Percentage of reported Covid-19 cases in Colombia: Estimating the true scale of the pandemic

Nicolás Parra, Vladimir Vargas, Juan Sebastián Flórez, and Leonel Ardila
Departamento de Física, Universidad Nacional de Colombia, Carrera 30 No.45-03, Bogotá D.C., Colombia and Laboratorio de Inteligencia Artificial y Computación de Alto Desempeño, Human Brain Technologies, Bogotá D.C., Colombia

Carlos Viviescas*
Departamento de Física, Universidad Nacional de Colombia, Carrera 30 No.45-03, Bogotá D.C., Colombia

(Dated: December 30, 2020)

Since March 6, when Colombia confirmed its first case of the coronavirus disease (Covid-19), the country healthcare system, with a limited testing capability, has struggled to monitor and report current cases. At the outbreak of a virus, data on cases is sparse and commonly severe cases, with a higher probability of a fatal resolution, are detected at a higher rate than mild cases. In addition, in an under-sampling situation, the number of total cases is under-estimated leading to a biased case fatality rate estimation, most likely inflating the virus mortality. Real time estimation of case fatality ratio can also be biased downwards by overlooking the delay between symptoms onset to death. In this communication, using reported data from Instituto Nacional de Salud up to December 28, we estimate the case fatality rate for Covid-19 in Colombia and some of its regions and cities adjusting for delay from onset to death. We then apply the method proposed by Russell *et al.* [1], and use our corrected case fatality rate to estimate the percentage of Covid-19 cases reported in Colombia as 42.95% (95% confidence interval: 42.50–43.41), which corresponds to a total of 3'661,621 estimated Covid-19 cases in the country.

I. INTRODUCTION

Keeping track of active cases in a pandemic is of paramount importance for epidemiological tracking in the early stages of the contagion spread. In more advanced stages of a pandemic where epidemiological routes cannot be constructed, effectively testing the population becomes a challenging task that must thrive to keep under-reporting as low as possible. The accuracy in measuring case incidence and prevalence, as well as mortality rates is decisive for high-quality epidemiological modelling that enables governments to propose and implement public policies to mitigate the impact of the pandemic [2–4].

Throughout the current Covid-19 pandemic, governments have mainly gathered databases of daily positive cases and positive deaths. Due to the finite resources of each country, tests for Covid-19 are performed only on a subset of the total country's population. Different countries have different strategies to select those subsets. No matter how effective those strategies are, it is expected that some infected people do not ever get tested for Covid-19, and therefore are not reported into the positive Covid-19 cases databases. This phenomenon is known as under-reporting, and its magnitude varies especially in relation with the effectiveness of epidemiological tracing strategies, as well as tests availability [5].

In this work we show how disease cases databases can be used to estimate the magnitude of this under-reporting, thus yielding better-quality figures for case incidence and mortality rates. In particular, we analyse and report data from Colombia to assess its current situation of cumulative positive cases. Unlike seroprevalence studies, the method that we use does not need to perform further tests than the ones already performed, and whose results are contained in the cases databases.

Our method is based on the idea of comparing case fatality ratios (CFRs) between a target country, which in our case is Colombia, and a benchmark country, which in our case is the Republic of Korea. The benchmark country should be one that has a very effective testing strategy. This means that the scientific community is confident in that strategy, which also means that the benchmark country is expected to have low under-reporting. The method answers the question of how many positive cases would be detected in the target country if it were using the testing strategy of the benchmark country? Related to this question, we also explain how to account for differences between the population of the target and the benchmark country.

This paper is divided as follows. Section II lays out the data used for our study, and also details the method to correct CFRs. Section III presents the estimated status of Covid-19 in Colombia. Then section IV compares our results with other studies, and also discusses some limitations of our study. Finally, we conclude in section V.

* clviviescasr@unal.edu.co

II. METHODS AND MATERIALS

Since the main objective of our work is to estimate the true number of cumulative positive cases of Covid-19 in Colombia, in this section we explain a method to answer the following question: for an arbitrary geographical region, can we account for under-reporting to estimate the true number of cumulative positive cases?

A. Data

We use two datasets for our analysis. The first one is the data published by the Instituto Nacional de Salud (INS) [6], which is updated every day and reports all known Covid-19 cases in Colombia. For each case it provides its date of notification, location city, state or district, current status (recovered, recovery from home, being treated in an intensive care unit and passed away), age, sex, country of provenance, symptom onset date, date of death, date of diagnosis, date of recovery and the web report date. From this information we derive the case incidence as well as the distribution of days between onset date and date of death for the cases that resulted in death. For this, we need the onset date for every record in the database, and their corresponding dates of death, if the cases resulted in death. We also keep information about age and location of each case. To construct the case incidence time-series (see fig. 1) as well as the onset-to-death distribution we used data up to December 28, 2020.

The other dataset is extracted from the demographic database maintained by the Departamento Administrativo Nacional de Estadística (DANE) [7] . We retrieve data containing the most recent projection for the year 2020 of the number of individuals per age group for all regions and cities of Colombia (see fig. 2). This demographic dataset will be needed in order to assess the vulnerability of each region to Covid-19, as it is well documented that death risk strongly depends on age [8–10].

B. Delay adjusted case fatality rate estimate

Dividing deaths-to-date by cases-to-death to compute the CFR (which we call from now on a naïve determination of the CFR, or nCFR) tends to yield a biased result, usually underestimating its true value [11, 12]. This happens because the outcome of active cases (i.e. recovery or death) is not known. An improved calculation of the CFR can be done by accounting for the delay from onset-to-death to estimate the number of cases expected to have a known outcome.

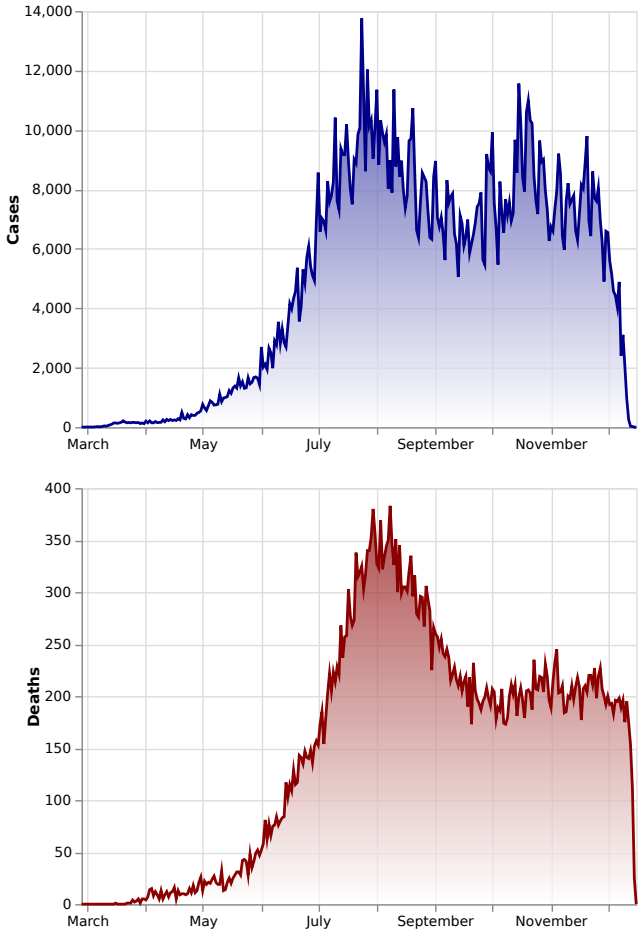

FIG. 1. Covid-19 cases and deaths incidence in Colombia. Up to December 28, 2020 there have been 1'594,497 confirmed cases and 47,175 confirmed deaths. Notice that there is a sudden drop at the latest days because data gathered in the future includes cases with onset and death dates in the past.

1. *Evaluating the delay time distribution between onset-to-death.*

We used the reported dates to calculate the time interval from illness onset to death of the confirmed 29,480 cases that resolved in death from March 16 to September 19 2020. The totality of cases in this period were already resolved (i.e. recovered or dead). We fit the conditional probability density $f(t)$ of the time between onset-to-death given death to Weibull, gamma, and lognormal distributions (as in the studies by Linton *et al.* [13], Verity *et al.* [14]). In all cases we use the maximum likelihood estimation (MLE) to determine the parameters and obtain credible intervals using PyMC3 [15]. Furthermore, we include a gaussian kernel density estimation (KDE) to fit the data more tightly. We select the best fit model by using the Akaike information criterion (AIC). Table I shows estimates for the three models plus the KDE. Although the lognormal distribution provided the best fit to data, we decided to use a gaussian

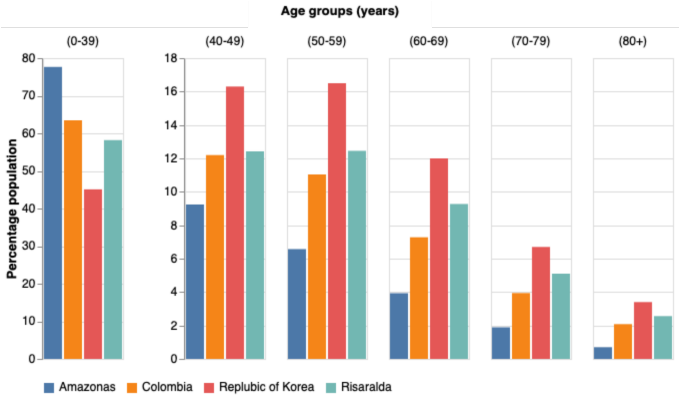

FIG. 2. Population demographic comparison between Amazonas (one of the youngest populations in Colombia), Risaralda (one of the oldest populations in Colombia), Colombia and the Republic of Korea.

KDE to approximate the onset-to-death distribution, as it is shown in fig. 3, because the lognormal fit was found to be largely inaccurate at the peak of the distribution (not shown). The mean time from illness onset to death given by the lognormal distribution was 22.8 days (95% CI: 22.5, 23.1), whereas with the KDE's was 22.4 days. As a comparison, Verity *et al.* [14] obtained a mean time of 18.8 days (95% CI: 15.7, 49.7), and Linton *et al.* [13] found a mean time of 15.0 days (95% CI: 12.8, 17.5).

TABLE I. Illness onset to death time-delay distribution for Covid-19 outbreak in Colombia.

	Gamma	Weibull	Lognormal	KDE
Mean (days)	22.5 (22.2-22.7)	22.6 (22.3-22.8)	22.8 (22.5-23.1)	22.4
SD (days)	19.4 (19.2-19.6)	22.9 (22.6-23.1)	35.2 (34.6-35.8)	12.7
AIC ($\times 10^3$)	250.0	251.0	247.2	247.1
Weight	0.0	0.0	0.0	1.0

2. *Adjusted case fatality ratio*

The CFR is adjusted following the method proposed by Nishiura *et al.* [11]. We estimate the proportion of cases resolved using the case and death incidence to adjust the CFR to account for delay outcome (cCFR). The underestimation factor [1, 11, 12, 16, 17],

$$u_t = \frac{\sum_{i=0}^{t}\sum_{j=0}^{\infty} c_{i-j} f_j}{\sum_{i=0}^{t} c_i}, \tag{1}$$

scales the cumulative number of cases in the denominator of the cCFR, and accounts for the adjustment. Here c_t is the daily case incidence (see top panel of fig. 1) at time t and $f_t = f(t)$ is the conditional probability density of the delay-time from onset-to-death (see fig. 3).

For Covid-19 the severity of the infection is highly correlated to the age of the infected individual [14], hence, for each region or city, we evaluate age-stratified estimates of the adjusted case fatality rate, cCFR_{r_i}, where r stands for the region and i labels the age group. The age aggregated cCFR for a region is adjusted for the population demographics, and is given by

$$\mathrm{cCFR}_r = \sum_i p_{r_i}\, \mathrm{cCFR}_{r_i}\,. \tag{2}$$

Here, p_{r_i} is the fraction of the population with age i for the region r.

C. Percentage of cases reported

The adjusted cCFR does not account for under-reporting. In order to obtain an estimate of the potential level of under-reporting in Colombia and its regions our model follows the simple method proposed by Russell *et al.* [1] further adjusting for the country demography. We assume a baseline CFR (blCFR), taken from a benchmark country, and compare it with the estimated cCFR for Colombia and some of its regions and cities. We do an age-stratified analysis. If the cCFR_{r_i} is higher than blCFR_i, it indicates that only a portion of the real number of cases in this age group have been reported so far. The fraction of reported cases in region r and age group i is given by

$$R_{r_i} = \frac{\mathrm{blCFR}_i}{\mathrm{cCFR}_{r_i}}\,. \tag{3}$$

We also evaluate the fraction of cases reported in a region aggregated over age R_r. For this, we introduce the region

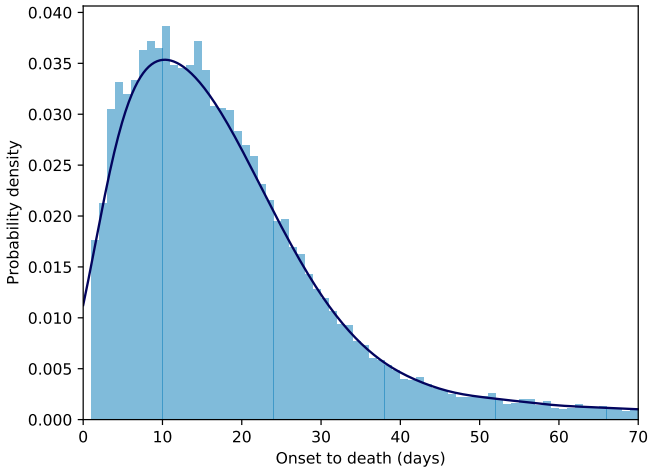

FIG. 3. Probability density distribution of the time from illness onset to death. The gamma distribution has mean delay of 22.4 days and standard deviation of 12.7 days.

baseline CFR,

$$\mathrm{blCFR}_r = \sum_i p_{r_i} \mathrm{blCFR}_i, \tag{4}$$

which accounts for the region population demographics through p_{r_i}, and obtain

$$R_r = \frac{\mathrm{blCFR}_r}{\mathrm{cCFR}_r}\,. \tag{5}$$

The most recent literature offers several estimates CFR for Covid-19. Among those adjusting or controlling for under-reporting, we mention those by Verity *et al.* [14], Russell *et al.* [16], Shim *et al.* [17], Guan *et al.* [18]. In our analysis, we use as benchmark country the Republic of Korea. As baseline we use for this work the age stratified CFR data of July 14, 2020 [19], listed in Table II, since most of the cases from the first peak had already been resolved by this date.

TABLE II. Covid-19 case fatality rates stratified by age groups as of July 14, 2020 in the Republic of Korea.

Age Groups (years)	Case Fatality Ratio (95% CI)
0-39	0.04 % (0.00-0.09)
40-49	0.17 % (0.00-0.36)
50-59	0.62 % (0.31-0.93)
60-69	2.32 % (1.62-3.02)
70-79	9.31 % (7.42-11.21)
≥80	24.96 % (21.43-28.49)

III. RESULTS

In our analysis we consider all regions and cities in Colombia that have reported at least 40 fatal Covid-19 cases as of December 28, 2020. Table III shows the percentage of cases reported, the cCFR, blCFR and the total cases and deaths for all regions of Colombia. Note that the value of the blCFR fluctuates from region to region, evidencing the differences in population demographics between regions. For regions and cities with a younger population than the average of the country, the blCFR is lower compared to the value for the country. On the contrary, in those regions or cities in which the fraction of the population with age above 60 years is more significant, the blCFR is higher than the value for the country. As an example, consider fig. 2 where the age distribution of Amazonas, the region with the youngest population, is compared to the age distribution of Risaralda, the region with the oldest population. The corresponding blCFRs mirror the demographics, with Amazonas' blCFR being 0.5% and Risaralda's being 1.4%.

We also present age-stratified reporting percentages for all the regions in Colombia (with at least 40 fatal Covid-19 cases) in table IV. A trend is notable: reporting percentage is much higher for the elder population, whereas it is quite low for the young population. We hypothesise that young people do not get tested as much as old people because if symptomatic, they only develop mild symptoms in most cases [10, 20]. This is an important source of uncertainty on the estimation of the virus propagation because young people are relevant vectors because of mobility reasons, especially in cities [21, 22].

Furthermore, the framework so far exposed can be used to evaluate reporting percentages and estimations of cCFRs in the past by truncating the datasets. However, there is a detail that has to be handled carefully. Notice that when we compute the cCFR, we are using the number of death people reported up to today, and we are using the case incidence curve reported up to today. This means that we are in the best state of knowledge that we can about today. However, tomorrow, we will have a state of knowledge about today better than the one we have today. This is because tomorrow there will be some cases with onset symptoms dates from today and from previous days. Similarly, tomorrow there will be reported cases of people that died today or in previous days. Therefore, we need to truncate the datasets based on the report day only. Thus, if we truncate the database a month ago, we ensure that we are reproducing the state of knowledge that we had one month ago.

The evolution of reporting percentage confidence intervals are shown in fig. 4. Qualitatively, almost all regions start with wide confidence intervals, and they become narrow as more people are reported positive or dead because of Covid-19. Also, in most regions the reporting percentage raises as time passes by. An important feature is that in some of the most inhabited regions, there are drops on the reporting percentage about July-August. This coincides with the peak of positive cases reported in Colombia. Even though the number of tests also incremented on those dates, they did not match the rate at which new cases and more importantly, new deaths rose.

These plots can also be useful to evaluate different public policies being carried out at each region. For instance, Boyacá closed their borders early on in the pandemic [23]. Later, they opened the borders, and naturally, the number of cases grew. The peak is found to be at the end of October and beginning of November. Nonetheless, despite a pronounced peak, the number of tests was enough to maintain the same reporting percentage over the course of those months, as seen in fig. 4. Another interesting case is Amazonas, the largest state in Colombia by area. It hit the headlines several times because it had the largest number of deaths per inhabitant at the start of the pandemic (around May) [24]. As a consequence, the government performed many tests (with respect to other states), making Amazonas the state with more performed tests per inhabitant. Despite this special attention that the state got, its reporting percentage is well below the average of Colombia.

TABLE III. Percentage of Covid-19 cases reported in Colombia and its regions until December 28, 2020 with 95% confidence intervals. The corrected and baseline CFRs are also shown, along with the total number of positive cases and deaths to date.

	Percentage of cases reported	cCFR	blCFR	Total cases	Total deaths
Colombia	43 % (43-43)	2.7 % (2.7-2.8)	1.2 % (0.9-1.5)	1594497	47175
AMAZONAS	24 % (20-28)	2.2 % (1.4-3.3)	0.5 % (0.4-0.7)	3231	134
ANTIOQUIA	56 % (55-58)	2.2 % (2.1-2.3)	1.2 % (0.9-1.5)	255845	5220
ARAUCA	31 % (26-36)	2.4 % (1.6-3.5)	0.7 % (0.5-1.0)	4521	150
ATLANTICO	29 % (27-30)	3.8 % (3.4-4.2)	1.1 % (0.8-1.4)	37024	1690
BARRANQUILLA	36 % (35-38)	3.5 % (3.1-3.8)	1.3 % (1.0-1.6)	54725	2082
BOGOTA	48 % (47-49)	2.3 % (2.3-2.4)	1.1 % (0.9-1.4)	452940	10865
BOLIVAR	37 % (33-41)	2.8 % (2.2-3.7)	1.0 % (0.8-1.3)	8124	333
BOYACA	55 % (51-59)	2.7 % (2.2-3.1)	1.5 % (1.1-1.8)	27620	681
CALDAS	66 % (62-71)	2.3 % (2.0-2.7)	1.5 % (1.2-1.9)	31604	765
CAQUETA	30 % (28-32)	2.8 % (2.3-3.3)	0.8 % (0.6-1.0)	14845	599
CARTAGENA	57 % (53-61)	1.8 % (1.5-2.1)	1.0 % (0.8-1.3)	42318	819
CASANARE	33 % (29-38)	2.2 % (1.5-3.0)	0.7 % (0.5-0.9)	8578	197
CAUCA	47 % (44-51)	2.4 % (2.0-2.9)	1.1 % (0.9-1.4)	18174	578
CESAR	34 % (32-36)	2.4 % (2.1-2.7)	0.8 % (0.6-1.0)	33161	1117
CHOCO	28 % (25-33)	2.7 % (1.9-3.7)	0.8 % (0.6-1.0)	4800	199
CORDOBA	27 % (26-28)	4.0 % (3.6-4.4)	1.1 % (0.8-1.4)	29427	1892
CUNDINAMARCA	42 % (40-44)	2.8 % (2.5-3.1)	1.2 % (0.9-1.5)	64803	1877
GUAINIA	35 % (24-53)	1.2 % (0.3-3.1)	0.4 % (0.3-0.6)	1229	23
GUAJIRA	29 % (26-31)	2.2 % (1.8-2.7)	0.6 % (0.5-0.8)	13938	563
GUAVIARE	36 % (27-49)	1.8 % (0.9-3.5)	0.7 % (0.5-0.9)	1939	40
HUILA	37 % (35-39)	3.0 % (2.6-3.4)	1.1 % (0.8-1.4)	34205	1190
MAGDALENA	21 % (19-23)	4.4 % (3.6-5.4)	0.9 % (0.7-1.2)	7525	575
META	39 % (37-42)	2.3 % (2.0-2.7)	0.9 % (0.7-1.2)	33144	830
NARIÑO	39 % (37-41)	3.2 % (2.8-3.7)	1.3 % (1.0-1.6)	30289	1064
NORTE SANTANDER	24 % (23-25)	4.4 % (3.9-4.8)	1.0 % (0.8-1.3)	38993	2048
PUTUMAYO	24 % (21-27)	3.3 % (2.5-4.4)	0.8 % (0.6-1.0)	5791	258
QUINDIO	54 % (50-58)	2.9 % (2.5-3.4)	1.6 % (1.2-2.0)	22919	671
RISARALDA	53 % (50-57)	2.7 % (2.3-3.2)	1.4 % (1.1-1.8)	32719	813
SAN ANDRES	78 % (58-100)	1.2 % (0.6-2.3)	0.9 % (0.7-1.2)	2431	40
SANTANDER	38 % (36-39)	3.3 % (3.0-3.6)	1.2 % (1.0-1.6)	65278	2530
STA MARTA D.E.	35 % (32-38)	2.6 % (2.1-3.1)	0.9 % (0.7-1.2)	16760	583
SUCRE	39 % (37-42)	2.8 % (2.4-3.3)	1.1 % (0.8-1.4)	17377	748
TOLIMA	46 % (43-48)	3.2 % (2.9-3.6)	1.5 % (1.1-1.8)	42407	1336
VALLE	41 % (40-43)	3.5 % (3.3-3.7)	1.4 % (1.1-1.8)	133536	4636
VAUPES	63 % (38-100)	0.8 % (0.2-2.3)	0.5 % (0.3-0.6)	1124	14
VICHADA	58 % (35-96)	0.8 % (0.3-1.6)	0.5 % (0.3-0.6)	1136	15

IV. DISCUSSION

An important point has to be remarked on the validity of our results. We saw that we corrected the CFR for today with the information that we possess today. In the future we will have more information about today, and a better correction of the CFR for today can be estimated. To account for this lag of information we would need to predict how many cases will be reported in future days. To perform this prediction, we need to use information about how many people have been infected so far. And calculating how many people have been infected so far needs the prediction of new cases. Thus, this becomes an auto-consistent problem which falls beyond the scope of this work. Nonetheless, the further correction that can be achieved by solving this auto-consistent problem can be important, especially in the early stages of the pandemic. For instance, a simple prediction of new cases done by linear regression on the case incidence time-series can yield statistically significant differences in the reporting percentage that we estimate (not shown). More accurate predictions done by fitting epidemiological models can improve the accuracy of our last statement.

A. Our study compared to others

Comparing our results with the extant literature, we find that it is qualitatively consistent (in the case of the complete country of Colombia) with the study by Russell *et al.* [1]. They find a peak in the reporting percentage around May-June, but we find it centered in June (c.f. fig. 4). This can be due to different onset-to-death distributions taken into account: ours is deduced from a

TABLE IV. Age-stratified percentage of Covid-19 cases reported in Colombia until December 28, 2020. For the country and its regions the age-stratified percentage of reported cases are shown with a 95% confidence interval.

	Percentange of cases reported by age group (years)					
	0-39	40-49	50-59	60-69	70-79	$\geq$ 80
Colombia	13 % (4-49)	15 % (5-44)	21 % (13-34)	26 % (20-36)	49 % (40-60)	75 % (65-87)
AMAZONAS	12 % (3-53)	7 % (2-22)	8 % (5-16)	21 % (13-35)	31 % (22-43)	64 % (47-90)
ANTIOQUIA	22 % (6-79)	29 % (10-85)	35 % (21-57)	38 % (28-51)	60 % (48-73)	81 % (70-94)
ARAUCA	12 % (3-49)	9 % (3-28)	12 % (7-23)	22 % (14-35)	36 % (26-50)	87 % (63-100)
ATLANTICO	7 % (2-27)	7 % (2-22)	12 % (7-19)	18 % (13-24)	35 % (28-44)	62 % (53-72)
BARRANQUILLA	8 % (2-30)	10 % (3-30)	15 % (9-25)	22 % (16-30)	46 % (37-57)	67 % (57-78)
BOGOTA	18 % (5-65)	19 % (6-56)	25 % (15-41)	31 % (23-43)	55 % (45-67)	78 % (67-89)
BOLIVAR	8 % (2-30)	7 % (2-21)	19 % (10-35)	23 % (16-33)	52 % (39-70)	72 % (59-89)
BOYACA	15 % (4-58)	18 % (6-55)	29 % (17-50)	31 % (22-43)	64 % (50-82)	84 % (70-100)
CALDAS	20 % (5-76)	46 % (14-100)	32 % (19-56)	41 % (29-57)	73 % (57-93)	100 % (84-100)
CAQUETA	8 % (2-30)	14 % (4-43)	12 % (7-21)	17 % (12-23)	40 % (32-52)	74 % (61-90)
CARTAGENA	17 % (4-62)	17 % (6-51)	29 % (17-49)	40 % (29-56)	73 % (57-93)	89 % (74-100)
CASANARE	12 % (3-49)	13 % (4-41)	13 % (7-24)	25 % (17-39)	39 % (28-53)	79 % (59-100)
CAUCA	13 % (4-51)	16 % (5-51)	27 % (15-48)	28 % (20-40)	50 % (39-63)	84 % (70-100)
CESAR	7 % (2-27)	12 % (4-35)	16 % (10-28)	24 % (17-33)	44 % (35-54)	75 % (63-89)
CHOCO	5 % (1-21)	6 % (2-20)	11 % (6-21)	17 % (11-25)	53 % (37-78)	78 % (59-100)
CORDOBA	5 % (1-17)	6 % (2-18)	11 % (7-18)	18 % (13-24)	37 % (30-46)	61 % (52-71)
CUNDINAMARCA	15 % (4-55)	17 % (6-50)	21 % (13-35)	23 % (17-32)	47 % (37-58)	79 % (67-93)
GUAINIA	10 % (2-49)	14 % (3-71)	12 % (5-29)	28 % (12-68)	100 % (41-100)	92 % (43-100)
GUAJIRA	7 % (2-27)	7 % (2-21)	13 % (7-22)	21 % (15-30)	44 % (34-56)	68 % (55-82)
GUAVIARE	46 % (6-100)	10 % (3-38)	100 % (19-100)	17 % (10-32)	40 % (24-71)	80 % (50-100)
HUILA	9 % (2-34)	13 % (4-38)	16 % (10-27)	22 % (16-31)	46 % (36-57)	69 % (58-81)
MAGDALENA	4 % (1-14)	6 % (2-17)	9 % (5-15)	13 % (10-19)	33 % (26-43)	50 % (42-60)
META	17 % (4-63)	17 % (6-52)	18 % (11-31)	23 % (16-32)	47 % (37-60)	79 % (66-94)
NARIÑO	9 % (2-34)	11 % (4-33)	19 % (11-33)	24 % (17-33)	43 % (34-54)	70 % (59-82)
NORTE SANTANDER	6 % (2-23)	6 % (2-18)	9 % (5-15)	14 % (10-19)	31 % (25-38)	60 % (51-70)
PUTUMAYO	9 % (2-34)	7 % (2-21)	10 % (6-18)	14 % (10-21)	29 % (22-39)	62 % (49-79)
QUINDIO	20 % (5-78)	20 % (7-63)	29 % (17-51)	36 % (25-50)	48 % (38-61)	99 % (82-100)
RISARALDA	19 % (5-71)	28 % (9-86)	25 % (15-43)	31 % (22-43)	58 % (46-73)	89 % (74-100)
SAN ANDRES	23 % (4-100)	100 % (100-100)	69 % (22-100)	29 % (17-48)	100 % (85-100)	100 % (93-100)
SANTANDER	12 % (3-45)	14 % (5-42)	15 % (9-25)	21 % (15-28)	43 % (35-54)	77 % (66-89)
STA MARTA D.E.	10 % (3-38)	8 % (3-24)	16 % (9-27)	23 % (16-33)	46 % (36-59)	81 % (66-100)
SUCRE	11 % (3-41)	11 % (4-33)	17 % (10-29)	22 % (16-31)	50 % (40-64)	77 % (65-92)
TOLIMA	14 % (4-51)	22 % (7-68)	22 % (13-37)	26 % (19-35)	51 % (41-64)	75 % (64-88)
VALLE	16 % (4-57)	15 % (5-44)	21 % (13-35)	25 % (18-33)	44 % (36-54)	68 % (59-79)
VAUPES	12 % (2-67)	100 % (100-100)	81 % (14-100)	29 % (14-62)	100 % (59-100)	100 % (58-100)
VICHADA	100 % (100-100)	100 % (100-100)	100 % (100-100)	21 % (10-44)	50 % (24-100)	100 % (54-100)

distribution built taking into account the onset-to-death, whereas theirs is a generalisation of a hospitalisation-to-death distribution from Wuhan. Then, the reporting percentage starts to grow in their study and ours. Furthermore, previous to May, we find an increase in the error, whereas their study still finds low error in the estimation of the reporting percentage. However, the magnitude of the estimation error rapidly decreases as more deaths are recorded. Furthermore, since our analysis includes the demographics of Colombia and its regions, this allows to identify their vulnerability. On average, Colombia has a younger population than the population from China [18], which is why our baseline CFR is lower than Russell's, which is 1.4%. This explains why the range of reporting percentages is lower in our study than in theirs, where reporting percentages reach 50% in August.

Another important resource when trying to estimate the true number of infected Covid-19 cases in Colombia and its regions is a seroprevalence study. In Colombia, some seroprevalence studies were recently published. In table V we show the results for some Colombian cities. By dividing the reported number of deaths by an estimate of the cases given by the seroprevalence studies, we computed the corresponding nCFR for each city. We also added the results of seroprevalence studies for two of the most affected zones in Europe and the New York State. We see a high seroprevalence and low values of nCFR in all of Colombia regions. On the other hand, European zones show an opposite situation. In the case of New York, the study was done just before the days with a high number of daily deaths (29 March), which is why the nCFR is very low. Rosenberg *et al.* [25] mentioned that if they considered an average of 19 days between onset-to-death as an average, the nCFR would be 0.6 %, which still doubles the nCFR found in Colombia. As a conclusion, the nCFR of Colombia deduced by seroprevalence studies in all its regions is smaller than in other countries. We hypothesise that this could be due to the age-distribution of the countries: Colombia's population is younger than USA's, Spain's and Italy's.

Also, Colombian seroprevalence studies are preliminary and the methodology details is not publish yet. Therefore, we do not know the possible biases. Particularly, the Montería study [26] has a possible bias because tested people were chosen randomly in different neighbourhoods and it is known that cities have intrinsically heterogeneous mobility patterns [21], thus making any uniform random samples problematic [27, 28].

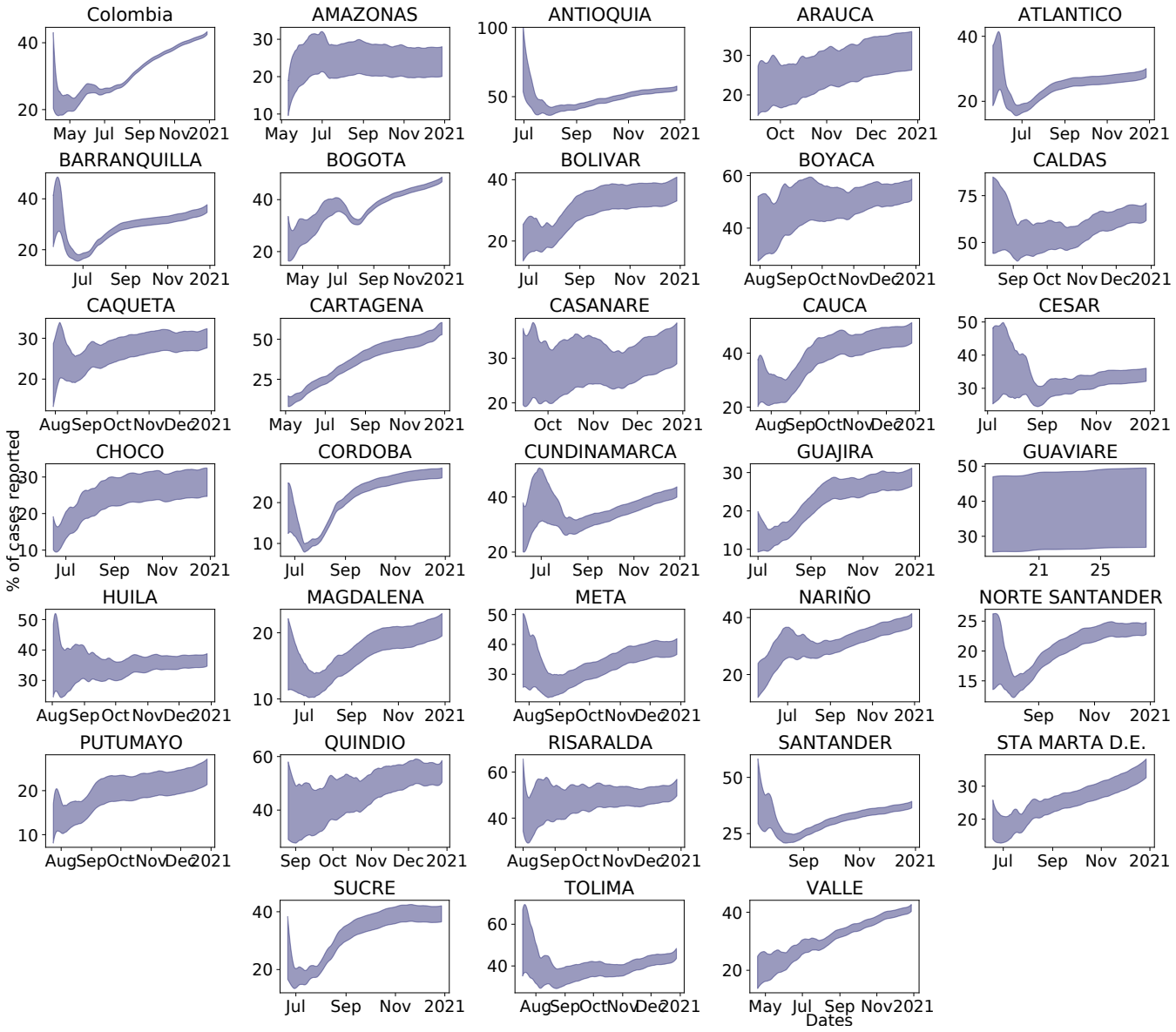

FIG. 4. 95% confidence interval of the evolution of age-aggregated percentage of reported cases in regions of Colombia with more than 40 confirmed deaths.

B. Limitations

Our study makes some assumptions that directly impact the results, therefore, the figures given in this work have to consider the following limitations:

- We do not account for comorbidities of the population which can influence the vulnerability (i.e. the baseline CFR) of a country or a region. The main comorbidities correlated with Covid-19 deaths are cerebrovascular disease, high-blood pressure, chronic obstructive pulmonar diseases, diabetes, among others [32].

- Also, we ignore the burden of the healthcare system among other regional and socio-economical factors that might influence the capacity of a country or a region to attend all important Covid-19 cases.

- Our results are liked and biased by the baseline CFR, which come from measured nCFRs from the Republic of Korea. An important effect is that reporting percentage (and therefore the estimation of seroprevalence) can change a lot for small errors in the estimation of the baseline CFR.

- Our results ignore deaths under-counting. In fact, as we discussed by comparing our results to seroprevalence studies in Colombia, it seems that nCFRs are too low when compared with many

TABLE V. nCFR of some cities in Colombia, as well as in the Lombardy region, Comunidad de Madrid, and New York State for comparison. The nCFR is computed using seroprevalence studies in those regions.

City	Total population	Seroprevalence %	Reported Deaths	nCFR %
Comunidad de Madrid, Spain	6747425	11.3 [29]	8683	1.139
Lombardy Region, Italy	10060574	2.7 [30]	9722	3.579
New York State, Usa	19453561	14 [25]	1722	0.064
Barranquilla, Colombia	1274250	55 [31]	1766	0.252
Bogotá, Colombia	7743955	30 [31]	9313	0.401
Bucaramanga, Colombia	607428	32 [31]	848	0.436
Cúcuta, Colombia	777106	40 [31]	1032	0.332
Leticia, Colombia	42280	59 [31]	105	0.421
Medellín, Colombia	2529403	27 [31]	2097	0,307
Montería, Colombia	505334	55,3 [26]	709	0.253
Villavicencio, Colombia	551212	34 [31]	474	0.253

other seroprevalence studies. Since it is expected that mortality is higher, it is a possibility that there is a strong death under-counting in Colombia.

- We do not take into account different strains of Covid-19 and the different CFRs for each type [33, 34].

V. CONCLUSIONS

In the present study we accounted for the delay between symptom's onset and death caused by the Covid-19 disease in order to estimate its case fatality ratio in Colombia and its regions. This estimation allows us to calculate the total number of infected people so far in Colombia and its regions, thus creating a clear picture of the magnitude of the pandemic in this country. In particular, we estimate that there have been a total number of 3'661,621 infected people in Colombia, in contrast with the confirmed 1'594,497, as of December 28, 2020. Moreover, the following are the states with the largest under-reporting in Colombia: Amazonas, Magdalena, Norte de Santander and Putumayo. Furthermore, the capital city of Colombia, Bogotá, presents a reporting percentage close to 48%.

A remarkable feature of the method to estimate CFRs in Colombia and its regions was the inclusion of demographics, as Covid-19 has proven to have different mortalities for populations with different ages. Therefore, we saw that older populations were more vulnerable than younger ones. Indeed, our corrected CFRs have a correlation of 0.51 with the baseline CFRs, which account for regions demographics, indicating that younger (older) populations are expected to have lower (higher) mortalities.

Another phenomenon linked with age distributions is that the reporting percentage is much lower in the young population, whereas the old population presents large reporting percentage. We hypothesise that this happens because young people usually have low to mild symptoms, and most of the time they do not require to go to the hospital, and prefer not to take a Covid-19 test. This behaviour is seen in every single region of Colombia.

Finally, we also compared our results to seroprevalence studies from Colombia. Although most of those studies have only released preliminary results, they all indicate mortalities much lower than those reported in Europe and the United States of America. Also, compared to our results, the seroprevalence studies in Colombia show lower estimates of mortality, and thus, estimate more true cases of Covid-19 in the country than we do.

VI. DATA AVAILABILITY

We developed a Python library that eases the extraction, load and transformation of raw data from the INS database. The code is freely available at `https://gitlab.com/hubrain/covid19`. Also, we designed a dashboard were we keep a daily record of how reporting percentages are changing at each region which can be found at `http://covid19.hubrain.co`. Furthermore, the code to reproduce the figures and tables here presented is in `https://gitlab.com/hubrain/covid19-paper`.

[1] T. W. Russell, J. Hellewell, S. Abbott, N. Golding, H. Gibbs, C. I. Jarvis, K. van Zandvoort, CMMID nCov working group, S. Flasche, R. Eggo, W. J. Edmunds, and A. J. Kucharski, Using a delay-adjusted case fatality ratio to estimate under-reporting, `https://cmmid.github.io/topics/covid19/global_cfr_estimates.html` (2020).

[2] O. Mitjà, À. Arenas, X. Rodó, A. Tobias, J. Brew, and J. M. Benlloch, Experts' request to the spanish government: move spain towards complete lockdown, The

Lancet **395**, 1193 (2020).

[3] A. Tobías, Evaluation of the lockdowns for the sars-cov-2 epidemic in italy and spain after one month follow up, Science of the Total Environment , 138539 (2020).

[4] M. Vinceti, T. Filippini, K. J. Rothman, F. Ferrari, A. Goffi, G. Maffeis, and N. Orsini, Lockdown timing and efficacy in controlling covid-19 using mobile phone tracking, EClinicalMedicine **25**, 100457 (2020).

[5] H. Rahmandad, T. Y. Lim, and J. Sterman, Estimating covid-19 under-reporting across 86 nations: implications for projections and control, Available at SSRN 3635047 (2020).

[6] Instituto Nacional de Salud, Casos positivos de COVID-19 en Colombia, `https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data` (2020).

[7] DANE: Departamento Administrativo Nacional de Estadística, DANE: ¿CUÁNTOS SOMOS?, `https://sitios.dane.gov.co/cnpv/#\protect\leavevmode@ifvmode\kern-.1667em\relax/` (2020), accessed: 2020-06-01.

[8] K. M. Jagodnik, F. Ray, F. M. Giorgi, and A. Lachmann, Correcting under-reported covid-19 case numbers: estimating the true scale of the pandemic, medRxiv (2020).

[9] S.-J. Kang and S. I. Jung, Age-related morbidity and mortality among patients with covid-19, Infection & chemotherapy **52**, 154 (2020).

[10] A. T. Levin, W. P. Hanage, N. Owusu-Boaitey, K. B. Cochran, S. P. Walsh, and G. Meyerowitz-Katz, Assessing the age specificity of infection fatality rates for covid-19: systematic review, meta-analysis, and public policy implications, European journal of epidemiology , 1 (2020).

[11] H. Nishiura, D. Klinkenberg, M. Roberts, and J. A. P. Heesterbeek, Early Epidemiological Assessment of the Virulence of Emerging Infectious Diseases: A Case Study of an Influenza Pandemic, PLoS ONE **4**, e6852 (2009).

[12] A. J. Kucharski and W. J. Edmunds, Case fatality rate for Ebola virus disease in west Africa, The Lancet **384**, 1260 (2014).

[13] N. M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita, and H. Nishiura, Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data, Journal of Clinical Medicine **9**, 538 (2020).

[14] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson, Estimates of the severity of coronavirus disease 2019: a model-based analysis, The Lancet Infectious Diseases **20**, 669 (2020).

[15] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, Probabilistic programming in python using pymc3, PeerJ Computer Science **2**, e55 (2016).

[16] T. W. Russell, J. Hellewell, C. I. Jarvis, K. van Zandvoort, S. Abbott, R. Ratnayake, CMMID nCov working group, S. Flasche, R. M. Eggo, W. J. Edmunds, and A. J. Kucharski, Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020, Euro Surveill **25** (2020).

[17] E. Shim, K. Mizumoto, W. Choi, and G. Chowell, Estimating the Risk of COVID-19 Death During the Course of the Outbreak in Korea, February–May 2020, Journal of Clinical Medicine 2020, Vol. 9, Page 538 **9**, 1641 (2020).

[18] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, *et al.*, Clinical characteristics of coronavirus disease 2019 in china, New England journal of medicine **382**, 1708 (2020).

[19] Ministry of Health and Welfare, Coronavirus disease-19 (COVID-19), `http://ncov.mohw.go.kr/tcmBoardView.do?brdId=&brdGubun=&dataGubun=&ncvContSeq=355414&contSeq=355414&board_id=140&gubun=BDJ#` (2020), accessed: 2020-07-15.

[20] J. R. Lechien, C. M. Chiesa-Estomba, D. R. De Siati, M. Horoi, S. D. Le Bon, A. Rodriguez, D. Dequanter, S. Blecic, F. El Afia, L. Distinguin, *et al.*, Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (covid-19): a multicenter european study, European Archives of Oto-Rhino-Laryngology , 1 (2020).

[21] A. Arenas, W. Cota, J. Gómez-Gardeñes, S. Gómez, C. Granell, J. T. Matamalas, D. Soriano-Paños, and B. Steinegger, Modeling the spatiotemporal epidemic spreading of covid-19 and the impact of mobility and social distancing interventions, Phys. Rev. X **10**, 041055 (2020).

[22] A. Arenas, J. Gómez-Gardeñes, C. Granell, and D. Soriano-Paños, Epidemic spreading: Tailored models for covid-19, Europhysics News **51**, 38 (2020).

[23] B. Ramiro, Boyacá anuncia cierre de fronteras por coronavirus, `https://www.elespectador.com/coronavirus/boyaca-anuncia-cierre-de-fronteras-por-coronavirus-articul` (2020).

[24] U. de datos El Tiempo, Amazonas, el de más casos de coronavirus por cada 10 mil habitantes, `https://www.eltiempo.com/datos/coronavirus-en-amazonas-cifras-de-contagio-y-muertes-por-c` (2020).

[25] E. S. Rosenberg, J. M. Tesoriero, E. M. Rosenthal, R. Chung, M. A. Barranco, L. M. Styer, M. M. Parker, S.-Y. John Leung, J. E. Morne, D. Greene, D. R. Holtgrave, D. Hoefer, J. Kumar, T. Udo, B. Hutton, and H. A. Zucker, Cumulative incidence and diagnosis of sars-cov-2 infection in new york, Annals of Epidemiology **48**, 23 (2020).

[26] S. Mattar, N. Alvis-Guzman, E. Garay, R. Rivero, A. García, Y. Botero, J. Miranda, K. Galeano, F. de La Hoz, C. Martínez, *et al.*, Severe acute respiratory syndrome coronavirus 2 seroprevalence among adults in a tropical city of the caribbean area, colombia: Are we much closer to herd immunity than developed countries?, in *Open Forum Infectious Diseases*, Vol. 7 (Oxford University Press US, 2020) p. ofaa550.

[27] A. Gamble, R. Garnier, T. Chambert, O. Gimenez, and T. Boulinier, Next-generation serology: integrating cross-sectional and capture–recapture approaches to infer disease dynamics, Ecology **101**, e02923 (2020).

[28] W. A. Link, Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities, Biometrics **59**, 1123 (2003).

[29] M. Pollán, B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M. A. Hernán, M. Pérez-Olmeda, J. L. Sanmartín, A. Fernández-García, I. Cruz, N. F. de Larrea, *et al.*, Prevalence of sars-cov-2 in spain (ene-covid): a nationwide, population-based seroepidemiological study, The Lancet **396**, 535 (2020).

[30] L. Valenti, A. Bergna, S. Pelusi, F. Facciotti, A. Lai, M. Tarkowski, A. Berzuini, F. Caprioli, L. Santoro, G. Baselli, *et al.*, Sars-cov-2 seroprevalence trends in healthy blood donors during the covid-19 milan outbreak, medRxiv (2020).

[31] Estudio Nacional de Seroprevalencia, `https://www.ins.gov.co/estudio-nacional-de-seroprevalencia/reporte.html`, accessed: 2020-12-21.

[32] Y. Ji, Z. Ma, M. P. Peppelenbosch, and Q. Pan, Potential association between covid-19 mortality and health-care resource availability, The Lancet Global Health **8**, e480 (2020).

[33] J. Wise, Covid-19: New coronavirus variant is identified in uk, BMJ **371**, 10.1136/bmj.m4857 (2020), https://www.bmj.com/content/371/bmj.m4857.full.pdf.

[34] A. Bal, G. Destras, A. Gaymard, H. Regue, Q. Semanas, C. d'Aubarde, G. Billaud, F. Laurent, C. Gonzales, M. Valette, *et al.*, Two-step strategy for the identification of sars-cov-2 variants co-occurring with spike deletion h69-v70, lyon, france, august to december 2020, medRxiv (2020).