Edinburgh Research Explorer

CoronaSurveys: Using Surveys with Indirect Reporting to Estimate the Incidence and Evolution of Epidemics

Link:
[Link to publication record in Edinburgh Research Explorer](#)

Document Version:
Peer reviewed version

CoronaSurveys: Using Surveys with Indirect Reporting to Estimate the Incidence and Evolution of Epidemics

Oluwasegun Ojo
Augusto García-Agundez
Benjamin Girault
Harold Hernández
Elisa Cabana
Amanda García-García
Payman Arabshahi
Carlos Baquero
Paolo Casari

Ednaldo José Ferreira
Davide Frey
Chryssis Georgiou
Mathieu Goessens
Anna Ishchenko
Ernesto Jiménez
Oleksiy Kebkal
For the authors' affiliations, see
https://coronasurveys.org/team/

Rosa Lillo
Raquel Menezes
Nicolas Nicolaou
Antonio Ortega
Paul Patras
Julian C Roberts
Efstathios Stavrakis
Yuichi Tanaka
Antonio Fernández Anta

ABSTRACT

The world is suffering from a pandemic called COVID-19, caused by the SARS-CoV-2 virus. National governments have problems evaluating the reach of the epidemic, due to having limited resources and tests at their disposal. This problem is especially acute in low and middle-income countries (LMICs). Hence, any simple, cheap and flexible means of evaluating the incidence and evolution of the epidemic in a given country with a reasonable level of accuracy is useful. In this paper, we propose a technique based on (anonymous) surveys in which participants report on the health status of their contacts. This indirect reporting technique, known in the literature as *network scale-up method*, preserves the privacy of the participants and their contacts, and collects information from a larger fraction of the population (as compared to individual surveys). This technique has been deployed in the CoronaSurveys project, which has been collecting reports for the COVID-19 pandemic for more than two months. Results obtained by CoronaSurveys show the power and flexibility of the approach, suggesting that it could be an inexpensive and powerful tool for LMICs.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; • **Applied computing** → **Health informatics**.

KEYWORDS

epidemics, surveys, crowdsourcing, privacy

1 INTRODUCTION

During the current SARS-CoV-2 virus pandemic, monitoring the evolution of COVID-19 cases is very important for authorities to make informed policy decisions, and for the general public to be informed of the reach of the problem. Official numbers of confirmed cases are periodically issued by health authorities [4]. Unfortunately, at the early stages of a pandemic outbreak there is usually only limited ability to test, as well as a lack of other resources. Hence, it is not possible to test all potential cases, and some eligibility criteria is applied to decide who is tested. Under these circumstances, the official confirmed cases are unlikely to represent the total number of cases (see [12]). This problem is more pressing in low and middle-income countries (LMICs), which may be limited in their ability to

deploy massive testing. Not having access to reliable data clearly prevents authorities from making informed decisions, putting the population at higher risk.

This motivates the need for other probing techniques, beyond laboratory testing, that can estimate the number of cases and their evolution. Information obtained using these alternative methods can be more timely, leaving more accurate estimates of the number of cases for later studies (e.g., massive serological studies [18]). Techniques that allow cheap and massive data collection, and lead to reasonably accurate estimates, are useful when testing is limited and can lead to improved data-driven decision making.

Direct surveys are an obvious approach to estimate incidence. A number of these direct surveys to obtain health data have been deployed in various countries in recent months [5, 10, 14]. While these surveys can gather useful data, a large number of participants is needed to achieve reliable estimates. Additionally, these surveys collect sensitive personal health information, which prevents the distribution of the raw data collected, and may also discourage privacy-concerned people from responding.

In this paper, as an alternative to direct surveys, we propose online surveys with *indirect reporting*, where the questions a participant answers are not about herself, but about her contacts. This technique is known in the literature as the *network scale-up method* [1, 2], and has been successfully used for public health. This approach has at least two major advantages with respect to direct surveys. First, the survey can be designed so that no personal information is collected from the participant (i.e., it is completely anonymous). Second, indirect reporting has a multiplicative effect, since it reduces the number of responses required to achieve a specific population coverage. The loss in accuracy, due to respondents not always having exact information about the health status of others, may be compensated by the significant increase in coverage (i.e., on average each respondent informs about the likely status of around 100 other people, and this number is rather stable across countries).

We have designed and deployed a system that implements the network scale-up method via online anonymous surveys with indirect reporting, as part of the CoronaSurveys project [15]. This system has surveys in multiple languages and allows reporting data on the incidence of COVID-19 in all countries. The project team has

promoted the survey via online social networks and personal contacts. The system has been collecting data for more than 2 months now, and has collected more than 13, 000 responses, reporting cases in 70 countries. All the collected data is available to be openly used[1]. In parallel with the data collection, the CoronaSurveys project has been developing statistical techniques to estimate the incidence of COVID-19 in different countries and geographical areas.

In the rest of this paper, we present the different elements of the CoronaSurvey project, and compare our resulting estimates with those obtained via other indirect methods and a wide systematic serology study conducted in Spain [6]. Our estimates are surprisingly close to the values obtained in the serological study[2]. From this, we conclude that anonymous open indirect surveys, in combination with our proposed statistical techniques, provide a cheap and flexible option for monitoring epidemics, especially in countries with limited infrastructure.

2 ESTIMATES VIA ANONYMOUS SURVEYS WITH INDIRECT REPORTING

This section describes the two main components of the Corona-Survey project: collection of survey responses and estimation of COVID-19 incidence from the collected responses.

2.1 Data Collection: The Survey

Our proposed surveys are carefully designed to avoid querying participants about their own health status[3], identity, or any personal data. In order to provide data for estimating incidence, participants answer three questions. First, they select a geographical area, which can be a whole country or a region within a country. (The participant does not need to be a resident of that area.) Then, they answer two simple questions about that area:

- How many people do you know in this geographical area? (Please, consider only people whose current health status you likely know.)
- As far as you know, how many of the above have had symptoms compatible with COVID-19 (or were diagnosed with the disease)?

We aim at increasing participation by not asking for any personal information (protecting the participant's privacy), and by having just two questions[4]. However, the lack of detailed information about the participants makes the estimation process more challenging. In particular, we do not control the spread of the survey and do not have means to ensure that there is an adequate coverage, in terms of regions, age groups and other demographic factors.

The main novelty of our proposed survey is that a participant does not report on her own health status but on those of others. This typically leads each participant to report on the health status of a large sample (around one hundred on average), which significantly increases our coverage of the population. We believe that this is the reason why, somewhat surprisingly, even with the limitations in the available data and few responses, we can still obtain estimates that are less than 4% away from the real value (see Section 3.3). We believe that indirect reporting compensates significantly the biases in the set of participants. Obvious advantages of this approach are that it is very simple to deploy and can give very timely results.

We started the project by running discrete surveys in Spain and evolved the system so that now it collects data continuously. The survey was initially done via Twitter, but was quickly moved to Google Forms. For privacy reasons, in May 2020 it has been moved to a dedicated server running the web-based Limesurvey system. We have been running the survey (starting with a simpler version) in some countries for more than 3 months (since March 13). The survey is now available in 57 languages. Participants can report at the regional level in 149 countries and at the country level in all countries of the world. We have already collected more than 15, 000 responses for almost 90 countries. This participation has been obtained by advertising and promoting the survey in online social networks, and via personal contacts. No incentive (economic or otherwise) has been used to promote participation. In some countries (e.g., Spain, Portugal, Ukraine, Brazil) the project received media attention, which led to bursts in participation at specific times. These burst do not seem to influence the results.

2.2 Computing Incidence Estimates

Country-wide and regional estimates. We use responses to our survey to estimate the incidence of COVID-19 in different countries. Respondents can report on the number of people they know for a specific region or for the whole country. Currently, when the data availability allows it, estimates for a given country are computed based only on the regional responses, because we observed that country-wide responses tend to introduce a geographical bias.

For instance, our initial estimates for Spain were ignoring regional information, and considering all responses as referring to the whole country. However, we observed that most of our responses were actually from the region of Madrid (very likely because the Spanish team members are based in Madrid), which is the most affected region in Spain, and thus tended to report high incidence ratios. This was in line with the prevalence in Madrid, but not necessarily a good reflection of country-wide conditions. We then computed new estimates by taking regional information into account and confirmed that we were initially over estimating the number of cases in Spain.

We also decided to use only the responses that specify a region as it is reasonable to assume that country-wide responses will share a similar bias (coming from the most affected region)[5]. Note that while we observed this bias in one specific dataset (Spain), similar problems could arise in other countries. COVID-19 outbreaks have been fairly localized in most countries, with different prevalence observed at the region or state level. Thus, answers that provide information within a region tend to lead to more reliable estimates

---

in general. We therefore plan to remove the country-wide option in future survey updates.

COVID-19 incidence estimation – Region based approach. Assume the country of interest is divided into k regions. In each region i, we get n_i responses, where each response j contains:

- a **reach** variable, r_i^j, which is the network size of participant j (i.e., the number of people whose health she knows; answer to the first survey question) and
- a **count** variable, c_i^j, which is the number of people (out of the r_i^j reached) that are known to the respondent to be showing symptoms compatible with COVID-19 (answer to the second survey question),

for $j \in \{1, \ldots, n_i\}$. Then, we obtain an estimate of the ratio of people infected with symptoms in region i, $\hat{p}_i$, using [1]

$$\hat{p}_i = \frac{\sum_{j=1}^{n_i} c_i^j}{\sum_{j=1}^{n_i} r_i^j}.$$

From the ratios $\hat{p}_i$ of the different regions, we compute an estimate of the proportion of those infected with symptoms in the country, $\hat{p}$, using a weighted sum of these ratios as follows.

$$\hat{p} = \sum_{i=1}^{k} \omega_i \hat{p}_i,$$

where ω_i is a proportional weight defined as

$$\omega_i = \frac{N_i}{N},$$

where N_i and N are the populations of region i and the whole country, respectively. To build a confidence interval for $\hat{p}$, we need to estimate its variance $V(\hat{p})$. To do so, we consider each region as a stratum in a post-stratified random-sampling setting, and estimate the variance of proportions for post-stratified sampling [8] as

$$V(\hat{p}) = \frac{1-f}{n} \sum_{i=1}^{k} \omega_i S_i^2 + \frac{1-f}{n^2} \sum_{i=1}^{k} (1 - \omega_i) S_i^2,$$

where $n = \sum_i^k n_i$ and $f = n/N$. The value S_i^2 for each region can be estimated using

$$S_i^2 = \sum_{j=1}^{n_i} \frac{(p_{ij} - \hat{p}_i)^2}{n_i - 1},$$

where $p_{ij} = \frac{c_i^j}{r_i^j}$. We can then construct a confidence interval for $\hat{p}$ as

$$\hat{p} \pm 1.96\sqrt{V(\hat{p})}.$$

Before generating an estimate for a given country at a given date, we first clean the survey responses by identifying and removing outliers. We declare a response to be an outlier if r_i^j, the number of persons that the participant claims to know, is unusually large (specifically, we remove entries where r_i^j is beyond 1.5 times the interquartile range above the upper quartile). We also consider to be outliers responses leading to a large ratio p_{ij} of symptomatic people reported (specifically, we remove entries in which the ratio c_i^j / r_i^j is above 0.3). We remove responses with large ratios because we aim at surveying the general population, and not individuals

(such as doctors or nurses) who may be in contact with a large number of symptomatic cases.

After outlier removal, for any given day we aggregate data from that day and from previous days until we have at least $A_{\min}$ responses. For the experiments we report here we set $A_{\min} = 300$, as we observed empirically that this provides enough data to make our estimate reliable. Since we usually do not get $A_{\min}$ responses on a given day for most countries, using data from previous days also provides a rolling estimate, which has an implicit smoothing effect on the estimate. From these responses, we exclude country-wide responses (as discussed earlier) and then compute the value of $\hat{p}_i$ for each region i represented in the responses. Then, the estimate of the proportion of people infected with symptoms in the country, $\hat{p}$, for the day is computed as shown above.

The described procedure produces reasonable estimates and works well as long as we have a sufficiently large number of responses (per day). Hence, we have only applied this procedure to generate estimates from data gathered in Spain, Portugal, and Ukraine, countries from which we got the highest number of responses.

COVID-19 incidence estimation – Country based approach. For countries where the number of survey responses is smaller (after removing outliers), we have used a simpler estimation procedure. Denote n_d the number of responses collected on day d for the given country (counting both regional and country-wide responses after removing outliers). If $n_d \geq a_{\min}$, we estimate the incidence of COVID-19 in the country of interest using

$$\hat{p}_d = \frac{\sum_{l=1}^{n_d} c_l}{\sum_{l=1}^{n_d} r_l},$$

where r_l is the number of people a participant l declares to know in the first question of the survey, and c_l is the number of people (out of r_l) showing symptoms compatible with COVID-19, for $l \in \{1, \ldots, n_d\}$. On the other hand, if $n_d < a_{\min}$, we do not compute an estimate for day d. We instead aggregate the responses for day d to the responses for the subsequent days $d + 1, d + 2, \ldots$ until we have a day $d' = d + m$ such that $n_m = n_d + n_{d+1} + \ldots + n_{d+m} \geq a_{\min}$ responses. In the experiments reported here we empirically chose $a_{\min} = 30$. We then compute the estimate $\hat{p}_{d'}$ for day d' as[6]

$$\hat{p}_{d'} = \frac{\sum_{l=1}^{n_m} c_l}{\sum_{l=1}^{n_m} r_l}.$$

The estimates $\hat{p}_d$ and $\hat{p}_{d'}$ are simple proportions and we construct a 95% confidence interval for $\hat{p} \in \{\hat{p}_d, \hat{p}_{d_m}\}$ using confidence intervals for binomial proportions as follows

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1 - \hat{p})}{r}},$$

where $r = \sum_l r_l$. A major limitation of this method is that we do not obtain estimates for every day. However, we are constantly refining these techniques, and trying new ways to obtain better estimates.

---

[6]We expect that better results may be achievable if $A_{\min}$ and $a_{\min}$ are selected as a function of country population. We plan to investigate this in future work.

3 VALIDATION

In the previous sections we introduced our strategy for indirect reporting of the number of cases. Given the nature of an open and anonymous survey, where data quality is much harder to enforce, we soon identified the need for an independent estimator to which we could compare the survey results. In most countries we had access to time series with the number of official, RT-PCR confirmed, cases and COVID-19 mortality data. Together these series allow the derivation of a naive *case fatality ratio* (CFR).

Up to early May 2020 our running option for an independent estimator was to use the CFR to estimate the current number of cases in each country, as detailed in the next section. On May 13, 2020, a large-scale serology study was reported for Spain [6] and this provided a more precise and direct data collection for calibration.

3.1 Inferring Cases from Reported Mortality

In an ongoing epidemic, the current CFR should be calculated by taking into account the number of deaths (d) over the number of current cases with known outcomes (c), since very recent cases can still evolve as fatalities or recover [13]. This correction yields a corrected CFR (cCFR), more accurate than the naive CFR that is often reported. To perform the correction we follow the same methodology that is described in [16] as well as their code in https://github.com/thimotei/CFR_calculation. We keep the same estimates for the delay from case confirmation to death, i.e., a Lognormal distribution with a mean delay of 13 days and a standard deviation of 12.7 days. (Our code is available under https://github.com/GCGImdea/coronasurveys.)

Under the assumption that the disease has similar mortality rates for similar populations, it is possible to use stable *cCFR* estimates to obtain a baseline, $cCFR_b$, for COVID-19 and check how each country's current *cCFR* compares to that baseline, and hence infer the proportion of cases that are being detected. In [16] the authors keep a frequently updated estimate of the level of under-reporting for several countries. As an example if $cCFR_b = 1\%$ and a country exhibits a $cCFR = 2\%$, they infer that only 50% of the cases are being detected. For our purposes, if we multiply the reported number of cases by $\frac{cCFR}{cCFR_b}$, we obtain an estimate for the likely true number of cases in that given country. In the example above we would have multiplied the reported number of cases by $\frac{cCFR}{cCFR_b} = 2$.

The next step is to select a reference baseline. In [17] the authors report, from a large sample of cases in China, a baseline $cCFR_b$ of 1.38% ($d_b = 1,023$; and $c_b = 74,130$), and we have chosen to use this baseline in our estimators. Still, given that several countries (North Korea, New Zealand, and others) have stabilized the growth of COVID-19 cases, it is possible to use the data in those countries to also define a baseline. Following the Ln-Method in [9], we also construct a confidence interval for the ratio $\frac{cCFR}{cCFR_b}$, at any given date. We model the number of deaths d using a binomial distribution with parameters $cCFR$ and c (analogously, we refer to d_b, $cCFR_b$, and c_b for our fixed baseline). Then, it can be shown that $\ln\left(\frac{d/c}{d_b/c_b}\right)$ is approximately normally distributed with mean $\ln\left(\frac{cCFR}{cCFR_b}\right)$ and estimated variance $\hat{\sigma}^2 = 1/d - 1/c + 1/d_b - 1/c_b$. Finally, a 95% CI for

our ratio is given by $(r \cdot \exp(-1.96 \cdot \hat{\sigma}),\ r \cdot \exp(1.96 \cdot \hat{\sigma}))$, where r is the observed value of $\frac{d/c}{d_b/c_b}$.

We observe that one limitation of resorting to reported mortality data is that some countries might not properly report or classify it as COVID-19 mortality. However, in countries with adequate reporting, it is a useful source of calibration and independent estimation.

3.2 Serology Study in Spain

On May 13, 2020 the Spanish Government published a first report from a large-scale serology study that looked for COVID-19 antibodies in the population [6]. Samples were collected from April 27 to May 11, and results made available from a group of 60, 897 participants in the study, selected according to demographic criteria, to obtain a representative sample of the population. Although both IgM and IgG antibodies were measured, the report focuses on the prevalence of SARS-Cov2 IgG antibodies. The overall prevalence was reported as 5.0% (95% CI: 4.7%-5.4%), and regional variations ranged from 1.1% in Ceuta to 11.3% in the *Comunidad de Madrid.*

The IgG test was found to have a sensitivity of 79% and specificity of 100%. Given that, we can correct for false negatives and estimate an overall infection rate of 6.33% (=0.05/0.79), with approximate 95% CI: 5.95%-6.84%. Assuming an average time span of two weeks since infection to the development of detectable levels of IgG [11], and with a population in Spain of 46, 934, 628 persons, this leads to an estimate of approximately 2, 970, 546 (95% CI: 2,792,313-3,208,190) cumulative cases around the weeks from April 13 to 27 (in the middle of this period, on April 20, the number of RT-PCR confirmed cases was 200, 210, almost 15 times lower). Using the cumulative mortality on May 11 (roughly two weeks after the likely infection dates) of 26, 744, this leads to an estimated *infection mortality rate* (IFR) of 0.9% (95% CI: 0.83%-0.96%)[7]. In Brasil, another large-scale serology survey (with 25,025 participants) [7] recently provided an estimated IFR of 1%, a value that is in line with our estimate for Spain.

The Spanish study includes data on the proportions of IgG positives that had one or more symptoms, or were completely asymptomatic. The proportion of IgG positives with some kind of symptoms (i.e., at least one symptom) was reported to be 66.27%. Fixing this percentage, and assuming that most of the (officially reported) RT-PCR confirmed cases were symptomatic, one can estimate about 1, 968, 550 persons with symptoms, and a CFR of around 1.36% (actually, very close to the value calculated for Wuhan at 1.38% [17]).

3.3 Comparing Estimates

Adopting as ground truth the serology-derived value of 6.33% cumulative infections in the period from April 13 to 27 (which lead to 79% IgG positives two weeks later, i.e. 5%), we compare this ground-truth to our estimates in the same period. Since our estimates, cCFR-based and survey-based, target symptomatic cases, we need to scale them to total infections by dividing the estimates by 0.66 (the ratio of symptomatic over total infections as reported in the study [6]). We also show, for comparison, the number of reported RT-PCR confirmed cases.

---

[7] The data of confirmed cases and mortality was extracted from [4] on the date of the submission (May 20, 2020).
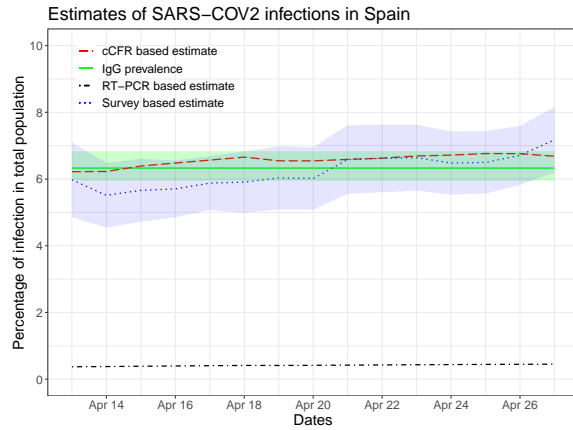
Figure 1: Case estimates with 95% confidence bands for Spain, April 13 to 27, 2020.
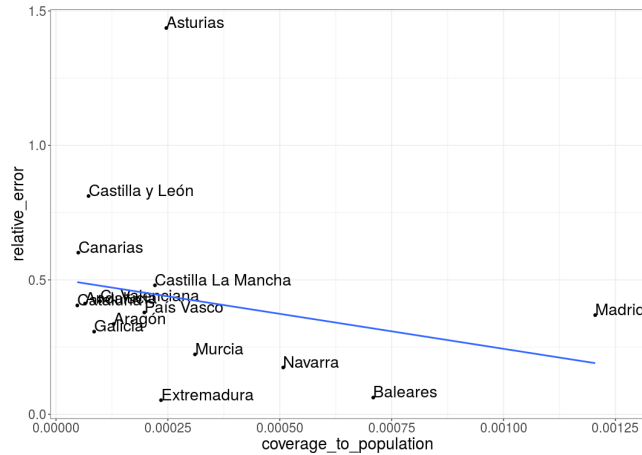

Figure 2: Relative accuracy vs. local coverage: For each region, we plot the relative error in our estimates as compared to serology data, shown as a function of the relative reach within the region. Higher reach (better coverage within a region) tends to lead to lower error.

As we can see in Figure 1 the official RT-PCR data is about an order of magnitude less than the likely true rate of infected people that is inferred from IgG prevalence, 6.33%. Figure 1 also shows that both the cCFR-based estimate and the estimate derived from the open survey closely track the IgG-prevalence value. Furthermore, the confidence bands intersect throughout the period, with a clear difference in amplitude due to the sample sizes. The average of the cCFR estimates in the 15-day period was 6.56%, while that of the survey-based estimates was 6.2%. This places them, respectively, only 0.24% and 0.13% apart from the IgG reference value (average relative difference of 3.72% and 2.05%, respectively). These results show that open surveys can bring relevant data on the size of a pandemic, which can be be useful when more reliable metrics are

not yet available or better estimates cannot be implemented in some regions.

For further evaluation of our method we compare our estimates to the serology data for each region in Spain as shown in Figure 2. We plot the relative error in our estimate for a region as a function of the relative coverage of our surveys for that region (reach divided by population). Although we can see some variability, we note that the trend is towards lower relative error as the reach increases.

4 EXPERIENCE USING OPEN SURVEYS IN LMI COUNTRIES

We have obtained estimates for a number of countries. In particular, we have received enough responses to estimate symptomatic cases in 3 LMICs, namely Brazil, Ecuador, and Ukraine (see Tables 1 and 2).

In Brazil we have recent estimates for May 17, when the official number of cumulative confirmed cases was 233, 142, and the official number of cumulative fatalities was 15, 633. For that same date, the estimate based on the cCFR is 2, 139, 681 (CI 95%: 2, 135, 408 - 2, 143, 963), 1.01% of the population, and the estimate based on 41 survey responses is 2, 120, 134 (CI 95%: 1, 195, 676 - 3, 044, 593), 1% of the population. Not surprisingly, the estimated number of cases is one order of magnitude larger than the number of confirmed cases. However, it is remarkable how close the cCFR and survey-based estimates are to each other, differing in less than 20, 000 cases (0.01% of the population).

In Ecuador we also observe that the number of estimated cases is at least one order of magnitude larger than the official number of confirmed cases. However, we observe that the estimates from cCFR and from the surveys are also very different. For instance, we have estimates dated April 15, when the official number of confirmed cases was 7, 603 and the official number of fatalities was 355. Our case estimate for that date from the cCFR is 53, 435 (CI 53, 069 - 53, 804), 0.31% of the population, and the one from 30 survey responses is 274, 668 (CI 190, 236 - 359, 100), 1.61% of the population. We observe a significant difference between the two estimates, the one from the surveys being 5 times larger than the cCFR-based one. This difference does not seem to be the result of geographical bias, since few survey responses came from the provinces with the largest number of cases (e.g., Guayas).

We observed a similar behavior in Ukraine where, again, the estimates are at least one order of magnitude larger than the number of confirmed cases, and the number of estimated cases from the surveys is one order of magnitude larger than the estimates from cCFR. For instance, the latest direct estimate from (30) survey responses was done on April 26, and has a value of 246, 646 (CI 107, 482 - 385, 811), 0.56% of the population, while the cCFR estimate is 32, 078 (CI 31, 734 - 32, 426), 0.07% of the population. The confirmed numbers of cases and fatalities on that date were 8, 617 and 209, respectively. This result is not due to a geographical bias, since the region-based estimate for that same day from 300 survey responses is 159, 529 (CI 62, 361 - 256, 696), 0.35% of the population, which is lower, but still five times higher than the cCFR-based estimate.

The results in Ecuador and Ukraine are puzzling, and we are not able to explain them yet. Our current hypothesis is that these

Country	Date	Cases	Fatalities	cCFR (CI 95%)	Resp.	CoronaSurveys (CI 95%)
Brazil	May 17	233, 142	15, 633	2, 139, 681 (2, 135, 408 - 2, 143, 963)	41	2, 120, 134 (1, 195, 676 - 3, 044, 593)
Ecuador	April 15	7, 603	355	53, 435 (53, 069 - 53, 804)	30	274, 668 (190, 236 - 359, 100)
Ukraine	April 26	8, 617	209	32, 078 (31, 734 - 32, 426)	30	246, 646 (107, 482 - 385, 811)

Table 1: Summary of estimates for Brazil, Ecuador and Ukraine. Cases and Fatalities correspond to the official data on that day, while cCFR and CoronaSurveys are estimates.

Country	Date	Cases	cCFR	CoronaSurveys
		% pop.	% pop. (CI 95%)	% pop. (CI 95%)
Brazil	May 17	0.11%	1.01% (1.00 - 1.01)	1.00% (0.56 - 1.43)
Ecuador	April 15	0.04%	0.31% (0.31 - 0.31)	1.61% (1.11 - 2.10)
Ukraine	April 26	0.02%	0.07% (0.07 - 0.07)	0.56% (0.25 - 0.88)

Table 2: Summary of estimates for Brazil, Ecuador and Ukraine in percentage of the country population.

countries use different criteria for reporting cases and fatalities than the countries we use as reference. This may cause the cCFR-based estimate to be unreliable. We have deployed additional questions in the survey that we believe can be used to track the difference. The good news is that this can be done with very little effort, and we do not need a lot of responses to have enough information to have solid conjectures.

5 DISCUSSION

By now it is clear that relying only on confirmed cases and fatalities to measure the true size of a growing pandemic is not a good idea. It is possible to use this data to derive estimates, like the one we obtain here based on the cCFR, that are reasonably reliable in countries with a good reporting system. This has been shown in the case of Spain using the ground truth provided by a serological study. The same study indicates that open anonymous surveys with indirect reporting also provide estimates that are close to the real values. The matching between cCFR-based and survey-based estimates has also been observed in Brazil. However, we have found countries in which there is no ground truth, and the cCFR-based and survey-based estimates differ significantly. We are investigating this further by adding new questions to the surveys in these countries, which will provide additional information on the causes of these discrepancies.

We are aware that having open anonymous surveys prevents a tight control of the population of participants. Hence, our responses may be suffering from strong biases: for example, they may be close geographically and socially to the team members, which are the ones promoting the survey. We are also aware that the set of people and cases participants report are possibly not disjoint. Interestingly, until now the only bias that we have observed to be relevant is the geographical bias. Initial studies done via simulation hint that the intersection of contacts among participants does not seem to have much influence on the estimates. However, more experiments are required. In fact, new questions have been added to the survey to evaluate the level of overlapping, and estimate the propagation graph of participation.

However, we plan to explore ways to influence and select the population of participants. One line to follow is the use of targeted campaigns in social networks and web ads (using Facebook Ads or

Google Ads), in which we have control over who sees a given ad promoting the survey. This will allow targeting participants from certain geographical areas and certain demographic profiles. In fact, we have started a exploratory campaign using Facebook Ads in Brazil.

While we only present here techniques for obtaining estimates on the number of people infected with COVID-19, we are designing new surveys that will hopefully allow us to estimate other important parameters of the pandemic, like the number of newly infected cases, the reproduction number, or the forecast needs of health equipment and infrastructures.

We believe that a survey system like the CoronaSurveys project is especially suited for LMICs, since the cost of preparing and deploying a survey is extremely small, participants can use very simple devices to fill the survey (since it is web based), and the number of participants required to have information on the pandemic is rather low. Having a uniform approach to obtain this information in many countries also makes the process especially interesting, since it avoids the current problems with different ways of counting and measuring.

6 ACKNOWLEDGEMENTS

REFERENCES

[1] H Russell Bernard, Tim Hallett, Alexandrina Iovita, Eugene C Johnsen, Rob Lyerla, Christopher McCarty, Mary Mahy, Matthew J Salganik, Tetiana Saliuk, Otilia Scutelniciuc, et al. 2010. Counting hard-to-count populations: the network scale-up method for public health. *Sexually transmitted infections* 86, Suppl 2 (2010), ii11–ii15.

[2] H Russell Bernard, Eugene C Johnsen, Peter D Killworth, and Scott Robinson. 1991. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social science research* 20, 2 (1991), 109–121.

[3] Crowdfight COVID-19. [n.d.]. An initiative from the scientific community to put all available resources at the service of the fight against COVID-19. https://crowdfightcovid19.org/. Accessed: 2020-05-20.

[4] European Centre for Disease Prevention and Control. [n.d.]. COVID-19. https://www.ecdc.europa.eu/en/covid-19-pandemic. Accessed: 2020-04-09.

[5] Facebook Data for Good. [n.d.]. COVID-19 Symptom Survey – Request for Data Access. https://dataforgood.fb.com/docs/covid-19-symptom-survey-request-for-data-access/. Accessed: 2020-05-20.

[6] Consumo y Bienestar Social Gobierno de España, Ministerio de Sanidad. 2020. *Estudio Nacional de sero-Epidemiología de la Infección por SARS-CoV-2 en España, Informe PRELIMINAR tras la primera ronda.* Technical Report. Gobierno de España, Ministerio de Sanidad, Consumo y Bienestar Social.

[7] Pedro Hallal, Fernando Hartwig, Bernardo Horta, Gabriel D Victora, Mariangela Silveira, Claudio Struchiner, Luis Paulo Vidaletti, Nelson Neumann, Lucia C

Pellanda, Odir A Dellagostin, Marcelo N Burattini, Ana M Menezes, Fernando C Barros, Aluisio J Barros, and Cesar G Victora. 2020. Remarkable variability in SARS-CoV-2 antibodies across Brazilian regions: nationwide serological household survey in 27 states. (2020). https://doi.org/10.1101/2020.05.30.20117531

[8] D. Holt and T. M. F. Smith. 1979. Post Stratification. *Journal of the Royal Statistical Society* (1979).

[9] P. A. R. Koopman. 1984. Confidence Intervals for the Ratio of Two Binomial Proportions. *Biometrics* 40, 2 (1984), 513–517. http://www.jstor.org/stable/2531405

[10] Manuel Linares, Ignacio Garitano, Laura Santos, and José M Ramos. 2020. Estimando el número de casos de COVID-19 a tiempo real utilizando un formulario web a través de las redes sociales: Proyecto COVID19-TRENDS. *Semergen* (4 2020).

[11] Quan-Xin Long, Bai-Zhong Liu, Hai-Jun Deng, Gui-Cheng Wu, Kun Deng, Yao-Kai Chen, Pu Liao, Jing-Fu Qiu, Yong Lin, Xue-Fei Cai, et al. 2020. Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nature Medicine* (2020), 1–4.

[12] A Maxmen. 2020. How much is coronavirus spreading under the radar? *Nature* (2020). https://www.nature.com/articles/d41586-020-00760-8.

[13] Hiroshi Nishiura, Don Klinkenberg, Mick Roberts, and Johan AP Heesterbeek. 2009. Early epidemiological assessment of the virulence of emerging infectious diseases: a case study of an influenza pandemic. *PLoS One* 4, 8 (2009). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006852.

[14] Nuria Oliver, Xavier Barber, Kirsten Roomp, and Kristof Roomp. 2020. The Covid19Impact Survey: Assessing the Pulse of the COVID-19 Pandemic in Spain via 24 questions. *CoRR* abs/2004.01014 (2020). arXiv:2004.01014 https://arxiv.org/abs/2004.01014

[15] The CoronaSurveys research team. [n.d.]. CoronaSurveys: Monitoring the Incidence of COVID-19 via Open Surveys. https://coronasurveys.org/. Accessed: 2020-05-20.

[16] T Russel, Joel Hellewell, S Abbot, et al. 2020. Using a delay-adjusted case fatality ratio to estimate under-reporting. *Available at the Centre for Mathematical Modelling of Infectious Diseases Repository, here* (2020). https://cmmid.github.io/topics/covid19/global_cfr_estimates.html.

[17] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick G T Walker, Han Fu, Amy Dighe, Jamie T Griffin, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Anne Cori, Zulma Cucunubá, Rich FitzJohn, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Daniel Laydon, Gemma Nedjati-Gilani, Steven Riley, Sabine van Elsland, Erik Volz, Haowei Wang, Yuanrong Wang, Xiaoyue Xi, Christl A Donnelly, Azra C Ghani, and Neil M Ferguson. [n.d.]. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases* (2020/04/11 [n. d.]). https://doi.org/10.1016/S1473-3099(20)30243-7

[18] Peng Yang, Chunna Ma, Weixian Shi, Shujuan Cui, Guilan Lu, Xiaomin Peng, Daitao Zhang, Yimeng Liu, Huijie Liang, Yi Zhang, et al. 2012. A serological survey of antibodies to H5, H7 and H9 avian influenza viruses amongst the duck-related workers in Beijing, China. *PLoS One* 7, 11 (2012). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0050770.