Irene Rocchetti, Dankmar Böhning, Heinz Holling and Antonello Maruotti*

Estimating the size of undetected cases of the COVID-19 outbreak in Europe: an upper bound estimator

https://doi.org/10.1515/em-2020-0024
Received July 21, 2020; accepted December 1, 2020; published online December 23, 2020

Abstract

Background: While the number of detected COVID-19 infections are widely available, an understanding of the extent of undetected cases is urgently needed for an effective tackling of the pandemic. The aim of this work is to estimate the true number of COVID-19 (detected and undetected) infections in several European countries. The question being asked is: How many cases have actually occurred?
Methods: We propose an upper bound estimator under cumulative data distributions, in an open population, based on a day-wise estimator that allows for heterogeneity. The estimator is data-driven and can be easily computed from the distributions of daily cases and deaths. Uncertainty surrounding the estimates is obtained using bootstrap methods.
Results: We focus on the ratio of the total estimated cases to the observed cases at April 17th. Differences arise at the country level, and we get estimates ranging from the 3.93 times of Norway to the 7.94 times of France. Accurate estimates are obtained, as bootstrap-based intervals are rather narrow.
Conclusions: Many parametric or semi-parametric models have been developed to estimate the population size from aggregated counts leading to an approximation of the missed population and/or to the estimate of the threshold under which the number of missed people cannot fall (i.e. a lower bound). Here, we provide a methodological contribution introducing an upper bound estimator and provide reliable estimates on the *dark number*, i.e. how many undetected cases are going around for several European countries, where the epidemic spreads differently.

Keywords: capture–recapture methods; COVID-19; geometric distribution; Chao's lower bound.

Introduction

The severe acute respiratory syndrome coronavirus 2 (COVID-19) has become a pandemic within few weeks. The number of detected cases increased day-by-day, at an exponential rate at the beginning, and now follows a logistic distribution (Petropoulos and Makridakis 2020; Sebastiani, Massa, and Riboli 2020). Cases of COVID-19 might have been vastly underreported in official statistics. It is widely acknowledged that the majority of the cases are asymptomatic and, thus, not observed or recorded (Böhning et al. 2020; Tuite et al. 2020; Yue, Clapham, and Cook 2020). In other words, the available data just tell us a part of the story: individuals may be

***Corresponding author: Antonello Maruotti**, Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne Libera Università Ss Maria Assunta, Rome, Italy; and Department of Mathematics, University of Bergen, Bergen, Norway, E-mail: a.maruotti@lumsa.it. https://orcid.org/0000-0001-8377-9950
Irene Rocchetti, Statistical Office - Consiglio Superiore della Magistratura, Rome, Italy
Dankmar Böhning, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK. https://orcid.org/0000-0003-0638-7106
Heinz Holling, Department of Methods and Statistics, Faculty of Psychology and Sports, University of Münster, Münster, Germany. https://orcid.org/0000-0002-0311-3970

already infected but are not aware of it, maybe because of the absence of symptoms, or cases may be under symptomatic suspicion but the disease has not been diagnosed yet (due to the delay in getting swab results).

The total number of cases is thus unknown, and general comments on the spread of the epidemic are thus partial as based on a (relatively small) fraction of the total cases. Some studies have used simulation-based approaches to infer reasonable estimates of total number of cases, but often these estimates are surrounded by poor uncertainty measures, leading to too wide confidence intervals (Flaxman et al. 2020). Many studies have also claimed that the number of undiagnosed cases is much higher than the official number (Li et al. 2020; Pollan et al. 2020; Phipps, Grafton, and Kompas 2020; Mukhopadhyay and Chakraborty 2020; Rothe et al. 2020; Yu et al. 2020). Here, we are proposing a simple and effective method to obtain reasonable point and interval estimates of the total number of COVID-19 infections in several European countries. In detail, we introduce a novel estimator based on a capture recapture (CR) approach. The capture–recapture method should be considered as the gold standard for counting when it is impossible to identify each case and large undercounts will occur (Lange and LaPorte 2003). CR methods were originally developed in the ecological setting with the aim of estimating the unknown size of a (possibly elusive) population and then they started to be applied also to epidemiological and health sectors (see Böhning, van der Heijden, and Bunge 2019; McRea and Morgan 2015). Many CR estimators have been proposed in the literature (see e.g. Tilling 2001; Wesson, Mirzazadeh, and McFarland 2018; Wesson et al. 2019), and some of them can be used to identify lower bounds (Chao and Colwell 2017) of the population size. In the analysis of COVID-19 infections, official data are available at the aggregated level, whereas individual data are not available to the general or the academic public. Hence, it is not possible to get the exact distribution of the number of infected individuals observed exactly one day, exactly two days and so on until m days. The population is open, subjected to deaths, and this may further complicate the analysis (McDonald and Amstrup 2001). A lower bound of the total number of infected cases is computed by Böhning et al. (2020) modifying the Chao estimator (Niwitpong et al. 2013) to address issues related to the data at hand. This is a relevant result as it provides reasonable information to the policy makers about the undetected cases and the magnitude this phenomenon may have at least, so that national health systems may be aware of the minimum number of cases that may demand health care services. At this stage of the spread of the epidemic, governments are willing to relax restrictive measures and several researches address issues related to the epidemic (Gregori et al. 2020; Khalatbari-Soltani et al. 2020; Lai 2020). To calibrate the new interventions, an estimate of the lower bound of the number of infections may not be enough, as COVID-19 has already shown to spread around the population very quickly (Li et al. 2020; Zhao et al. 2020; Zhou et al. 2020). This contribution aims at providing an approximated upper bound for the total number of COVID-19 cases, to better appreciate the dimension of the epidemic, under the worse scenario. Such an estimate is obtained from a non-parametric CR model, providing an upper bound estimate of the total number of infections regardless of the true data generating process.

This contribution is organized as follows. In “Methods” section, we introduce the basic notation and how we are going to work with the data at hand. A brief summary of the modified Chao lower bound is also discussed. These notions are then used to compute the upper bound, details of which are provided in “Data analysis” section, along with the computation of the uncertainty surrounding the estimates. In “Conclusions” section, we show the empirical application of the proposal on data from several European countries. A discussion showing other interesting insights concludes.

Methods

Preliminaries

Let us denote with $N(t)$ the cumulative count of infections at day t where t=t_0, …, t_m. Hence $\Delta N(t)$=$N(t)$ − $N(t$–1) are the number of new infections at day t where t=t_0 + 1, …, t_m. Also, let $D(t)$ denote the cumulative count of deaths at day t where t=t_0, …, t_m. t_0 defines the beginning of the observational period and t_m defines the end. We assume the trivial assumption t_m>t_0, so that the observational window is not empty. Again, we denote with $\Delta D(t)$=$D(t)$–$D(t$–1) the count of new deaths at day t where t=t_0 + 1, …, t_m.

The question arises how this can be linked to a capture–recapture approach. Let X_i denote the number of identifications for each infected individual i typically provided by the days the individual will surely remain infected. Let denote τ_x the probability of identifying an individual x times where x=0, … A lower bound estimator of the unobserved frequency f_0, say $\widehat{f}_0$, can be estimated by using the observed frequency of those identified exactly once, f_1, and of those identified twice, f_2 (Chao and Colwell 2017; Niwitpong et al. 2013):

$$\widehat{f}_0 = f_1^2 / f_2. \tag{1}$$

It is thus crucial to relate f_1 and f_2 with the data at hand. In detail, at each day t, $f_1(t)$ represents the infected people identified just once, i.e. the new infections, whose number is given by $\Delta N(t)$. Similarly, $f_2(t)$ represents the infected people detected at time $(t-1)$ and still infected at time t. This can be computed as $\Delta N(t-1) - \Delta D(t)$. Hence the estimate for the number of hidden infections at day t is

$$\widehat{f}_0(t) = \frac{[\Delta N(t)]^2}{\Delta N(t-1) - \Delta D(t)}. \tag{2}$$

By applying the estimator (1) day-wise we get the modified Chao lower bound estimator (see Böhning et al. 2020):

$$\widehat{f}_0 = \sum_{t=t_0+1}^{t_m} \frac{[\Delta N(t)]^2}{\Delta N(t-1) - \Delta D(t)}. \tag{3}$$

In practice, however, the bias-corrected form of (3) suggested by Chao (1989) is used:

$$\widehat{f}_0 = \sum_{t=t_0+1}^{t_m} \frac{\Delta N(t)[\Delta N(t) - 1]}{1 + \Delta N(t-1) - \Delta D(t)}. \tag{4}$$

We define the understanding that $\Delta N(t-1) - \Delta D(t)$ is set to 0 if it becomes negative, in other words we use $\max\{0, \Delta N(t-1) - \Delta D(t)\}$. The final estimate of lower bound (LB) of the total number of infection is then given as what has been observed at the end of the observational window t_m and the estimate of the hidden numbers:

$$N_{LB} = N(t_m) + \widehat{f}_0 \tag{5}$$

The upper bound estimator

The lower bound is helpful as an indication of the minimum number of people having had COVID-19 and answers to a fundamental open question: “How many undetected cases are at least going around?”. Nevertheless, this information may be treated as a starting point whenever interventions and tools to dampen the spread of the epidemic are rolled out. The proposed upper bound estimator extends the research on the undetected cases and helps policy makers to evaluate the COVID-19 epidemic situation locally and at the current phase of its development. An estimate of the worse possible scenario is provided.

Following a similar strategy as in “Preliminaries” section, this is achieved by firstly estimating daily-specific upper-bounds and then summing up all the estimates to get the final point-estimate of the maximum number of undetected cases. This daily-wise based upper bound approach provides an approximation of the data generation process.

Let us introduce the cumulative distribution function

$$\pi_{ij} = Pr(X_i \le j) = Pr(X_i = 0) + Pr(0 < X_i \le j) = \pi_{i0} + (1 - \pi_{i0})p_{ij}, \tag{6}$$

where homogeneity in the probability of being infected at a certain date t is assumed, i.e. $\pi_{ij}=\pi_j$, with $p_{ij}=p_j$ being the cumulative zero-truncated probability distribution. Equation (6) represents the probability that an individual is infected for at most j days, and it is function of π_0 and p_j; but π_0 is not observed. The quantities $p_j(j$=1, 2, 3) in Eq. (6) at each time t may be approximated as

$$\begin{aligned} p_1(t) &= f_1(t)/n^*_{\text{obs}}(t), \\ p_2(t) &= (f_1(t) + f_2(t))/n^*_{\text{obs}}(t), \\ p_3(t) &= (f_1(t) + f_2(t) + f_3(t))/n^*_{\text{obs}}(t) \end{aligned}$$

where $f_1(t)$ and $f_2(t)$ have been introduced in the previous section and

$$f_3(t) = \Delta N(t-2) - \Delta D(t-1) - \Delta D(t).$$

and $n^*_{\text{obs}}(t)$ is the number of current infected individuals observed at each time. We think that it is reasonable, for each day t, to consider the number of individuals affected by COVID-19 for the day t, for day t and the day before, and, for day t and the two days before, as m=3 is the minimum number of consecutive days of new infections necessary for the upper bound estimator to be computed. Furthermore, considering more than three days for an individual to be observed as affected by COVID-19 would lead to the risk of not observing the number of people affected by COVID-19 for exactly four, five and so on times because of the higher risk of overlapping cases.

Since π_0 is unknown, to compute the probabilities in (6), we substitute it with

$$\widehat{\pi}_0(t) = \frac{\widehat{f}_0(t)}{f_1(t) + f_2(t) + \widehat{f}_0(t)}.$$

where $\widehat{f}_0(t)$ is the *lower bound* probability of undetected cases derived from the Chao estimator in its bias corrected form, computed at each time t (see Eq. (2)). This also explains why a lot of detail was devoted to the lower bound estimator in the previous section as it is very much needed here. In other words, based on the Chao lower bound estimator of the undetected cases, we derive the *complete* count distribution and calculate the upper bound for the population size on such a complete distribution. Now, it follows that Eq. (6) takes the form

$$\widehat{\pi}_j(t) = \widehat{\pi}_0(t) + (1 - \widehat{\pi}_0(t))p_j(t)$$

when theoretical probabilities are replaced by their now available estimates. In order to provide an upper bound estimator we use the main results of Alfó, Böhning, and Rocchetti (2020):

$$\pi_j \leq p_j \left[1 - \left(1 - p_j\right) \left(\frac{p_{j+1} - p_j}{p_{j+1} - p_j \frac{\widehat{\pi}_j}{\widehat{\pi}_{j+1}}} \right) \right]^{-1}.$$

For j=m − 2, and by some algebra, we get the equivalent condition

$$\pi_0 \leq \frac{p_{m-1} - p_{m-2}}{\left(1 - \frac{\widehat{\pi}_{m-2}}{\widehat{\pi}_{m-1}}\right) + p_{m-1} - p_{m-2}} = \widehat{\pi}_0^{UB};$$

that makes clear why at least m=3 days should be considered. The right-hand side $\widehat{\pi}_0^{UB}$ of the above inequality provides an upper bound estimate of the population size based on the Horvitz–Thompson estimator:

$$\widehat{f}_0^{UB}(t) = n^{\star}_{\text{obs}}(t) \frac{\widehat{\pi}_0^{UB}}{1 - \widehat{\pi}_0^{UB}}.$$

However we deal with a day-wise upper bound approximation of π_0(t) which is given by

$$\widehat{\pi}_0^{UB}(t) = \frac{p_2(t) - p_1(t)}{\left(1 - \frac{\widehat{\pi}_1(t)}{\widehat{\pi}_2(t)}\right) + p_2(t) - p_1(t)}.$$

To get an estimate for the missed COVID-19 infections $\widehat{f}_0(t)$ at each time t we compute the Horvitz–Thompson (HT) estimator at each time t and ultimately we sum it up over all times, reaching thus the final upper bound for the missed COVID-19 cases n_0 as follows

$$\widehat{f}_0^{UB} = \sum_{t=t_0+2}^{t_m} \left(\frac{n^{\star}_{\text{obs}}(t)}{1 - \pi^{\star}_0(t)} - n^{\star}_{\text{obs}}(t) \right). \tag{7}$$

Hence, the approximated upper bound of the total number of infected people, $\widehat{N}_{UB}$, in the time window from t_0 to t_m is then given by

$$\widehat{N}_{UB} = \widehat{f}_0^{UB} + N_{t_m}.$$

Uncertainty estimation

A fundamental issue in general CR analyses is the quantification of uncertainty surrounding the estimates of the unknown population size. An estimation of the population size can be correctly computed, but if the associated estimation of variance is poor, then coverage by the 95% confidence interval may falsely indicate poor estimation by the point estimator, i.e. the point estimator may result in a poor coverage rate. Focusing on the proposed upper-bound estimator, we attempt here to investigate bootstrap methods as a robust and general approach to estimate variances and confidence intervals. Various bootstrap methods have been considered to estimate uncertainty in CR analyses with respect to other estimators (Anan, Böhning, and Maruotti 2017; Buckland and Garthwaite 1991; Norris and Pollock 1996; Zwane and van der Heijden 2003). In the following, we consider two different bootstrap approaches to approximate the uncertainty surrounding the point estimate: the imputed and the reduced bootstrap approaches.

Under the imputed bootstrap approach, we draw 1,000 bootstrapped samples of size N_{UB} generated according to a multinomial model whose probabilities are given by

$$\left\{\widehat{\pi}_0^{UB}(t) = \frac{\widehat{f}_0^{UB}(t)}{N_{UB}(t)}, \frac{f_1(t)}{N_{UB}(t)}, \frac{f_2(t)}{N_{UB}(t)}, \frac{f_3(t)}{N_{UB}(t)}\right\},$$

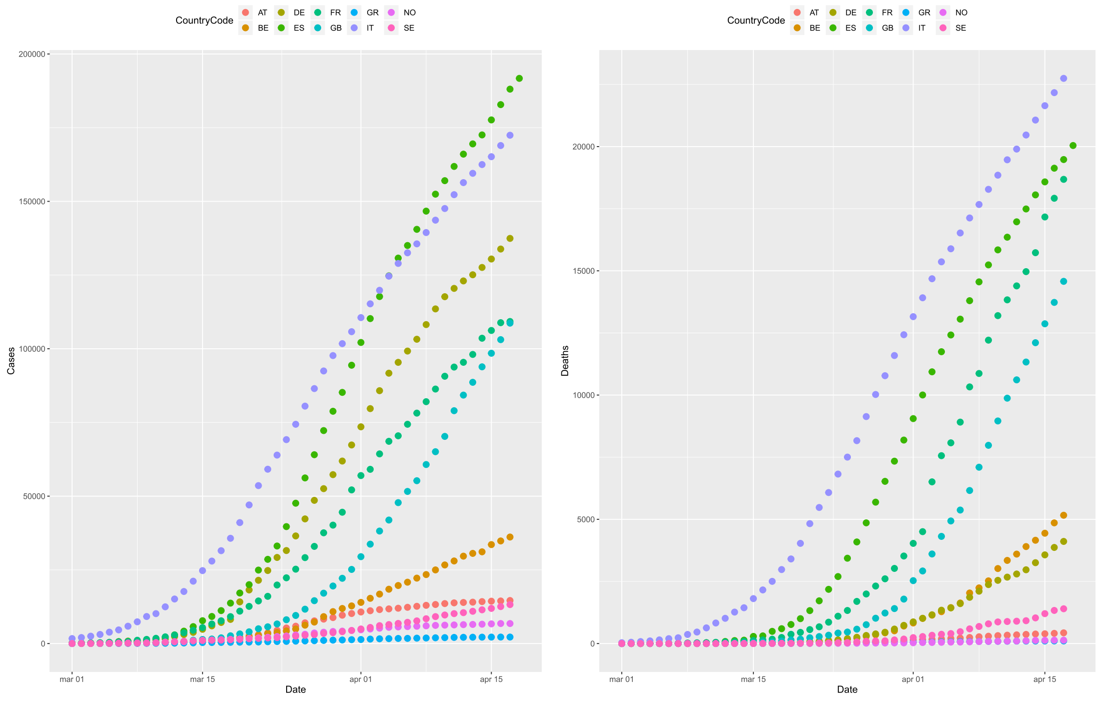
where $N_{UB}(t) = \widehat{f}_0^{UB}(t) + f_1(t) + f_2(t) + f_3(t)$.

Differently, under the reduced bootstrap approach, each of the bootstrapped samples contains $n^*_{\text{obs}}(t) = f_1(t) + f_2(t) + f_3(t)$ observations generated according to a multinomial model whose probabilities are given by $\left\{\frac{f_1(t)}{n^*_{\text{obs}}(t)}, \frac{f_2(t)}{n^*_{\text{obs}}(t)}, \frac{f_3(t)}{n^*_{\text{obs}}(t)}\right\}$. For each of the two approaches, the upper bound N_{UB} is computed for each bootstrapped sample, by summing up over the time period. Of course, for the imputed bootstrap the fraction of undetected cases is dropped and considered unknown when computing the population size. We report the 2.5 and 97.5% values of N_{UB} distribution. This allows us to overcome issues often encountered in the construction of the symmetric confidence intervals (Chao 1987): the sampling distribution could be skewed, the coverage probabilities may be unsatisfactory, etc.

Data analysis

The example provided here relies on European data. The time series of cumulative cases and deaths up to 17/04/2020 are considered and are taken from https://github.com/open-covid-19/data. A graphical representation of the data at hand is shown in Figure 1. The method is implemented in the **asymptor** package (Gruson 2020) of the R software.

Data from the day which we record the first death are analyzed only. We obtain the estimates of an upper bound for undetected cases for several European countries (see Table 1). The last column in Table 1 shows the ratio of the total estimated cases to the observed cases. The ratio of the total estimated cases (in the worse scenario) to the observed cases is interesting in itself. A ratio of 4.5 would mean that for every observed patient



(a) Cases

(b) Deaths

Figure 1: Cases and deaths for the analyzed countries.

Table 1: Estimated hidden and total cases of COVID-19 for several European countries, at 17/04/2020.

Country	Observed cases	Upper bound for total number of cases	(2.5–97.5%) Bootstrap values (IB)	(2.5–97.5%) Bootstrap values (RB)	Total/ observed	Total/observed (2.5–97.5%) IB	Total/observed (2.5–97.5%) RB
Italy	172,434	780,704	(777,690–784,121)	(778,080–783,895)	4.53	(4.51–4.58)	(4.51–4.55)
Austria	14,603	62,403	(61,631–63,465)	(61,549–63,474)	4.27	(4.22–4.37)	(4.21–4.35)
Germany	137,439	650,841	(647,138–655,236)	(646,974–655,056)	4.74	(4.71–4.77)	(4.71–4.77)
Spain	188,068	871,660	(868,136–875,570)	(868,615–874,953)	4.63	(4.61–4.66)	(4.62–4.65)
France	109,252	867,214	(814,767–944,686)	(811,082–952,137)	7.94	(7.46–8.65)	(7.42–8.72)
UK	108,692	504,652	(501,972–508,031)	(501,982–507,713)	4.64	(4.62–4.67)	(4.62–4.67)
Greece	2,207	9,586	(9,262–10,311)	(9,243–10,316)	4.34	(4.20–4.67)	(4.19–4.67)
Belgium	36,138	186,633	(182,715–191,609)	(182,744–191,383)	5.16	(5.06–5.30)	(5.06–5.30)
Norway	6,791	26,680	(26,199–27,456)	(26,197–28,344)	3.93	(3.86–4.04)	3.86–4.03
Sweden	13,216	56,917	(56,120–58,001)	(56,103–58,004)	4.31	(4.25–4.39)	(4.25–4.39)

there are 3.5 infected persons unseen. The reason for this can be manifold as these unseen cases might be without symptoms or show very mild signs of infection.

As expected, the undetected cases represent a relevant portion of the total number of cases. This is in line with a few existing works and discussions on the topic, see e.g. (Day 2020; La Stampa 2020; WHO 2020). The number of total number of cases are at most approximately 4.5 times the observed cases. Of course, differences arise at the country level, and heterogeneous estimates ranging from the 3.93 times of Norway to the 7.94 times of France, see Table 1. A recent study by the LMU Munich lead by Professor Hölscher shows that the amount of total infections (including those not identified) is four times higher than the officially register number of infections (http://www.klinikum.uni-muenchen.de/Abteilung-fuer-Infektions-und-Tropenmedizin/de/COVID-19/KoCo19/Aktuelles/index.html). These differences are due to different heterogeneity structures in the cases and deaths time series at the country level. These results are telling us that COVID-19 outbreak was more prevalent than described by the official data, though a significant number of individuals that are infected actually remain asymptomatic.

Point estimates can be used to synthetically describe the COVID-19 outbreak, but they may be rather uncertain. In Table 1, we also provide uncertainty measures, based on the bootstrap procedures described in Section "Uncertainity estimation". It is also possible to compare the two employed bootstrap approaches. They perform rather similarly (see also Anan, Böhning, and Maruotti 2017) and the bootstrap intervals are rather narrow, with France only showing a rather wide interval to indicate that its point estimate should taken with caution.

These results can be easily compared with those obtained in the works cited in the introduction. Results from Li et al. (2020), Pollan et al. (2020) and Yu et al. (2020) are in line with ours, and approximately 60–80% of the total cases are estimated as undetected. Mukhopadhyay and Chakraborty (2020) analyze data from India (not included in our sample), and estimates that the *total* number of infections is eight times the observed ones; this is a bit higher ratio than the one estimated for western countries. At last, substantial differences for some countries arise with the work of Phipps et al. (2020), though similar results are obtained for Austria and Norway; but the assumptions underlying their approach are different from ours.

Conclusions

Different capture–recapture approaches have been used to estimate the size of a partially observed population; many parametric or semi-parametric models have been developed to estimate the population size from aggregated counts leading to an approximation of the missed population and/or to the estimate of the threshold under which the number of missed people cannot fall (i.e. a lower bound). While several proposals for the latter exist, the estimation of an upper bound in capture recapture methods has been often overlooked, with the exception of the recent work of Alfó et al. (2020). We propose an extension of the upper bound estimator under cumulative data distributions, in an open population, such that a day-wise estimator varying over time. The approach results in a time-aggregated approximation for f_0 and thus for N. The proposed upper bound estimator has been applied to registered cases in some European countries; confidence intervals for N have been provided by employing bootstrap approaches. We consider, for each country, data up to the 17 of April, by assuming, given also the day wise nature of the estimator, that the recoveries are negligible. The main result is that, during the first wave, we are under-diagnosing the amount of infection in the population; and because of this, a large amount of the population moves normally, thus transmitting the disease. The hidden number is more important for transmission than the observed infections as these will be going in self-isolation. The larger the hidden amount, the more difficult is the control of the disease. Accordingly, we need more testing to have a better understanding of the amount of infection in the population. We should also test persons without symptoms, but it is no feasible to test everyone. Indeed, the economic and human resources required to continuously monitoring, at regular short intervals, the entire population may be not realistic. We have already seen difficulties in tracing and testing direct contacts of infected people in many European countries. A possible alternative is to apply sampling techniques, based on e.g. pooling or adaptive strategies. Comparison of molecular testing strategies for COVID-19 control are given, for example, in Grassly et al. (2020). To summarize existing results from the literature, the most effective approach to reduce onwards transmission is self-isolation of symptomatic individuals. Screening of high-risk groups (e.g. health-care workers) irrespective of symptoms by use of PCR testing may also play a fundamental role. The effectiveness of test and trace depends strongly on coverage and the timeliness of contact tracing. Antibody tests performance has been highly variable and may produce considerably uncertain results to judge for the effectiveness of this strategy. However when dealing with cases and deaths at a more recent date, given the increased percentage of immune people, recoveries should be taken into account in the computation. Another issue which should be considered is the one concerning the role of the deaths: even when the number of confirmed cases for two different countries are close to each other the upper bounds can be different according to the deaths size with respect to the cases (i.e. France and Spain). The length of the observation window plays an important role in this context and according to the distribution of COVID-19 cases observed more than once, the distribution can be less or more stable. It appears necessary to analyze this issue more deeply and we propose to do this in a future work.

Research funding: None declared.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.

References

Alfó, M., D. Böhning, and I. Rocchetti. 2020. "Upper Bound Estimators of the Population Size Based on Ordinal Models for Capture–Recapture Experiments." *Biometrics*, https://doi.org/10.1111/biom.13265.

Anan, O., D. Böhning, and A. Maruotti. 2017. "Uncertainty Estimation in Heterogeneous Capture–Recapture Count Data." *Journal of Statistical Computation and Simulation* 87: 2094–114.

Böhning, D., P. G. M. van der Heijden, and J. Bunge. 2019. *Capture–Recapture Methods for the Social and Medical Science*. Boca Raton: CRC Press.

Böhning, D., I. Rocchetti, A. Maruotti, and H. Holling. 2020. "Estimating the Undetected Infections in the COVID-19 Outbreak by Harnessing Capture–Recapture Methods." *International Journal of Infectious Diseases* 97: 197–201.

Buckland, S., and P. Garthwaite. 1991. "Quantifying Precision of Mark-Recapture Estimates Using the Bootstrap and Related Methods." *Biometrics* 47: 255–68.

Chao, A., and R. K. Colwell. 2017. "Thirty Years of Progeny from Chao's Inequality: Estimating and Comparing Richness with Incidence Data and Incomplete Sampling." *SORT Statistics and Operations Research Transactions* 41: 3–54.

Chao, A. 1987. "Estimating the Population Size for Capture–Recapture Data with Unequal Catchability." *Biometrics* 43: 783–91.

Chao, A. 1989. "Estimating Population Size for Sparse Data in Capture–Recapture Experiments." *Biometrics* 45: 427–38.

Day, M. 2020. "COVID-19: Identifying and Isolating Asymptomatic People Helped Eliminate Virus in Italian Village." *BMJ* 368: m1165.

Flaxman, S., S. Mishra, A. Gandy, H. Unwin, H. Coupland, T. Mellan, H. Zhu, T. Berah, J. Eaton, P. Perez Guzman, and N. Schmit. 2020. *Report 13: Estimating the Number of Infections and the Impact of Non-pharmaceutical Interventions on COVID-19 in 11 European Countries*. http://hdl.handle.net/10044/1/77731.

Grassly, N. C., M. Pons-Salort, E. P. Parker, P. J. White, N. M. Ferguson, K. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. Brazeau, and L. Cattarino. 2020. "Comparison of Molecular Testing Strategies for COVID-19 Control: A Mathematical Modelling Study." *The Lancet Infectious Diseases* 20: 1381–9.

Gregori, D., D. Azzolina, C. Lanera, I. Prosepe, N. Destro, G. Lorenzoni, and P. Berchialla. 2020. "A First Estimation of the Impact of Public Health Actions Against COVID-19 in Veneto (Italy)." *Journal of Epidemiology & Community Health* 74: 858–60.

Gruson, H. 2020. Asymptor: Estimate the Lower and Upper Bound of Asymptomatic Cases in an Epidemic Using the Capture/Recapture Methods. https://CRAN.R-project.org/package=asymptor.

Khalatbari-Soltani, S., R. G. Cumming, C. Delpierre, and M. Kelly-Irving. 2020. "Importance of Collecting Data on Socioeconomic Determinants from the Early Stage of the COVID-19 Outbreak Onwards." *Journal of Epidemiology & Community Health* 74: 620–3.

La Stampa. 2020. Castiglione d–Adda – un caso di studio: –Il 70% dei donatori di sangue – positivo–. lastampa.it 2020. https://www.lastampa.it/topnews/primo-piano/2020/04/02/news/coronavirus-castiglione-d-adda-e-un-caso-di-studio-il-70-dei-donatori-di-sangue-e-positivo-1.38666481.

Lai, F. T. T. 2020. "Association Between Time from SARS-CoV-2 Onset to Case Confirmation and Time to Recovery across Socio-Demographic Strata in Singapore." *Journal of Epidemiology & Community Health* 74: 678.

Lange, J. H., and R. E. LaPorte. 2003. "Capture–Recapture Method Should Be Used to Count How Many Cases of SARS Really Exist." *BMJ* 326: 1396.

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. 2020. "Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (SARS-CoV2)." *Science* 368: 489–93.

McDonald, T. L., and S. C. Amstrup. 2001. "Estimation of Animal Abundance and Related Parameters." *Journal of Agricultural, Biological, and Environmental Statistics* 6: 206–20.

McRea, R. S., and B. J. T. Morgan. 2015. *Analysis of Capture–Recapture Data*. Boca Raton: CRC Press.

Mukhopadhyay, S., and D. Chakraborty. 2020. "Estimation of Undetected COVID-19 Infections in India." *medRxiv*, https://doi.org/10.1101/2020.04.20.20072892.

Niwitpong, S. A., D. Boehning, P. G. van der Heijden, and H. Holling. 2013. "Capture–Recapture Estimation Based upon the Geometric Distribution Allowing for Heterogeneity." *Metrika* 76: 495–519.

Norris, J. L., and K. H. Pollock. 1996. "Including Model Uncertainty in Estimating Variances in Multiple Capture Studies." *Environmental and Ecological Statistics* 3: 235–44.

Petropoulos, F., and S. Makridakis. 2020. "Forecasting the Novel Coronavirus COVID-19." *PLoS ONE* 15 (3): e0231236.

Phipps, S.J., R. Q. Grafton, and T. Kompas. 2020. Estimating the True (Population) Infection Rate for COVID-19: A Backcasting Approach with Monte Carlo Methods. *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.05.12.20098889v1.

Pollan, M., B. Perez-Gomez, R. Pastor-Barriuso, J. Oteo, M. A. Hernán, M. Pérez-Olmeda, J. L. Sanmartín, A. Fernández-García, I. Cruz, N. F. de Larrea, and M. Molina. 2020. "Prevalence of SARS-CoV-2 in Spain (ENE-COVID): A Nationwide, Population-Based Seroepidemiological Study." *The Lancet* 396: 535–44.

Rothe, C., M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, T. Zimmer, V. Thiel, C. Janke, W. Guggemos, and M. Seilmaier. 2020. "Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany." *New England Journal of Medicine* 382: 970–1.

Sebastiani, G., M. Massa, and E. Riboli. 2020. "Covid-19 Epidemic in Italy: Evolution, Projections and Impact of Government Measures." *European Journal of Epidemiology* 35: 341–5.

Tilling, K. 2001. "Capture–Recapture Methods – Useful or Misleading?" *International Journal of Epidemiology* 30: 12–14.

Tuite, A. R., V. Ng, E. Rees, and D. Fisman. 2020. "Estimation of COVID-19 Outbreak Size in Italy." *The Lancet Infectious Diseases* 20: 537.

Wesson, P. D., A. Mirzazadeh, and W. McFarland. 2018. "A Bayesian Approach to Synthesize Estimates of the Size of Hidden Populations: The Anchored Multiplier." *International Journal of Epidemiology* 47: 1636–44.

Wesson, P. D., W. McFarland, C. C. Qin, and A. Mirzazadeh. 2019. "Software Application Profile: The Anchored Multiplier Calculator–A Bayesian Tool to Synthesize Population Size Estimates." *International Journal of Epidemiology* 48: 1744–9.

WHO. 2020. Q&A: Similarities and Differences – COVID-19 and Influenza. https://www.who.int/news-room/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza.

Yu, Y., Y. R. Liu, F. M. Luo, W. W. Tu, D. C. Zhan, G. Yu, and Z. H. Zhou. 2020. "COVID-19 Asymptomatic Infection Estimation." *medRxiv*, https://doi.org/10.1101/2020.04.19.20068072.

Yue, M., H. E. Clapham, and A. R. Cook. 2020. “Estimating the Size of a COVID-19 Epidemic from Surveillance Systems.” *Epidemiology* 31: 567–9.

Zhao, S., Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, and M. H. Wang. 2020. “Preliminary Estimation of the Basic Reproduction Number of Novel Coronavirus (2019-nCoV) in China, from 2019 to 2020: A Data-Driven Analysis in the Early Phase of the Outbreak.” *International Journal of Infectious Diseases* 92: 214–17.

Zhou, T., Q. Liu, Z. Yang, J. Liao, K. Yang, W. Bai, X. Lu, and W. Zhang. 2020. “Preliminary Prediction of the Basic Reproduction Number of the Wuhan Novel Coronavirus 2019-nCoV.” *Journal of Evidence-Based Medicine* 13: 3–7.

Zwane, E., and P. van der Heijden. 2003. “Implementing the Parametric Bootstrap in Capture–Recapture Studies.” *Statistics & Probability Letters* 65: 121–5.