CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

ARTIGO
ARTICLE

Estimated prevalence of COVID-19 in Brazil with probabilistic bias correction

Prevalência estimada de COVID-19 no Brasil com correção probabilística de vieses

Prevalencia estimada de COVID-19 en Brasil con corrección probabilística de sesgo

Erik Alencar de Figueiredo [1]
Démerson André Polli [2]
Bernardo Borba de Andrade [2]

doi: 10.1590/0102-311X00290120

Abstract

Using data collected by the Brazilian National Household Sample Survey – COVID-19 *(PNAD-COVID19) and semi-Bayesian modelling developed by Wu et al., we have estimated the effect of underreporting of COVID-19 cases in Brazil as of December 2020. The total number of infected individuals is about 3 to 8 times the number of cases reported, depending on the state. Confirmed cases are at 3.1% of the total population and our estimate of total cases is at almost 15% of the approximately 212 million Brazilians as of 2020. The method we adopted from Wu et al., with slight modifications in prior specifications, applies bias corrections to account for incomplete testing and imperfect test accuracy. Our estimates, which are comparable to results obtained by Wu et al. for the United States, indicate that projections from compartmental models (such as SEIR models) tend to overestimate the number of infections and that there is considerable regional heterogeneity (results are presented by state).*

Herd Immunity; Selection Bias; Quantitative Analysis

Correspondence
B. B. Andrade
Universidade de Brasília.
Campus Universitário Asa Norte – Prédio CIC/EST, Brasília, DF 70910-900, Brasil.
bbandrade@unb.br

[1] *Universidade Federal da Paraíba, João Pessoa, Brasil.*
[2] *Universidade de Brasília, Brasília, Brasil.*

Introduction

The *Brazilian National Household Sample Survey – COVID-19* (PNAD-COVID19) [1] is a nationwide, complex, survey aimed at "*estimating the number of persons with symptoms associated with the flu syndrome and at following up the impact of the COVID-19 pandemic in the Brazilian labor market. The data collection of the Brazilian National Household Sample Survey – PNAD COVID-19 began on May 4, 2020, including interviews by telephone in nearly 48,000 households per week, adding up to nearly 193,000 households per month in the entire country. The sample is permanent, i.e., the households interviewed in the first month of data collection will remain in the sample along the next months, up to the end of the survey*". Considering its latest release, in December 2020 (survey end date: December 5th), the survey indicates a total of 22.7% positive results for COVID-19, that is, 3.1% of the Brazilian population has tested positive. Official tallies by the Brazilian Ministry of Health are also at 3.1%. Two states are above 7% (confirmed) infection rate: states of Roraima and Amapá both in the North Region of Brazil. Most states have rates between 3% and 5%. According to our results presented by state in the section *Results from Probabilistic Bias Analysis*, the prevalence of COVID-19 in Brazil at the end of 2020 is estimated to be around 15% of the population, slightly more than 30 million individuals.

The number of confirmed cases is well known as being only a fraction of the actual number of infected individuals. Firstly, testing is not always available and most moderate or asymptomatic cases go undetected even when testing is accessible. Secondly, COVID-19 testing in Brazil has been carried out with a several testing kits and survey respondents may have used procedures with high rates of false negatives. Finally, and somewhat related to the first issue, there is the problem of selection bias as only those with stronger symptoms will seek medical attention and testing [2,3,4,5]. Some earlier studies have indicated that the true number of cases was about 12 times the number of confirmed cases [6,7]. The September issue of *The Economist* [8] publicized some estimates for the share of the population with COVID-19 antibodies obtained by serosurveys in June and the implied ratio of cases to confirmed cases: it was reported that Moscow (Russia) could -have 27 times more cases than reported ones, Stockholm (Sweden) 17 times, London (England) 14 times, Madrid (Spain) 10 times, and New York City (United States) 7 times. More recent estimates from Wu et al. [9] suggest a ratio of nine times for the United States and that the ratio would vary from three to 20 across the 50 states in the country. In late 2020, studies from the São Paulo University [10] and the Federal University of Pelotas [11], in Brazil, indicated a much higher share of antibodies in the population, as high as 60%, close to standard herd immunity thresholds, but recent increases in cases across the country suggest that antibodies may not be as prevalent. In any event, common sense and empirical evidence strongly suggest that the actual number of infected individuals is much higher than the number of confirmed (cumulative) cases.

For estimating the prevalence of COVID-19 in Brazil at a given point in time we can resort to at least three major approaches. Firstly, there are simple analyses using data from other years on similar respiratory diseases and some demographical considerations: for instance, Ribeiro et al. [7] concluded in May 2020 that 3.8 times more hospitalizations in Brazil due to COVID-19 were identified than reported by analyzing hospitalization patterns of acute respiratory distress syndrome between 2012 and 2019 as compared to 2020.

Another approach is based on dynamic mathematical models, such as SEIR models, which have been made widely available for COVID-19. Such models yield projections that are higher than what is observed as illustrated by the compilation of four popular models in OurWorldInData.org [12]: recent estimates regarding SARS-CoV-2 from the SEIR model by Youyang Gu (YYG) indicate that Brazil has reached 16.9% infection rate, eight times the confirmed cases. The model from the London School of Hygiene & Tropical Medicine (LSHTM) estimates that between 28% and 42% of symptomatic cases are unreported. The models by the Imperial College London (ICL) and the Institute for Health Metrics and Evaluation (IHME) place estimates at much higher values, especially around June.

A third approach, adopted by Wu et al. [9] who developed a bias correction model for the estimation of COVID-19 cases, is to perform a quantitative bias analysis [13]. This approach is entirely data-based and does not aim to model transmission mechanisms or dynamics. Quantitative bias analysis aims to quantify the effects of systematic error (due to selection bias, unmeasured confounders, information bias, etc.) on estimates derived from nonrandomized epidemiologic studies.

The next section briefly describes the approach of bias correction developed by Wu et al. [9] and reproduced in this study. Section *Results from Probabilistic Bias Analysis* provides the results we obtained for Brazil using the national survey mentioned above (PNAD-COVID19) and Section *Discussion* concludes with a summary and final remarks briefly discussing the related issue of herd immunity.

Methods

Probabilistic bias analysis by Wu et al.

Along the lines of probabilistic bias analysis [13], Wu et al. [9] have developed a simulation-based bias correction method aimed to adjust count estimates on confirmed cases for selection bias (preferential testing of moderate/severe cases) and imperfect test accuracy. Despite additional computational cost, empirical, simulation-based methods provide the flexibility needed for the kind of multiple correction desired. In this study, the simple estimate based on confirmed cases is biased away from the true value due to preferential testing and imperfect test accuracy. The parameters affecting the distributions used to correct for bias are treated as random variables and, hence, the procedure is known as probabilistic bias analysis. Even though the only modification we have made to the method proposed by Wu et al. [9] is the selection of (hyper-)parameter values in the prior models (their specifications reflect the U.S. reality), we explain their method for completeness.

More specifically, we want to estimate N^* which is the number of cases in the population (for each of the 27 Brazilian states, including the Federal District). The starting point, which is reported in surveys or official reports, is N_T^+, the number of confirmed cases among tested individuals which we identified to be just a fraction of N^* due to selection bias and imperfect test accuracy. These two figures are connected by an epidemiological identity which provides a correction for imperfect test accuracy [13],

$$N^* = \frac{N^+ - (1 - S_p) \text{ x } N}{S_e + S_p - 1},$$

where: N is the population size (known), $N_{T^c}^+$ is the number of infected individuals not tested, $N^+ = N_T^+ + N_{T^c}^+$, and S_e and S_p are test sensibility and specificity, respectively.

The value of $N_{T^c}^+$ is unknown. It may be obtained as the sum of the number of untested individuals who have moderate or severe symptoms and would result positive if tested,

$$N_{T^c,S_1}^+ = P(S_1|T^c) \text{ x } P(+|S_1,T^c) \text{ x } N_{T^c},$$

and the number of untested individuals who have mild or no symptoms and would result positive if tested,

$$N_{T^c,S_1}^+ = P(S_1|T^c) \text{ x } P(+|S_1,T^c) \text{ x } N_{T^c},$$

The above expressions provide correction for incomplete testing. Following Wu et al. [9], we considered them to be binomially distributed with size N_{T^c} and success probability equal to $P(+,S_j,T^c) = P(+|S_j,T^c) \text{ x } P(S_j|T^c) \text{ x } P(T^c)$, $j = 0$ or 1. In simulations, these two quantities are held fixed at their mean values since their variability is negligible compared to other sources of uncertainty and the population size is large [9]. However, the probabilities involved in the above expressions are unknown. They are modelled and simulated: thus the set of parameters for which prior information must be provided contains the probability of having moderate to severe symptoms among tested individuals, $P(S_1|T)$ and also $P(S_1|T^c)$, and the probability of having mild symptoms among positive cases, $P(S_0|+)$. It also contains the sensibility and the specificity of testing procedures used for COVID-19 and the ratios α and β which refer, respectively, to $P(+|S_1,T^c)$ and $P(+|S_0,T^c)$ divided by $P(+|T)$.

A probabilistic identity is used to coherently connect these probabilities,

$$P(S_0|+) = \frac{\beta(1 - P(S_1|T^c))}{\beta(1 - P(S_1|T^c)) + \alpha P(S_1|T^c)}$$

where the ratios α and β have been defined above. Because priors are specified on both sides of the equation above, a sample for the vector $(P(S_0|+), \alpha, \beta, P(S_1|T^c))$ is obtained by a technique known as Bayesian melding [14]. Bayesian melding is a procedure that uses the fact that one parameter can be expressed as a deterministic function of other parameters. It can be useful to simplify complex models when a deterministic relation is present. In this sense, it is enough to define the prior for one of them with the other being fully determined by the deterministic relation among them.

Simulations

Considering the above framework, seven quantities are present in the probabilistic bias analysis just described whose uncertainties must be assessed with simulations.

For each state, the empirical estimate $P(+|T)$ (cumulative number of cases divided by the cumulative number of tests) from the latest release of PNAD-COVID19 is fixed. The ratio of tests performed to the state population ranged from 9% to 26% across the 27 states and the point estimates of positive rates range from 18% to 50 across states.

With those probabilities fixed the relevant quantities are simulated and a distribution of expected cases is obtained for each state as described next. Table 1 shows all parameters modelled as truncated beta random variables such that their moments and bounds, agreeing with estimated values in the survey (PNAD-COVID19) or test kit specifications obtained from the Brazilian Ministry of Health.

Finally, a decomposition of the two sources of biases, incomplete testing and imperfect test accuracy, can be obtained through

$$p_1 = \frac{N^* - N^+}{N^* - N^+_{T^c}},$$

and $p_2 = 1 - p_1$, where p_1 is attributable to the inaccuracies of testing and its complement to incomplete testing.

Initially, 10^4 values are sampled from the distributions of $P(S_1|T)$, $P(S_1|T^c)$, α, β, S_e and S_p with these six variables independent and identically distributed across states. Then, values of $P(S_1|T^c)$ and $P(+|S_0,T^c)$ are simulated based on $P(+|T^c)$ and simulated values of α and β. Despite the theoretical possibility that some parameters may be correlated between some states, we and Wu et al. [9] do not have robust evidence to inform an appropriate model of correlation structure.

Table 1

Prior specifications for truncated beta models used in the probabilistic bias analysis.

Parameter	Minimum	Mean	Maximum	SD
$P(S_1 \mid T)$	0.0000	0.5000	1.0000	0.2887
$P(S_1 \mid T^c)$	0.0000	0.1500	0.3000	0.2000
α	0.7000	0.9000	1.0000	0.2000
β	0.0020	0.2000	0.5000	0.4000
$P(S_0 \mid +)$	0.2500	0.7000	0.9000	0.4000
Sensibility (S_e)	0.6500	0.8500	1.0000	0.3000
Specificity (S_p)	0.9800	0.9995	1.0000	0.0100

SD: standard deviation.

In Wu et al. [9], and in this study, Bayesian melding is used to relate the components of $(P(S_0|+), \alpha, \beta, P(S_1|T^c))$. As stated by Poole & Raftery [15] the use of pooling weight to be ½ makes the combined prior distribution of $P(S_0|+)$ and $P(S_1|T^c)$ to be the geometric mean of each prior distribution. We do not have any evidence to support any of the prior distributions, therefore using ½ as pooling weight is the natural choice and it is employed in this study. Bayesian melding is performed with 10^5 iterations of sampling-importance-resampling algorithm (SIR); a simulation size greater than 10^4, since this stage involves a more complex generator as opposed to independent univariate sampling. These simulation sizes were also used by Wu et al. [9].

After all samples are simulated, point estimates are obtained as sample medians. The analysis does not involve a probability and only the quantities of interest are sampled. No sampling of likelihood parameters was necessary.

Before we report our results based on the methodology described, we argue that a suitable way to estimate the number of infected individuals is to try to emulate a natural experiment based on the data provided by PNAD-COVID19. Many employers and local governments have requested mass testing for certain groups of individuals. We took a subsample from the national survey considering only working individuals, aged from 18 to 60 years old who have declared absence of previous COVID-19-like symptoms. This should partially eliminate selection bias. By calculating the percentage of those individuals who tested positive we reached 14.3%.

Results from probabilistic bias analysis

We will report our results in terms of percentage of the population infected by COVID-19 (prevalence),

$$\textit{Estimated Infection Rate} = \textit{Estimated Number of Cases/Population},$$

and the associated correction factor F,

$$F = \textit{Estimated Number of Cases/Confirmed Cases.}$$

Figure 1 shows our main results, where the estimated percentage of cases decomposed into confirmed cases (blue bar) and unreported cases (orange part) can be identified. Table 2 brings the corresponding credibility intervals. Northern states have higher number of estimated cases. Note that, reported cases account for 8% of the population of the state of Roraima. After bias corrections this value increases to 28%. Amazonas was one of the states with the highest number of infected individuals. However, the local government has reported that only 5% of its population was infected with COVID-19 (official cases). A recent study [10] suggests that the prevalence of COVID-19 antibodies is much higher based on donated blood sample, as high as 60%. We estimate it to be around 21%.

Figure 2 shows our results in terms of correction factors (F) including lower and upper bounds (2.5% and 97.5% quantiles). For instance, the state of Ceará has, according to these estimates, 4.3 times more cases than officially reported within an interval 3-5.3.

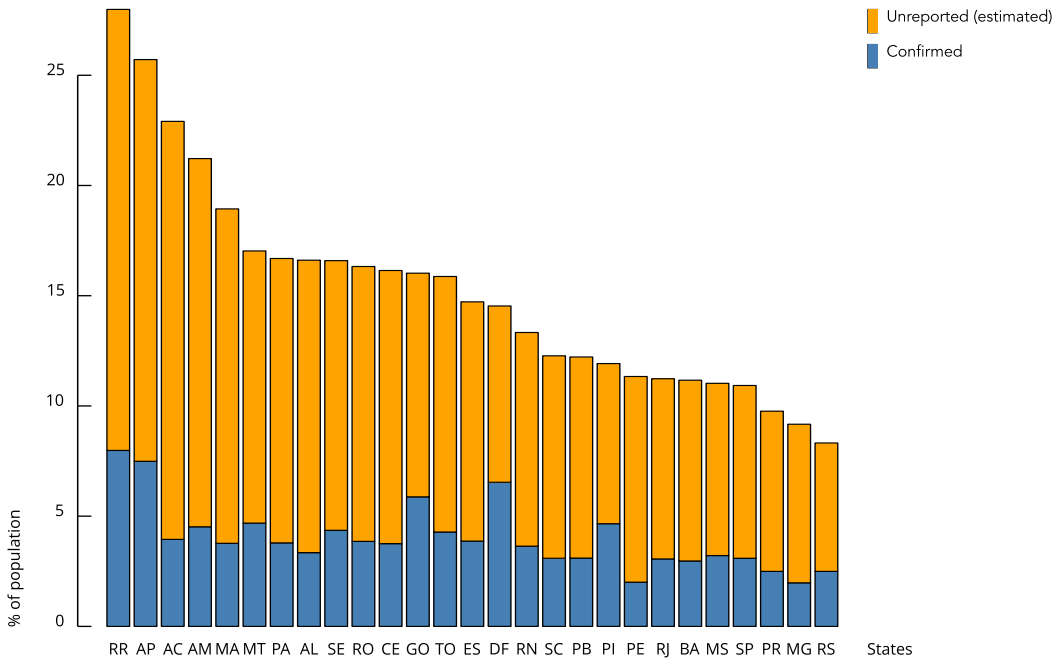
Finally, our results indicate a strong linear association between correction factors and testing coverage in the log-log scale (Figure 3), which can be written as:

$$\mathrm{F} \approx 0.671 \left(\frac{\#\text{tests}}{\text{population size}} \right)^{-0.867}.$$

Therefore, we can estimate F for a given region, say a large city, which has not been directly targeted by the survey. For instance, if 5% of the population has been tested then we estimate that F ≈ 9.0; for 10% coverage we estimate F ≈ 4.9 and for 20% testing we should obtain F ≈ 2.7.

Figure 1

Confirmed cases and estimated percentage of infected individuals by Brazilian state.



States: AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceará; DF: Federal District; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; RR: Roraima; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

Discussion

Estimating the number of infected individuals during a pandemic is essential to decision-makers and researchers. National surveys such as PNAD-COVID19 have been a significant source of information to adjust official counts of reported cases for different biases. Our use of the simulation methods developed by Wu et al. [9] show that Brazilian official estimation is about five times lower than the expected number of cases. It must be understood that this number is static and represents the magnitude of underreporting in August and that this number is likely to have been higher in the peak of the pandemic. We have presented national and state figures but for practical policy use, local (municipality level) estimates may be necessary.

The city of Manaus is located in Brazil and is the largest city in the state of Amazonas which accounts for slightly more than half of the state's population. The COVID infection in this region presented two peaks (with peaks in late July/2020 and in mid-March/2021) and has thus attracted much attention as an example of a largely unmitigated epidemic. A recent study [10] has estimated that, by October of 2020, the cumulative incidence in Manaus was 76% (95%CI: 67% to 98%), whereas our estimate for the state stands at 21% (95%CI: 15% to 26%). The fact that, in March, the city experienced a stronger second peak suggests that the 76% estimate is, most likely, unreliable and too high. Considering such high infection rate, a strong second peak would have to be assigned to the new variant found in Manaus (P.1 lineage) and to antibody waning. Reinfection would have then been very common which, as far as we know, has not been well documented. Despite these possibilities we still find the 76% estimate to be too high. Buss et al. [10] have indicated that their results rely on a certain

Table 2

Estimated prevalence and corresponding 95% credibility bounds.

State	Prevalence	Lower 95%	Upper 95%
Roraima	28.0	20.1	34.0
Amapá	25.7	18.6	31.2
Acre	22.9	15.5	28.6
Amazonas	21.2	14.7	26.2
Maranhão	18.9	13.0	23.5
Mato Grosso	17.0	12.2	20.7
Pará	16.7	11.6	20.5
Alagoas	16.6	11.4	20.6
Sergipe	16.6	11.8	20.2
Rondônia	16.3	11.4	20.0
Ceará	16.1	11.3	19.8
Goiás	16.0	12.0	19.1
Tocantins	15.9	11.3	19.3
Espírito Santo	14.7	10.5	18.0
Federal District	14.5	11.4	16.9
Rio Grande do Norte	13.3	9.5	16.2
Santa Catarina	12.3	8.7	15.0
Paraíba	12.2	8.6	14.9
Piauí	11.9	9.1	14.1
Pernambuco	11.3	7.7	14.1
Rio de Janeiro	11.2	8.0	13.7
Bahia	11.2	7.9	13.6
Mato Grosso do Sul	11.0	8.0	13.4
São Paulo	10.9	7.8	13.3
Paraná	9.8	6.9	11.9
Minas Gerais	9.2	6.3	11.3
Rio Grande do Sul	8.3	6.0	10.1

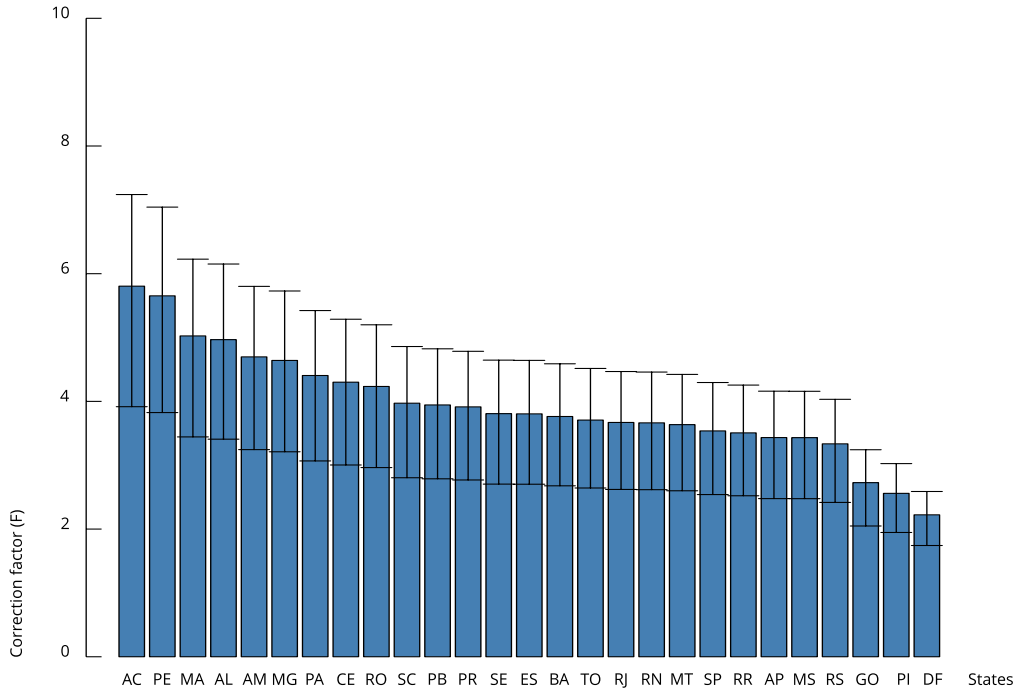
assumption about the dynamics of seroreversion and that their sample is not a random sample but, rather, a sample of blood donors which they argue to be representative of the population of Manaus. Notably, without adjusting for seroreversion, Buss et al. [10] report an estimated prevalence, adjusted for specificity and sensitivity, of 26% (95%CI: 21% to 31%) which is close to what is reported in this study. The disparity appears after adjustment for seroconversion. Whether this adjustment is reliable is open for research it certainly suggests that the estimates around 20-30% are too conservative. Similar arguments may be applied to the case of the state of São Paulo. Buss et al. [10] report an estimated prevalence of 29% (95%CI: 26% to 37%) adjusting for seroconversion and 14% (95%CI: 11% to 17%) without adjustment, for the capital city which accounts for slightly less than 30% of the state's population. Our state estimate for the state of São Paulo is 11% (95%CI: 7.8% to 13%). We conclude that our estimates, or any estimate not adjusted for seroconversion, must be considered conservative.

Since the prevalence (or a lower bound for it) has been estimated, it is natural to ask about herd (or collective) immunity, the level at which contagion becomes under control (herd immunity threshold – HIT). Despite lacking a precise definition, the concept of herd immunity is inevitable in discussions about COVID-19 and infectious diseases in general.

The possibility that collective immunity is not applicable to SARS-CoV-2 could be based on the epidemiology of the coronavirus HCoV-NL63 [16] for which long term individual immunization is not achieved and reinfection is common. However, our understanding of the epidemiological literature [17] and of expert opinions available in the media is that collective immunity is applicable to COVID-19. Nonetheless, the actual value to be targeted has been a topic for discussion.

Figure 2

Estimated correction factors for Brazilian states.

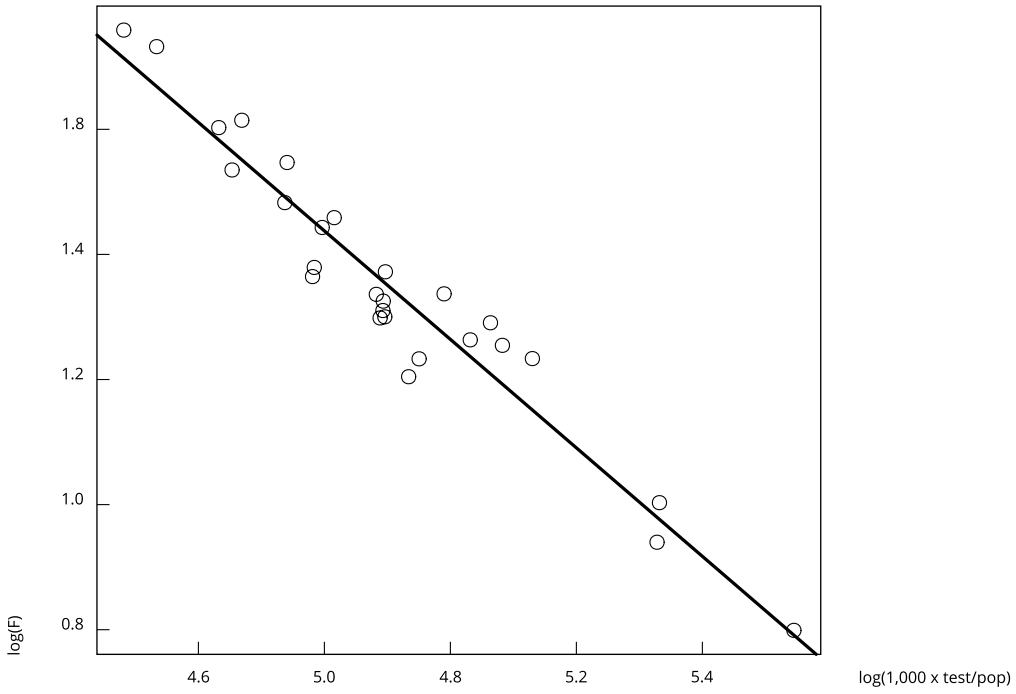


States: AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceará; DF: Federal District; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; RR: Roraima; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

The standard formula for the threshold is $1 - (1/R_0)$, where R_0 is the number of persons infected, on average, by a given individual carrying the virus. It assumes uniform susceptibility and it results from an idealized scenario. For COVID-19 it has been argued that, approximately, $R_0 = 3$, and thus HIT = 67%. Such calculations have been questioned in the epidemiological literature and, regarding SARS-CoV-2, Aguas et al. [18] have argued that random immunization is far from reality and considerations about non-uniform susceptibility considering the individual coefficient of variation (CV) would lead to a more realistic formula: in an epidemic that takes its natural course, by contrast, the virus very specifically infects the people that are most susceptible first. This removes all of the strongest vectors early on, and continues to selectively remove the vectors until the herd immunity threshold is reached. The new parameter, CV, plays an opposite role than that of the reproduction number. The larger the CV the lower the HIT. It is proposed by Aguas et al. [18] that $\mathrm{HIT} = 1 - (1/R_0)^{1/(1+CV^2)}$. Just like R_0, the CV needs to be estimated. Setting aside (relevant) discussions about the validity of proposed values for de R_0 and CV for SARS-CoV-2, if we consider $R_0 \approx 2.5$ and recent estimates of the CV for SARS-CoV-1 for Singapore and Beijing (China), that is CV ≈ 2.6, we obtain HIT = 11%. For comparison, CV values for malaria in the Amazon and tuberculosis in Brazil are 1.8 and 3.3 respectively. First versions of Aguas et al. [18] concluded that HIT values could be much lower than 50% in many cases but this seems to be valid without mutations and other changes in the dynamics of the epidemic as can be observed throughout the world with cases still on the rise in places where infection has passed the 10%. Nevertheless, the work of Aguas et al. provides an important discussion in the ongoing research on HIT.

Figure 3

Approximate linear relationship (log-log) between correction factors and testing coverage.



Another source of uncertainty must be addressed before answering the herd immunity question. The question of pre-existing immunity [19]. An aspect much harder to be measured. We thus expect more research to be conducted along these lines, not only regarding the best expression for HIT but also on best estimates for R_0 and CV and more insights into pre-existing immunity.

If we assume that the HIT for SARS-CoV-2 is in fact less than 67% (due to the effect of CV and pre-existing immunity), say 50%, then, considering the estimated prevalence of 8% to 28% (Figure 2; Table 2), most of Brazil is still some time away from achieving some sort of collective immunity but not too far if vaccination efforts are successful. For HIT around 50% some states are half way towards herd immunity but for HIT around 60% to 70% most states would be less than halfway in reaching the threshold.

Contributors

All authors participated in the literature review, PNAD-COVID19 data processing, code adaptation, and manuscript writing.

Additional informations

ORCID: Erik Alencar de Figueiredo (0000-0002-3479-3665); Démerson André Polli (0000-0002-5904-2315); Bernardo Borba de Andrade (0000-0003-4688-9733).

Acknowledgments

We thank the revisors for their insightful comments and careful revisions. All remaining errors and imprecisions are the responsibility of the authors.

References

1. Instituto Brasileiro de Geografia e Estatística. PNAD-COVID19: informativo para a mídia. https://www.ibge.gov.br/en/statistics/social/health/27975-weekly-release-pnadcovid1.html?=&t=o-que-e (accessed on 02/Oct/2020).
2. Pearce N, Vandenbroucke JP, VanderWeele TJ, Greenland S. Accurate statistics on COVID-19 are essential for policy guidance and decisions. Am J Public Health 2020; 110:949-51.
3. Lan L, Xu D, Ye G, Xia C, Wang S, Li Y, et al. Positive RT-PCR test results in patients recovered from COVID-19. JAMA 2020; 323:1502-3.
4. Yang Y, Yang M, Shen C, Wang F, Yuan J, Li J, et al. Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. medRxiv 2020; 17 feb. https://www.medrxiv.org/content/10.1101/2020.02.11.20021493v2.
5. Angrist J, Pischke J-S. Mostly harmless econometrics: an empiricist's companion. Princeton: Princeton University Press; 2009.
6. Rahmandad H, Lim TY, Sterman J. Estimating COVID-19 under-reporting across 86 nations: implications for projections and control. SSRN 2020; 1 jul. https://ssrn.com/abstract=3635047.
7. Ribeiro LC, Bernardes AT. Nota técnica: atualização da estimativa de subnotificação em casos de hospitalização por síndrome respiratória aguda e confirmados por infecção por Covid-19 no Brasil e estimativa para Minas Gerais. Belo Horizonte: Centro de Desenvolvimento e Planejamento Regional, Universidade Federal de Minas Gerais; 2020.
8. The covid-19 pandemic is worse than official figures show. The Economist 2020; 26 sep. https://www.economist.com/briefing/2020/09/26/the-covid-19-pandemic-is-worse-than-official-figures-show.
9. Wu SL, Mertens AN, Crider YS, Nguyen A, Pokpongkiat NN, Djajadi S, et al. Substantial underestimation of SARS-CoV-2 infection in the United States. Nature Commun 2020; 11:4507.
10. Buss LF, Prete Jr. CA, Abrahim CM, Mendrone A, Salomon T, Almeida-Neto C, et al. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. Science 2021; 371:288-92.

11. Universidade Federal de Pelotas. 4ª fase do EPICOVID19 mostra desaceleração do coronavírus no Brasil. http://www.epidemio-ufpel.org.br/site/content/sala_imprensa/4-fase-do-epicovid19-mostra-desaceleracao-do-coronavirus-no-brasil.php?noticia=3149 (accessed on 02/Oct/2020).
12. Giattino C. How epidemiological models of COVID-19 help us estimate the true number of infections. https://ourworldindata.org/covid-models (accessed on 28/Sep/2020).
13. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer; 2011.
14. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
15. Poole D, Raftery AE. Inference for deterministic simulation models: the Bayesian melding approach. J Am Stat Assoc 2000; 95:1244-55.
16. Kiyuka PK, Agoti CN, Munywoki PK, Njeru R, Bett A, Otieno J, et al. Human coronavirus NL63 molecular epidemiology and evolutionary patterns in rural coastal Kenya. J Infect Dis 2018; 217:1728-39.
17. Gudbjartsson DF, Norddahl GL, Melsted P, Gunnarsdottir K, Holm H, Eythorsson E, et al. Humoral immune response to SARS-CoV-2 in Iceland. N Engl J Med 2020 383:1724-34.
18. Aguas R, Corder RM, King JG, Gonçalves G, Ferreira MU, Gomes MGM. Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics. medRxiv 2020; 24 jul. https://www.medrxiv.org/content/10.1101/2020.07.23.20160762v1?versioned=true.
19. Doshi P. Covid-19: do many people have pre-existing immunity? BMJ 2020; 370:m3563.

Resumo

Estimamos o efeito da subnotificação de casos de COVID-19 no Brasil até dezembro de 2020, com base nos dados coletados pela Pesquisa Nacional de Amostra de Domicílios sobre COVID-19 *(PNAD-COVID19) e a modelagem semi-bayesiana desenvolvida por Wu et al. O número total de indivíduos infectados é cerca de 3 a 8 vezes o número de casos notificados, a depender do estado do país. No final de 2020, os casos confirmados representavam 3,1% da população total, enquanto nossa estimativa aponta para quase 15% dos cerca de 212 milhões de brasileiros no mesmo período. O método de Wu et al., que adotamos com pequenas modificações nas especificações, aplica correções de vieses para compensar pela testagem incompleta e pela acurácia imperfeita dos testes. Nossas estimativas, que são comparáveis aos resultados obtidos por Wu et al. para os Estados Unidos, indicam que projeções a partir de modelos compartimentais (tais como modelos SEIR) tendem a superestimar o número de infecções, e que há uma heterogeneidade regional considerável (resultados apresentados por estado).*

Imunidade Coletiva; Viés de Seleção; Análise Quantitativa

Resumen

Usando los datos recogidos por la Encuesta Nacional por Muestra de Domicilios – COVID-19 *(PNAD-COVID19) y un modelado semibayesiano desarrollado por Wu et al., hemos estimado el efecto del subregistro de casos de COVID-19 en Brasil en diciembre de 2020. El número total de individuos infectados es de entre 3 a 8 veces más el número de casos informados, dependiendo del estado. Los casos confirmados son un 3,1% del total de población y nuestra estimación del total de casos es al menos un 15% de aproximadamente 212 millones de brasileños en 2020. El método que se tomó fue el de Wu et al., con leves modificaciones en las especificaciones previas, es aplicable a las correcciones de sesgo para tener en cuenta los test incompletos y la imprecisión de los tests. Nuestras estimaciones, que son comparables a los resultados obtenidos por Wu et al. para los Estados Unidos, indican las proyecciones de los modelos compartimentales (tales como los modelos SEIR), que tienden a sobreestimar el número de infecciones, así como la considerable heterogeneidad regional (los resultados se presentan por estado).*

Inmunidad Colectiva; Sesgo de Selección; Análisis Cuantitativo

Submitted on 06/Oct/2020
Final version resubmitted on 26/Apr/2021
Approved on 29/Apr/2021