Peer-reviewed version available at *Public Health in Practice* **2020**; doi:10.1016/j.puhip.2020.100013

On the Importance of Population-Based Serological Surveys of SARS-CoV-2 Without Overlooking Their Inherent Uncertainties

Evangelos I. Kritsotakis, PhD, FHEA, CStat

Associate Professor of Biostatistics

School of Medicine, University of Crete,

71003, Heraklion, Crete, Greece

e.kritsotakis@uoc.gr

ORCID: 0000-0002-9526-3852

May 3, 2020

Abstract

This brief note aims to explain the scope in conducting large-scale serological surveys of SARS-CoV-2 to define the landscape of population immunity without overlooking the inherent uncertainty steaming from sampling design and diagnostic validity. The note completes with a succinct statistical appendix of simple methods for estimating prevalence from random population samples using imperfect diagnostic tests.

Keywords

Serosurvey, seroepidemiology, seroprevalence, sampling, imperfect diagnostic test, sensitivity, specificity, Coronavirus.

The problem

To date we know little about the SARS-CoV-2 virus spread into the general population. Our great uncertainty stems from the fact that the virus spreads easily between people but many COVID-19 infections are extremely mild, subclinical or asymptomatic and therefore go unnoticed. The actual number of people already exposed to SARS-CoV-2 may be much higher than the number of confirmed COVID-19 patients who have been seriously ill and/or tested positive for SARS-CoV-2. Most experts would agree that it is reasonable to assume that we are at least 10 times off in reported numbers, but a recent report suggests that the actual number of infections may be as much as 85 times higher than that reported [1].

From a public health standpoint, knowing how many and who have already been exposed to SARS-CoV-2 gives a clearer picture of how widespread the virus is in local populations. This is extremely useful because public health measures depend on how far Coronavirus has already penetrated into the general population. In the absence of precise estimates from a random sample of the general population, we are essentially operating in the dark and likely to continue taking restrictive measures without being able to assess their effectiveness.

Seroprevalence surveys

Population-based serological surveys, commonly referred to as seroprevalence studies or serosurveys, can generate much needed data [2]. They use serological tests to examine a large number of blood samples from people without a confirmed SARS-CoV-2 infection to detect signs that they were once infected with the virus. That is, serological tests detect our body's response to the virus but not the virus itself (as opposed to molecular tests). Therefore, they cannot be used early in infection before the patient's body has already developed an antibody response. Thus, serological tests are not much helpful for clinicians to diagnose infection in individual persons. However, they are extremely useful for epidemiological purposes to understand the immunity landscape of the population at large.

Estimating the true rate of SARS-CoV-2 infection allows epidemiologists to predict the likely future course of the epidemic in specific locations or populations and helps public health authorities to better design interventions to control the epidemic. This is because we expect, although no one is entirely certain yet, that once we have antibodies to the virus, they will provide us with immunity, that is, we will be protected for some period of time. Detecting people who are potentially immune to SARS-CoV-2 could even play an important role in when and how social distancing restrictions are lifted. The results of serological surveys can also be useful in guiding strategic decisions on essential staffing in hospitals and other health care facilities - for example, by assigning to the forefront those who are probably immune. It is therefore desirable to conduct targeted serological studies of healthcare workers.

Inherent uncertainties

The results of serological surveys come with uncertainty, but it is important to note that this can be assessed. Uncertainty stems from two main sources: (a) sampling variability, that is, from the fact that we examine only a small part of the overall population, and (b) diagnostic validity, that is, imperfect accuracy of the immunoassay test in detecting the presence or the absence of antibodies. Therefore, it is critical that serological surveys are based on both appropriate sampling designs assuring population representation and accurate serological tests. Due to urgency and demand, several serological tests have been developed and placed on the market recently. Manufacturer's own data [3] and independent evaluations [4] indicate that accurate enough tests are currently available: their probability of successfully detecting people exposed to SARS-CoV-2 (sensitivity) exceeds 90% a few days after the infection and their success in detecting non-infected individuals (specificity) reaches 99%.

An example

Available serological tests are not perfect but are acceptable for use in the context of surveying populations for SARS-CoV-2 antibodies, because survey estimates can be corrected for imperfect diagnostic performance. For example, let us assume that a serological survey of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2020 doi:10.20944/preprints202005.0194.v1

Peer-reviewed version available at Public Health in Practice 2020; doi:10.1016/j.puhip.2020.100013

$n = 1{,}000$ people found that $a = 100$ are positive for SARS-CoV-2 antibodies, meaning that $P_A = a/n = 10\%$ were infected. The test used was imperfect, say with known sensitivity $S_e = 92\%$ and specificity $S_p = 98\%$, but we can correct our estimate for these inaccuracies. The corrected estimate of the true prevalence of SARS-CoV-2 turns out to be $P_T = 8.9\%$. We can express the uncertainty associated with this estimate using a 95% confidence interval, which in this case is from 6.7% to 11.1%. In this way, we get a fairly precise idea of the extent of the virus spread into the population.

Conclusion

Large-scale seroprevalence surveys are an important tool in combating COVID-19 disease as they can provide much-needed estimates of the fraction of the population with antibodies against SARS-CoV-2. The quality of the antibody prevalence estimates depends on the sampling design and the diagnostic accuracy of serological tests.

Statistical Appendix

This appendix provides a summary of simple methods to estimate prevalence using imperfect diagnostic tests.

Assume that the prevalence of infection (π_T) in the target population is a fixed, but unknown quantity. To estimate π_T, we do a diagnostic test on n randomly sampled individuals from the target population and a individuals test positive. However, the test is imperfect, with sensitivity (S_e) and/or specificity (S_p) that are below 100%. Thereby, the apparent prevalence $P_A = a/n$ is a biased estimate of π_T.

Let P_T denote the true prevalence proportion that we would observe if the diagnostic test was perfect. It is easy to confirm that the apparent prevalence $P_A = a/n$ and the true prevalence P_T are related by:

$$P_T = \frac{P_A + S_p - 1}{S_e + S_p - 1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2020 doi:10.20944/preprints202005.0194.v1

Peer-reviewed version available at Public Health in Practice 2020; doi:10.1016/j.puhip.2020.100013

P_T is known as the Rogan–Gladen-estimator. [5] Assuming S_e and S_p are known with certainty, P_T is an unbiased estimate of the true population prevalence π_T. It is also a maximum likelihood estimate of π_T .[6] Note that P_T is meaningful under the reasonable requirement that the diagnostic test is better than the flip of a coin $(S_e + S_p > 1)$. Nevertheless, P_T is not guaranteed to lie between 0 and 1 (especially when P_A is very small) and a "clipped" estimate may need to be used: $P_{TC} = min[max(P_T, 0), 1]$.

The standard error of P_T is:

$$SE(P_T) = \frac{SE(P_A)}{S_E + S_p - 1}$$

where $SE(P_A)$ depends on the sampling design used. For a simple random sample from a large population:

$$SE(P_A) = \sqrt{\frac{P_A(1 - P_A)}{n}}$$

For large *n*, the statistic $(P_T - \pi)/SE(P_T)$ can be treated as a standard normal variate. Thus, an approximate 95% confidence interval for π is obtained as:

$$P_T \mp 1.96 \cdot SE(P_T)$$

The "clipped" estimate $P_{TC} = min[max(P_T, 0), 1]$ is asymptotically equivalent to P_T ,[7] so the large sample theory is valid in that case too.

Essentially, for fixed S_e and S_p, a 95% confidence interval $[l, u]$ for the apparent prevalence π_A, can be converted to a 95% confidence interval for the true prevalence π_T by

$$\left[\frac{l + S_p - 1}{S_e + S_p - 1}, \frac{u + S_p - 1}{S_e + S_p - 1}\right]$$

Consequently in situations where asymptotic assumptions are not met (e.g. small sample size and/or very low prevalence), exact methods (e.g. Clopper-Pearson) can be applied to calculate confidence limits for the apparent prevalence that can be converted to confidence limits for the true prevalence using the formula above. [8]

Peer-reviewed version available at Public Health in Practice 2020; doi:10.1016/j.puhip.2020.100013

If S_e and S_p are not known with certainty, but independent binomial estimates are available from a validation study on persons whose infection status is known, then P_T is biased but to a much lesser degree than P_A [5]. In that case, a more valid quantification of standard error that captures the uncertainty in S_e and S_p is given by:

$$SE(P_T) = \frac{1}{S_e + S_p - 1}\sqrt{\left[SE(P_A) + \frac{S_e(1-S_e)}{n_1}P_T^2 + \frac{S_p(1-S_p)}{n_2}(1-P_T)^2\right]}$$

where n_1 and n_2 denote the numbers of infected and non-infected individuals in the validation study. [5] A double sampling design that partly utilises a more definitive diagnostic test can also be used [9]. Using a binomial distribution model for the number of positive tests a out of the n individuals tested, a Bayesian approach may also be used to estimate π_T that does not yield explicit formulae but is computationally easy [10,11].

References:

1. Bendavid, E. *et al.* COVID-19 Antibody Seroprevalence in Santa Clara County, California. *medRxiv* 2020.04.14.20062463 (2020). doi:10.1101/2020.04.14.20062463
2. Metcalf, C. J. E. *et al.* Use of serological surveys to generate key insights into the changing global landscape of infectious disease. *Lancet (London, England)* **388**, 728–30 (2016).
3. Cairns, E. Covid-19 antibody tests face a very specific problem. *Evaluate Vantage* (2020).
4. Kontou, P. I., Braliou, G. G., Dimou, N. L., Nikolopoulos, G. & Bagos, P. G. Antibody tests in detecting SARS-CoV-2 infection: a meta-analysis. *medRxiv* 2020.04.22.20074914 (2020). doi:10.1101/2020.04.22.20074914
5. Rogan, W. J. & Gladen, B. Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.* **107**, 71–76 (1978).
6. Levy, P. S. & Kass, E. H. A three-population model for sequential screening for bacteriuria2. *Am. J. Epidemiol.* **91**, 148–154 (1970).
7. Gastwirth, J. L. The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test Data. *Stat. Sci.* **2**, 213–222 (1987).

8. Reiczigel, J., Földi, J. & Ózsvári, L. Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. *Epidemiol. Infect.* **138**, 1674–1678 (2010).

9. Levy, P. S. & Lemeshow, S. *Sampling of Populations: Methods and Applications: Fourth Edition*. *Sampling of Populations: Methods and Applications: Fourth Edition* (John Wiley & Sons, Inc., 2011). doi:10.1002/9780470374597

10. Diggle, P. J. Estimating Prevalence Using an Imperfect Test. *Epidemiol. Res. Int.* **2011**, 1–5 (2011).

11. Lewis, F. I. & Torgerson, P. R. A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. *Emerg. Themes Epidemiol.* **9**, 9 (2012).