $n_j - 1$ members of household are infected with probability 1.

(ii) Once infected, the response to infection of each of the $n_j - 1$ individuals are independent, that is, the number of deaths among the remaining susceptible members in a household follows a binomial distribution with parameters $n_j - 1$ and p.

Observe that (i) implies that when two or more individuals are infected in the household, the probability that any one of the remaining susceptible will be infected is not increased. Also, it implies that all infected individuals are equally infectious, regardless of their symptomatic response to infection. Observe also that by excluding the *index case* of each household we avoid any bias.

Estimation of IFR and the total number of infections

Suppose a sample of m confirmed individuals led us to m households of size n_j, $j = 1, 2, 3, \ldots, m$. Let $n = \sum_j^m n_j$ be the sum of all members in all households in the sample. Let x_j be the number of deaths in household j (excluding all possible deaths of *index cases*) and let $x = \sum_j^m x_j$. The estimate of p, the IFR measured at the household level is $x_j/(n_j - 1)$. Using all households data in the sample, the estimate of p is:

$$\hat{p} = \frac{x}{n - m} \tag{1}$$

with variance $\hat{p}(1 - \hat{p})/(n - m)$.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

With one further assumption, one can estimate the number of infections for the total population from these same data. If we assume that the number of COVID-19 deaths recorded includes all deaths from COVID-19, we can simply estimate the number of infected people in the population by expanding the fraction of infected people estimated from the sample of observed households. This should provide a simple but statistically sound estimate of the total number of infected people in the population.

The estimate of the total number of infections per death is about $\theta = 1/\hat{p}$. The approximate variance of $\hat{\theta}$ is:

$$\mathrm{Var}(\hat{\theta}) = \frac{1-\hat{p}}{(n-m)\hat{p}^3}$$

Let M be the total number of deaths from COVID-19 in the population, the estimate of the total number of infected individuals in the population, N is:

$$\hat{N} = M\frac{(1-\hat{p})}{\hat{p}} + M = M/\hat{p} \tag{2}$$

with approximate variance:

$$\hat{\sigma}_N^2 = \frac{M^2}{n-m}\frac{1-\hat{p}}{\hat{p}^3} \tag{3}$$

It is important to stress that our model does not assume that all infections among the remaining $n_j - 1$ members of the household were caused by the same individual. In fact, our approach only requires that the remaining $n_j - 1$ individuals in the household have had enough infectious pressure to guarantee they are infected. Thus, it works even if one or more infections among the members in a household were caused outside the household.

Example

In this example we build an approximation to (1) using a database from Mexico's IMSS (Instituto Mexicano de Seguro Social), the Mexican Institute for Social Insurance.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

The database has 9939 confirmed SARS-CoV-2 cases from March 2 to May 4, 2020. In an attempt to consider only households with final outcomes we excluded cases with symptoms onset in the last 21 days, that is, we considered only cases from March 2 to April 19, 2020. The final dataset has 3232 cases.

We grouped the cases in households. If there were more than one case in a household, we only considered the household if all cases were already solved as deaths or recoveries. In every household with more than one case we consider the index case as the individual with the earliest symptom onset and counted the number of deaths among the remaining members of the house. From the final set of 3193 households, there were 3185 with no deaths among the remaining members of the household and 8 houses with one additional death. The mean age of this final set was 46.0 years with a standard deviation of 14.79 years with median 45 years. From these, there were 57.4 % males and 42.6 % females. In this set, 37 % were at least 50 years old.

The total number of households was $m = 3193$ and there were a total of $x = 8$ deaths. Since the total number of individuals in all households in the sample (n) i s not known, we vary the average household size in the sample (μ) to calculate $n = m\mu$ and estimate $\hat{p}$ using (1). The results are summarized in Figure 1.

Discussion

First we must mention that our goal here is not to provide precise estimates of p for Mexico since the total number of individuals in all households is not known and we used an approximation according the average household size. Our goal is to illustrate a simple methodology to estimate the true number of infections in a population using available information on confirmed individuals.

Our estimate from the IMSS data at the average household size $\mu = 3.7$ is $p = 0.00092$, which is 13 times smaller than the IFR for the *Diamond Princess* with IFR= 0.012 and mean age of 58 years [14] and about the same as the reported so far for the *USS Theodore Roosevelt*, with IFR= 0.001 with an evident

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

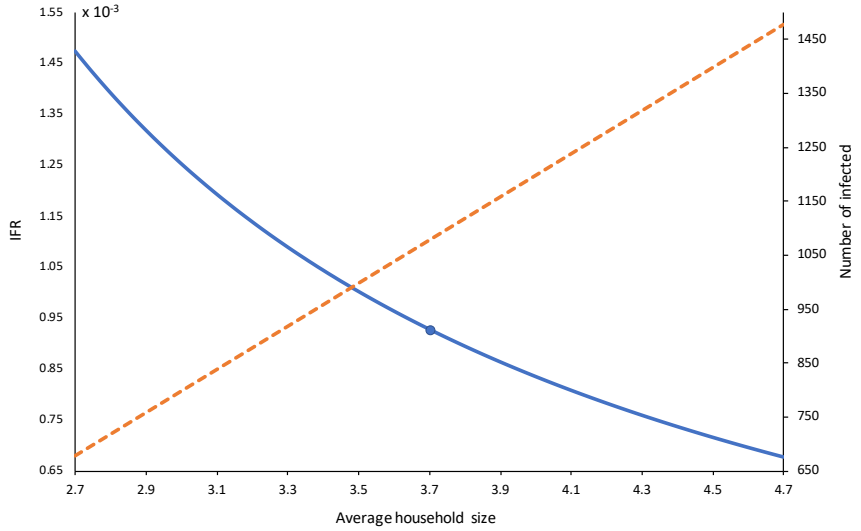

Figure 1: Plot of the average household size vs estimated IFR and K=1/IFR^{-1}, the ratio of the total infected to deaths. The point marks the IFR at the average family size for Mexico, 3.7 At this average, IFR= 0.00092 and $K = 1077$.

lower mean age [15]. In conclusion, we estimate one death per 1000 infected individuals.

Our method is simple enough to be applied in countries with relatively few tracking capabilities. All it is needed is a list of households with at least one confirmed case of COVID-19 (a sample may suffice) with the total number of members in the household and the number of deaths for COVID-19 in each household. The precision of estimate (1) depends on the sample size m, and the precision of estimate (2) depends in addition on how good is our estimate of the actual number of deaths from COVID-19 to date. Overall, the precision will depend on our ability to diagnose COVID-19 related deaths.

Assumption (i) is central for this proposal, but there is a way to avoid it although clearly at a larger economic cost: this consists in testing all the members of the household of a confirmed case. The estimate (1) can still be applied using only data of confirmed cases, but now x is the number of deaths among all confirmed cases in all households (excluding the *infected zero*) and

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

n the total number of confirmed cases in all households (including the *infected zero*).

In a following step, we can obtain the same probabilities for the whole population of positive cases by matching the household sample of tested households with households in the census. In other words, we only need to make sure that the sample of households retained from the interviews is representative of the national sample of households. This can be done, *ex-ante* with a sample of available infected households or, if this information is not available, *ex-post* by matching the interviewed sample of households with the national census of households. Something that can be done with matching or machine learning methods. This provides the distribution of cases between any categorization of symptoms for the population of infected people in a population. A direct approach from stratified sampling may use some demographic knowledge of the population which would allow us to weight for differential response to the infection. Suppose that we classify a population in K categories (e.g., age) at relative frequencies f_i. Let $x^{(i)}$ and $n^{(i)}$ be respectively the total number of deaths and total number of individuals in category i in all households in the sample of size m, then a better estimate of p would be:

$$\hat{p} = \sum_{i=1}^{K} f_i \theta_i, \qquad \text{with } \theta_i = \frac{x^{(i)}}{n^{(i)} - m} \tag{4}$$

with variance

$$\hat{p} = \sum_{i=1}^{K} f_i^2 \; \frac{\theta_i}{n^{(i)} - m} \tag{5}$$

This $\hat{p}$ must be plugged in (2), with variance (3). We can divide then population in Mexico in two categories: age ≤ 50 years and age > 50 years, at respective proportions $f_1 = 0.9$ and $f_2 = 0.1$ [16]. The IFR in the first category was 0.002 and in the second 0.0052. From (4) we have $\hat{p} = 0.0023$ for the whole population, the weighted estimate suggests the number of total infected is about 400 times larger than the number of deaths.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

One of the most important sources of bias in this method, is that some observations may be censored. Perhaps death has not occurred yet in a given household and thus the probability of death is underestimated. In our analysis of IMSS data, we tried to control this by using only data where the onset of symptoms was at least 21 days old so that the outcome is very likely observed, but in principle, we should use households were there is enough evidence to believe that we can observe final outcomes.

Conflict of interest

Authors declare no conflict of interest.

Funding

This work is part of the program "Building the Evidence on Protracted Forced Displacement: A Multi-Stakeholder Partnership". The program is funded by UK aid from the United Kingdom's Department for International Development (DFID), it is managed by the World Bank Group (WBG) and was established in partnership with the United Nations High Commissioner for Refugees (UNHCR). The scope of the program is to expand the global knowledge on forced displacement by funding quality research and disseminating results for the use of practitioners and policy makers. This work does not necessarily reflect the views of DFID, the WBG or UNHCR. This study had approval R-2020-601-07 by the Health Research Ethics Committee (601) of the IMSS.

References

[1] Y. Liu, L.-M. Yan, L. Wan, T.-X. Xiang, A. Le, J.-M. Liu, M. Peiris, L. L. Poon, W. Zhang, Viral dynamics in mild and severe cases of covid-19, The Lancet Infectious Diseases.

[2] K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

the diamond princess cruise ship, yokohama, japan, 2020, Eurosurveillance 25 (10) (2020) 2000180.

[3] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, et al., Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19), medRxiv.

[4] J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, G. M. Leung, Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china, Nature Medicine 26 (4) (2020) 506–510. doi:10.1038/s41591-020-0822-7.
URL https://doi.org/10.1038/s41591-020-0822-7

[5] F. Brauer, C. Castillo-Chavez, Z. Feng, Mathematical Models in Epidemiology, Springer, 2019.

[6] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang, Presumed asymptomatic carrier transmission of covid-19, Jama.

[7] J. F.-W. Chan, S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C.-Y. Yip, R. W.-S. Poon, et al., A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, The Lancet 395 (10223) (2020) 514–523.

[8] Z. Hu, C. Song, C. Xu, G. Jin, Y. Chen, X. Xu, H. Ma, W. Chen, Y. Lin, Y. Zheng, et al., Clinical characteristics of 24 asymptomatic infections with covid-19 screened among close contacts in nanjing, china, Science China Life Sciences (2020) 1–6.

[9] P. J. Lillie, A. Samson, A. Li, K. Adams, R. Capstick, G. D. Barlow, N. Easom, E. Hamilton, P. J. Moss, A. Evans, et al., Novel coronavirus disease (covid-19): the first two patients in the uk with person to person transmission, Journal of Infection.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.23.20077446; this version posted May 30, 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-NC-ND 4.0 International license .

[10] X. Pan, D. Chen, Y. Xia, X. Wu, T. Li, X. Ou, L. Zhou, J. Liu, Asymptomatic cases in a family cluster with sars-cov-2 infection, The Lancet Infectious Diseases 20 (4) (2020) 410–411.

[11] G. Qian, N. Yang, A. H. Y. Ma, L. Wang, G. Li, X. Chen, X. Chen, A covid-19 transmission within a family cluster by presymptomatic infectors in china, Clinical Infectious Diseases.

[12] P. Yu, J. Zhu, Z. Zhang, Y. Han, A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period, The Journal of infectious diseases.

[13] C. Rothe, M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, T. Zimmer, V. Thiel, C. Janke, W. Guggemos, et al., Transmission of 2019-ncov infection from an asymptomatic contact in germany, New England Journal of Medicine 382 (10) (2020) 970–971.

[14] T. W. Russell, J. Hellewell, C. I. Jarvis, K. Van-Zandvoort, S. Abbott, R. Ratnayake, S. Flasche, R. M. Eggo, A. J. Kucharski, C. nCov working group, et al., Estimating the infection and case fatality ratio for covid-19 using age-adjusted data from the outbreak on the diamond princess cruise ship, medRxiv.

[15] New York Times. Sailor on Roosevelt, whose captain pleaded for help, dies from coronavirus [online] (April 13, 2020) [cited April 22,2020].

[16] Encuesta Intercensal, INEGI, Recovered from: http://www. beta. inegi. org. mx/proyectos/enchogares/especiales/intercensal.