frontiers

Estimating the COVID-19 Prevalence in Spain with Indirect Reporting via Open Surveys

Augusto Garcia-Agundez [1,*], Oluwasegun Ojo [2], Harold Hernandez [3], Carlos Baquero [4], Davide Frey [5], Chryssis Georgiou [6], Mathieu Goessens [7], Rosa Lillo [3], Raquel Menezes [8], Nicolas Nicolaou [9], Antonio Ortega [10], Efstathios Stavrakis [9] and Antonio Fernandez Anta [2]

[1] *Multimedia Communications Lab, etit, TU Darmstadt, Darmstadt, Germany*
[2] *IMDEA Networks Institute, Madrid, Spain*
[3] *Department of Statistics, University Carlos III de Madrid, Madrid, Spain*
[4] *Departmento de Informatica, University of Minho, Braga, Portugal*
[5] *Inria Centre de Recherche Rennes Bretagne Atlantique, Rennes, France*
[6] *Department of Computer Science, University of Cyprus, Nicosia, Cyprus*
[7] *IMT Atlantique, Nantes, France*
[8] *Departmento de Matematica, University of Minho, Braga, Portugal*
[9] *Algolysis Ltd, Limassol, Cyprus*
[10] *Department of Electrical and Computer Engineering, USC Viterbi School of Engineering, Los Angeles, CA, USA*

Correspondence*:
Augusto Garcia-Agundez
augusto.garcia@kom.tu-darmstadt.de

ABSTRACT

During the initial phases of the COVID-19 pandemic, accurate tracking has proven unfeasible. Initial estimation methods pointed towards case numbers that were much higher than officially reported. In the CoronaSurveys project, we have been addressing this issue using open online surveys with indirect reporting. We compare our estimates with the results of a serology study for Spain, obtaining high correlations (R squared 0.89). In our view, these results strongly support the idea of using open surveys with indirect reporting as a method to broadly sense the progress of a pandemic.

Keywords: COVID-19, pandemic, serology, survey, indirect reporting, sensing

1 INTRODUCTION

During the initial phases of the COVID-19 pandemic, progress tracking via massive serology testing has proven to be unfeasible. However, initial estimation methods suggested that the real numbers of COVID-19 cases were significantly higher than those officially reported (1). For instance, by April 30th, 2020, the number of confirmed fatalities due to COVID-19 in the US was $66,028$, and the number of confirmed cases was $1,080,303$. However, with that number of fatalities the number of cases must have been no less than $4,784,637$, by simply using the Case-fatality Ratio (CFR) of 1.38% measured in Wuhan (2).

In the case of Spain, the discrepancy seems to be even higher. Preliminary studies point towards only one in 53 cases being reported during the first days of the pandemic (3). Although recent availability of massive

testing has reduced this discrepancy, demographic statistics still indicate a degree of underreporting to this day, which can be seen among others in mortality numbers: all-cause mortality statistics in Spain point to two periods of significant excess of deaths in the country over the predicted values in 2020: March and April ($44,599$ deaths in excess) and August to December ($26,186$ deaths in excess) (4). These numbers contrast with the officially reported number of deaths due to COVID-19, which rests at $50,837$ (5). This discrepancy is corroborated in publications from official government authorities, which indicate an ongoing estimated underreporting of 20% to 40% (6).

In the CoronaSurveys project, (7) we aim to track the progress of the pandemic using online, open, anonymous surveys with indirect reporting. Recent articles have also suggested the use of surveys to monitor the pandemic, both for Spain (8, 9) and globally (10). However, to our knowledge, all surveys conducted in Spain have employed direct reporting only, asking participants about themselves. CoronaSurveys implements the network scale-up method of indirect reporting instead, allowing us to collect data on a wide fraction of the population with a small number of responses and in a very short time-frame (11). In this article, we compare the accuracy of CoronaSurveys with a gold standard: serology testing data collected by the Spanish government in the ENE-COVID study (12).

2 METHODS

The survey deployed in the CoronaSurveys project, which can be answered via browser or mobile app, includes two questions:

1. *How many people do you know in your area for which you know their health condition?* The answer to this question by participant i is the *Reach* r_i.
2. *How many of those were diagnosed with or have symptoms of COVID-19?* The answer to this question by participant i is the *Cumulative Number of Cases* c_i.

In the CoronaSurveys project we have focused on simplicity and brevity to maximize interest and retain users that would consistently provide data every few days. For that reason the total number of questions in the survey has been kept small at all times. Our approach yielded good initial results with about 200 responses per week. The survey has been promoted via social networks via direct contacts and, more recently, with paid advertising. To ensure total anonymity, the surveys are hosted on a private instance of LimeSurvey (13). Data is aggregated daily, and in this process the responses are shuffled so no single entry can be back-traced to its user. All the data is published in a public Github repository. The study design was reviewed and approved by the ethics committee of the IMDEA Networks Institute. The survey includes an informed consent.

Once the data is collected, we remove outlier responses. A response is considered an outlier if (1) r_i is outside 1.5 times the interquartile range above the upper quartile (which for the data in this paper means $r_i > 175$) or if (2) c_i/r_i is greater than $1/3$ (to exclude participants with an exceptionally high contact with cases). For this paper we only consider responses in which participants provide information for their region. Hence, the data is aggregated by region for all participants, to obtain the estimator of COVID-19 prevalence $(\sum_i c_i)/(\sum_i r_i)$ (11).

3 RESULTS

To evaluate the accuracy of this method to sense the cumulative number of cases of COVID-19, we compare our estimates with the results of the serology study of Pollán et al. (12) for Spain. We exclude Ceuta and Melilla due to lack of data on our part. Conducted between April 27 and May 11, 2020, the serology study

provides data for $n = 61,075$ participants ($0.1787\% \pm 0.0984\%$ of the regional population, and 0.1299% of the national population). We consider as positive cases those that tested positive to the point-of-care or immunoassay IgG tests (Supplementary Table 6 in Pollán et al. (12), column *Either test positive*).

For our estimates, we consider the (up to) 100 most recent survey responses per region on April 20. The date is chosen because the mean period between illness onset and a 95% confidence of IgG antibodies presence is 14 days (14). This results in $n = 999$ responses (59 ± 35 per region) across Spanish regions, with a cumulative reach of $\sum_i r_i = 67,199$ ($0.1827\% \pm 0.0701\%$ of the regional population, and 0.1434% of the national population).

4 DISCUSSION

The Bland-Altman plot in Figure 1A shows a high correlation between the CoronaSurveys estimates and the gold standard. A direct comparison of crude percentages, depicted in Figure 1B, also yields excellent results ($R^2 = 0.8994$). The linear regression equation points to CoronaSurveys very consistently underestimating the number of cases by a factor of approximately 46%, possibly due to asymptomatic cases. This ratio is consistent with the estimates of the Covid19Impact study of Oliver et al. (9), which used more than $140,000$ direct survey responses collected on March 28th-30th. It is also consistent with the reported data on asymptomatic cases reported by Pollán et al. (12), which found that around a third of the seropositive participants were asymptomatic. Table 1 presents a detailed comparison of the estimates per region obtained in the different studies.

Figure 2A presents how the number of replies per region affects the resulting value of R^2. This analysis indicates that 50 responses per region can already offer a reasonable estimation of cases. Including more replies may increase accuracy further, but the numbers remain reasonably stable. Naturally, it is important that replies are well distributed across all regions. Figure 2B depicts the effect of the day limit on R^2 if we consider a date of $\pm$ one week. Theoretically, a bell curve centered on the 20th should be expected, as estimating too early would imply too few cases are reported, and estimating too late would include more cases. We indeed observe an impact on accuracy, and the left half of the bell curve is more visible. The change in accuracy is mostly due to responses collected on April 16th. The lack of the right half of the bell curve is due to the low number of daily responses after April 16th, which implies that the daily estimates are computed with sets of responses with large intersections.

Interestingly, a similarly high number of responses was collected on April 14th, with nearly no impact on accuracy. We believe this is due to the distribution of the responses. As depicted in Figure 3, additional responses from regions where many are already available will barely have an impact on the global result. As the great majority of contributions for April 14th were for Madrid, where we already had many responses available, the 77 new responses on April 14th barely had any impact.

Our study presents a number of limitations. Firstly, as presented in Table 1, our number of responses in some regions was limited (e.g., 9 responses in La Rioja or 16 in Navarra and Cantabria). Our own analysis suggests this is not enough to offer reliable data for these three regions. Additionally, our criteria to eliminate outliers is heuristic, and may change in the future as we collect more data.

Nevertheless, despite these limitations, the estimates obtained in CoronaSurveys show high correlation with serology tests. Moreover, since the underestimation of our estimates over all regions is homogeneous, and consistent with the one third fraction of asymptomatic reported by Pollán et al. (12), these estimates can be "corrected" to provide an accurate cumulative number of cases for each region. We will further

evaluate the robustness of our model as Pollán et al. publish the results of their three additional serology studies.

In summary, we believe these results strongly support using open surveys with indirect reporting as a method to broadly sense the progress of a pandemic.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

The analysis presented in this article was conducted by Augusto Garcia-Agundez and Antonio Fernandez Anta with support and feedback from all remaining co-authors. The data acquisition and processing techniques were developed by all co-authors.

FUNDING

At the time of writing this article, CoronaSurveys has received no public funding. Social networks surveys have been partially funded via donations through our website. CoronaSurveys received an award from the UMD/CMU COVID-19 Symptom Data Challenge.

ACKNOWLEDGMENTS

We would like to thank all CoronaSurveys researchers and collaborators for their contribution to this project: https://coronasurveys.org/team/.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found in the CoronaSurveys Github Repository at https://github.com/GCGImdea/coronasurveys.

REFERENCES

1 .Maxmen A. How much is coronavirus spreading under the radar. *Nature* **10** (2020).

2 .Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases* **20** (2020) 669–677.

3 .Krantz SG, Rao ASS. Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. *Infection Control & Hospital Epidemiology* (2020) 1–3.

4 .[Dataset] Centro Nacional de Epidemiologia, Instituto de Salud Carlos III. Informe MoMo. situación a 30 de diciembre de 2020. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/MoMo/Paginas/Informes-MoMo-2020.aspx (2020).

5 .[Dataset] Ministerio de Sanidad Gobierno de España. Actualización nº 282. enfermedad por el coronavirus (COVID-19). https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documActualizacion_282_COVID19.pdf (2020).

6 .Moros MJS, Monge S, Rodríguez BS, San Miguel LG, Soria FS. COVID-19 in Spain: view from the eye of the storm. *The Lancet Public Health* (2020).

7 .Ojo O, García-Agundez A, Girault B, Hernández H, Cabana E, García-García A, et al. Coronasurveys: Using surveys with indirect reporting to estimate the incidence and evolution of epidemics.

KDD Workshop Humanitarian Mapping, San Diego, California USA, August 24, 2020. ArXiv preprint:2005.12783 (2020).

8 .Linares M, Garitano I, Santos L, Ramos JM. 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* (2020).

9 .Oliver N, Barber X, Roomp K, Roomp K. Assessing the impact of the COVID-19 pandemic in Spain: Large-scale, online, self-reported population survey. *Journal of medical Internet research* **22** (2020) e21319.

10 .[Dataset] Facebook Data for Good. COVID-19 symptom survey – request for data access. `https://dataforgood.fb.com/docs/covid-19-symptom-survey-request-for-data-access/` (2020). Accessed: 2021-01-24.

11 .Bernard HR, Hallett T, Iovita A, Johnsen EC, Lyerla R, McCarty C, et al. Counting hard-to-count populations: the network scale-up method for public health. *Sex. Transm. Infect.* **86** (2010) ii11–ii15.

12 .Pollán M, Pérez-Gómez B, Pastor-Barriuso R, Oteo J, Hernán MA, Pérez-Olmeda M, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *Lancet* **396** (2020) 535–544.

13 .LimeSurvey Project Team / Carsten Schmitz. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany (2012).

14 .Pallett SJ, Rayment M, Patel A, Fitzgerald-Smith SA, Denny SJ, Charani E, et al. Point-of-care serological assays for delayed SARS-CoV-2 case identification among health-care workers in the UK: a prospective multicentre cohort study. *Lancet Respir. Med.* **8** (2020) 885–894.

FIGURE CAPTIONS

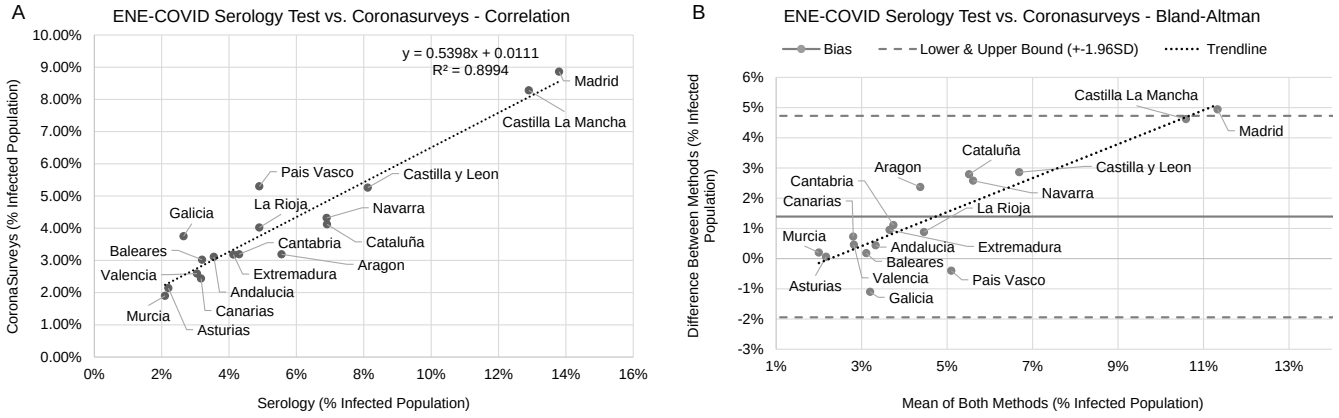

Figure 1. Comparison between the serology test and CoronaSurveys, Bland-Altman (**A**) and direct correlation (**B**)

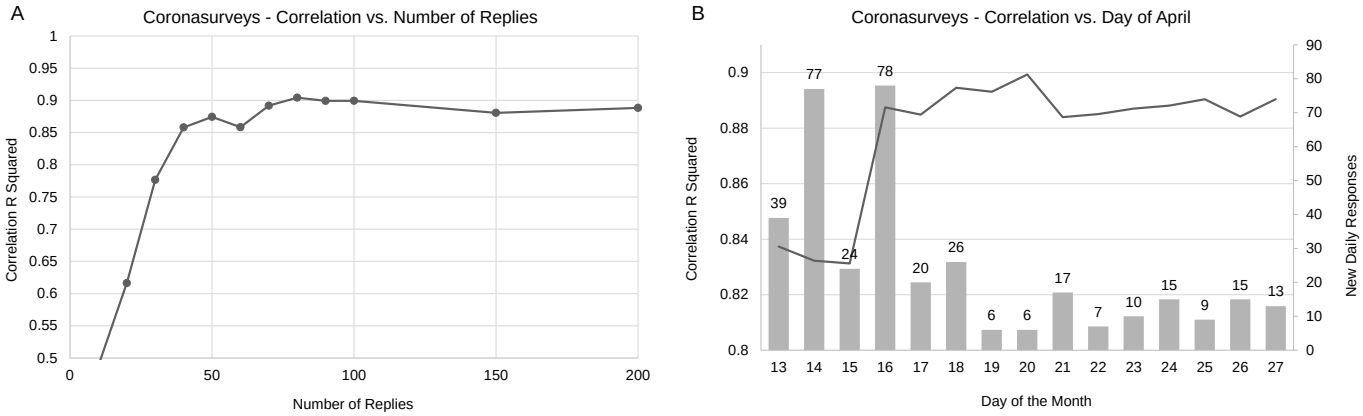

Figure 2. Convergence of correlation with number of replies (**A**) and day of the month (**B**)

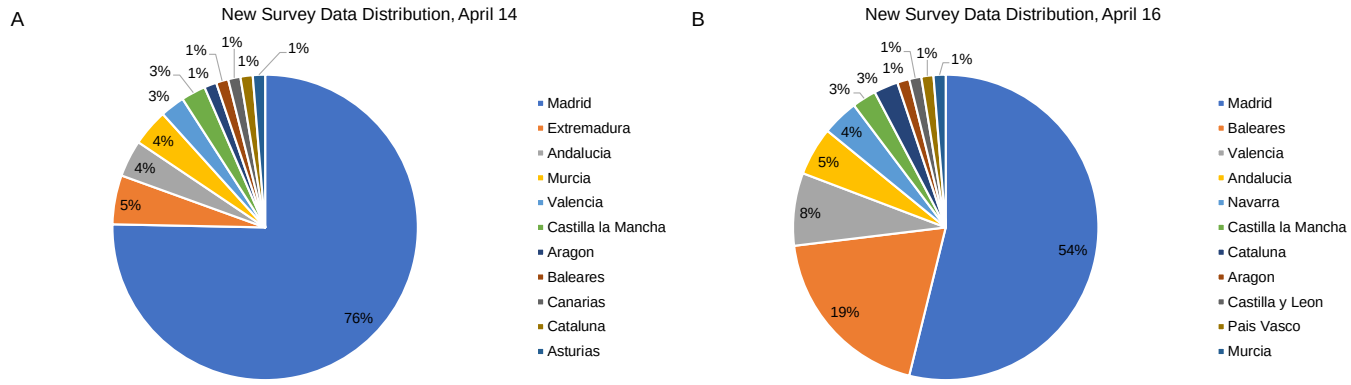

Figure 3. Distribution of new survey responses on April 14 (**A**) and April 16 (**B**)

Region	ENE-COVID (% Infected)	CoronaSurveys (% Infected)	CoronaSurveys Responses	CoronaSurveys Reach	Covid19Impact (% Infected)	Covid19Impact Responses
Andalucia	3.55	3.11(±0.41)	100	6,721	2.2(±0.3)	5,691
Aragon	5.56	3.19(±0.41)	44	3,045	2.0(±0.3)	1,463
Asturias	2.20	2.14(±0.52)	42	2,987	1.5(±0.3)	655
Cantabria	4.30	3.19(±0.96)	16	1,285	2.8(±0.3)	497
Castilla y Leon	8.12	5.26(±0.58)	86	5,763	3.7(±0.4)	1,994
Castilla La Mancha	12.90	8.28(±0.68)	100	6,399	8.0(±0.3)	3,469
Canarias	3.17	2.44(±0.74)	26	1,678	1.4(±0.2)	1,052
Catalonia	6.91	4.12(±0.49)	100	6,310	2.8(±0.3)	5,088
Extremadura	4.13	3.18(±0.74)	32	2,168	2.3(±0.4)	656
Galicia	2.65	3.75(±0.49)	85	5,781	1.3(±0.3)	2,257
Baleares	3.20	3.02(±0.76)	33	1,955	1.9(±0.3)	1,222
Murcia	2.10	1.90(±0.50)	45	2,835	1.5(±0.3)	3,566
Madrid	13.8	8.86(±0.67)	100	6,850	6.1(±0.4)	10,365
Navarra	6.90	4.32(±1.16)	16	1,180	3.6(±0.4)	580
Basque Country	4.90	5.30(±0.65)	65	4,511	1.9(±0.4)	1,007
La Rioja	4.90	4.02(±1.72)	9	498	1.8(±0.4)	220
Valencia	3.05	2.59(±0.37)	100	7,233	1.6(±0.3)	102,021

Table 1. Percentage (and 95% confidence interval) of infected population per region according to the ENE-COVID serology study (12), CoronaSurveys and Covid19Impact (9) (symptom-only model).