Snowball Sampling Study Design for Serosurveys in the Early COVID-19 Pandemic

Snowball sampling study design for serosurveys in the early COVID-19 pandemic

William P. Hanage[1*], Xueting Qiu[1] and Lee Kennedy-Shaffer[1]

1. Center for Communicable Disease Dynamics and Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA

*Corresponding author
William P. Hanage whanage@hsph.harvard.edu

ABSTRACT

Serological surveys can provide evidence of cases that were not previously detected, depict the spectrum of disease severity and estimate the proportion of asymptomatic infection. To capture these parameters, survey sample sizes may need to be very large, especially when the overall infection rate is still low. Therefore, we describe a novel method of "snowball sampling" to enrich serological surveys by using contact networks identified in the early SARS-CoV-2 pandemic. By testing all contacts of known index cases, snowball sampling efficiently builds a sample to answer many key questions about a new outbreak, such as estimating asymptomatic proportion of all infected cases, the probability of a given clinical presentation for a seropositive individual, or the association between characteristics of either the host or the infection and seropositivity among contacts of index individuals. Although clustering effects need to be considered since identified cases have common exposures, snowball sampling can be a more efficient way to achieve adequate statistical power than random sampling, as demonstrated in the COVID-19 example. We hope such study designs can be applied to provide valuable information to slow the onward spread of the pandemic as it enters its next stage.

KEY WORDS

asymptomatic infection, design effect, contact tracing, COVID-19, SARS-CoV-2, serosurvey, study design, transmission chain

There is great interest in the results of serosurveys based on antibodies against the SARS-CoV-2 virus, to indicate the true numbers of people infected so far in the pandemic, the proportion that might be immune in future waves of infection, and the proportion of people infected who experience mild or no symptoms. We are still in the early stages of the pandemic of COVID-19 caused by SARS-CoV-2 and these and many other important parameters for the public health response remain obscure. In addition, while asymptomatic or pre-symptomatic cases are known to be capable of transmission, the proportion of infections caused by such cases is not clear. We also do not know whether infections that cause a larger than average number of secondary cases have any distinctive characteristics in common. This phenomenon gives rise to an overdispersed R0, which has been observed in other beta-coronaviruses and is likely to be the case for SARS-CoV-2.[1] Identifying the properties of such 'superspreader events' is a key element of control. Here we propose a survey design with the potential to rapidly gather data capable of addressing these questions.

The majority of serosurvey studies seek to define the amount of population immunity, and how it might impact the future progress of the pandemic. There is also interest in determining the amount of infections that might have been undetected because they were either minimally symptomatic or exhibited symptoms that did not lead to testing. While there is a focus on random testing, the rates of population immunity in different places will vary greatly depending on the stage of the pandemic, and a very large sample may be necessary to include enough cases to capture less common disease presentations. An alternative study design based on 'snowball sampling' offers a route to collect data on the spectrum of clinical severity and impact on transmissibility. This sampling approach is familiar to researchers attempting to enrich their

sample with people who are otherwise hard to reach.[2] While enriched sampling might typically focus on a marginalized or underrepresented community, in this case we are enriching for the presence of seropositivity. This approach is best used in research seeking the full set of transmission events, including individuals infected with SARS-CoV-2 who were not previously identified, either because they did not receive a test or because the test result was a false negative.

The method assumes that a serological assay with high specificity and sensitivity for previous SARS-CoV-2 infection is available. The approach is based on contact tracing. One use of serology in contact tracing for SARS-CoV-2 was demonstrated recently as it identified a previously-unrecognized infection that connected two otherwise unlinked transmission chains.[3] This approach instead uses contact tracing to inform serological sampling. By taking people who are known to have been infected and tracing forward in time in order to identify who they infected, we are able to both identify how many secondary cases were infected by an index case and obtain a larger dataset of individuals who have been infected. This latter consequence allows us to better determine the range of clinical presentations of infection, even when current population disease prevalence remains low.

The goal of snowball sampling is not to determine the amount of population level immunity, which is best addressed by conventional serosurveys, but to obtain many individuals who have been exposed, in order to estimate the range of clinical presentations and their relation to transmission. We begin with known cases, occurring at a point in the past and with known clinical courses. These should test positive and provide an estimate of the distribution of

antibody titers resulting from infection. We then proceed to test the reported contacts of these cases, both those to whom the primary case is known to have transmitted and other potential contacts who may not have been previously identified as infected. Given limited contact tracing early in the pandemic we expect to identify both previously known transmission events and others that escaped initial surveillance, for instance because of a false negative on initial test. For each case identified, we collect a full history of symptoms, and their contacts are identified and tested for serological evidence of prior SARS-CoV-2 infection. We continue to build out the sample in this way until it has sufficient statistical power to answer the question(s) of interest.

STATISTICAL ANALYSIS AND POTENTIAL APPLICATIONS

There are several types of scientific questions of interest that may be answered through this design. For example:

1. Identify the possible clinical presentations for an individual with a positive serology test (seropositive individual).
2. Identify the probability of a given clinical presentation for a seropositive individual.
3. Identify the association between some characteristics of interest (e.g., types of contact or personal characteristics) and seropositivity among contacts of index individuals.
4. Identify the association between some characteristics of interest (e.g., clinical severity or number of contacts) and the number of secondary cases due to one infection.

Accordingly, these require different analysis approaches:

1. Assess all clinical presentations of all sampled seropositive individuals.

2. Identify the probability of each clinical presentation among sampled seropositive individuals.
3. Compute the risk difference, risk ratio, or odds ratio for seropositivity by the characteristics of interest.
4. Compute a measure of association for the number of secondary cases by the characteristics of interest.

For questions 2, 3, and potentially 4, estimation and inference must account for the clustered nature of the data, as the contacts of an individual have a shared exposure (and perhaps other latent shared characteristics) and thus can be viewed as a cluster.[4,5] In these cases, we assume that the index individuals we select are representative of all possible index individuals from some larger population of interest (e.g., all individuals with confirmed infection in a given time range at a given geographic location, workplace, etc.) This allows us to view the sample as a cluster sample from a larger population of clusters, allowing inference to proceed.[6]

Question 2 can then be viewed as ratio estimation from a one-stage cluster sampling survey, and estimation and inference can proceed accordingly (noting the different cluster sizes).[6] Questions 3 and 4 can be viewed as regression questions, and estimation and inference can proceed according to various methods, including mixed effects models or robust variance estimation.[4,5]

SAMPLE SIZE AND POWER CALCULATIONS

Making the assumptions described in the Statistical Analysis section, we can calculate the required sample size according to the analysis method. Because of the clustering, we must inflate

the variance (and thus the sample size required) by an appropriate design effect. This design effect can be estimated by $DE = 1 + (\overline{m} - 1)\rho$, where $\overline{m}$ is the average number of seropositive contacts per index individual and ρ is the intracluster correlation coefficient (ICC) for the outcome of interest.[5,6] When the number of seropositive contacts per index individual varies substantially, better (and more conservative) estimates of the design effect can be obtained by replacing $\overline{m}$ in this formula by either $\overline{m}_h$, the harmonic mean of the number of seropositive contacts per index individual, or $\overline{m}\times(1 + CV_m^2)$, where CV_m is the coefficient of variation (the standard deviation divided by the mean) of the number of seropositive contacts per index individual. Note that for questions 1 and 2, the outcome of interest is not seropositivity, but rather the specific clinical presentation. This likely has a lower ICC than seropositivity itself.

To compare the required sample size for a question of interest to the sample size required from a simple random sample of individuals, however, we account for the inflated variance through the design effect but we also must account for the higher percentage of tested individuals who are seropositive due to the enriched sample from this design. That is, if we perform the serology tests on a fixed sample size of individuals, *N*, then our effective sample size is *Np* for a simple random sample of individuals, where *p* is the overall percentage of the population that is seropositive. In the snowball sampling design, however, the effective sample size is Nq/DE, where q is the marginal probability of a close contact of an index individual testing positive, and *DE* is the design effect defined above. If $q/DE \geq p$, then the snowball sampling will be at least as efficient as the simple random sampling.

Example. Assume that 1) we identify 20 index individuals, who are independent of one another; 2) the infection has mean $R_0 = 2$ and 3) the population seropositive rate is 5% (i.e., $p = 0.05$), as was estimated by seroprevalence surveys conducted in April and May 2020 in Los Angeles County and Spain.[7,8] Other European studies have generally also shown seroprevalence rates of 5–10% following the outbreaks in the spring.[9] Now suppose that each index individual has 10 close contacts, each with an equal probability of being infected by the index individual. To get $R_0 = 2$, the probability of infection for each close contact is 20%, so $q \geq 0.20$. Note that the probability is greater than or equal to 0.20 because any individual not infected by the index individual still has a chance of becoming infected later in the outbreak. Testing 200 individuals, under random sampling, will yield an expected 10 seropositive individuals. Testing 200 individuals, under snowball sampling, will yield an expected 40 seropositive individuals (not including the index cases). In the absence of superspreading events, the mean cluster size (excluding clusters of size 0) is 2.24 with a variance of 1.25, giving $\overline{m}\times(1 + CV_m^2) = 2.8$. So for any value of ρ, $q/DE \geq p$ and the snowball sampling design is more efficient than random sampling. For $\rho = 0.05$ used in the design of the Ebola ring vaccination trial,[4] the effective sample size of the snowball sampling design is more than three times that of the random sampling design. If the snowball design additionally tests the close contacts of all seropositive individuals found among the first ring of close contacts, then an additional 400 tests will yield 80 more seropositive individuals, provided the transmission events occurred sufficiently prior to testing for seroconversion to occur. That is, 600 tests yield 120 seropositive individuals total, whereas randomly sampling 600 individuals to test would be expected to yield 30 seropositive individuals. Treating all individuals identified in the close contact chain of an initial index case as a cluster, this method will increase mean cluster size and thus the design effect, leading to a

lower relative benefit of the snowball sampling design (although the ICC will likely decrease for this larger cluster). But it may be a more feasible method of getting a larger snowball sample than identifying more index cases.

In the absence of serological surveys, the relative effective sample sizes of the methods would need to be estimated by the timing of the survey in the course of the outbreak. Using the generation interval, reproduction number, population size, and the number of introductions to a population, the cumulative incidence can be estimated at any point in the outbreak.[10,11] This estimate can be used as p for a serological survey conducted once those infections have seroconverted, which in the case of SARS-CoV-2 infection is estimated to occur within three weeks after symptom onset.[12]

DISCUSSION

This study design has a number of advantages since its contact-based testing method enriches the sample for cases of infection. This allows us to more rapidly and efficiently determine the range of clinical presentations, including within the community rather than those that have had contact with healthcare. In a sample of sufficient size, we would also be able to compare the numbers of onward infections associated with different clinical presentations. More data on the role of asymptomatic and less severe clinical presentations is critical. A recent study of an outbreak in a Korean call center successfully identified cases of asymptomatic infection but concluded that they did not transmit,[13] but this could have been influenced by changing contact patterns in response to the outbreak. There is also uncertainty about the antibody response among people with less severe clinical presentations, which is essential for long-term predictions of population

immunity. Data on this is vital, and snowball sampling offers a way of obtaining it comparatively rapidly and efficiently. This sampling approach could also inform estimates of the secondary attack rate of symptomatic and asymptomatic cases, improving future modeling studies and providing context for tailored public health interventions.

In addition to being unable to estimate overall population seroprevalence, this approach faces potential limitations for the analyses for which it is designed. The labor in contact tracing is not trivial and may increase the cost of the survey. The relative effective sample size of the design depends on many factors. In particular, if there are transmissions outside of the identified close contacts, it decreases the relative benefit of the snowball design (in the example, if half of the transmission from index individuals is to non-identified contacts, q is reduced by half). A more mature epidemic, where the infection has been circulating in the community for some time, will also increase the probability of infection of a randomly sampled individual compared to an identified close contact. And a high proportion of super-spreading events, where a few index individuals are responsible for many secondary cases while many index individuals lead to no secondary cases, will increase the mean cluster size and the variation of the cluster sizes, leading to a higher design effect and lower benefit of snowball sampling.[14] Interventions that reduce R_t will also reduce the number of seropositive individuals identified among close contacts. However, if the intervention also reduces the number of identified close contacts (e.g., social distancing), then it may preserve the relative benefit of snowball sampling. From the practical side, if there is a cost of identifying and reaching index individuals and their close contacts above that of identifying and reaching randomly sampled individuals, the benefit of snowball sampling will decrease for index individuals whose close contacts overlap more.

Nevertheless, the snowball sampling survey design can collect samples in a more rapid and efficient manner than conventional serosurveys, especially in the early stage of an epidemic. Studies using this design can then provide vital information on important parameters, including the range and likelihood of clinical disease severity among infected individuals.

ACKNOWLEDGEMENT

We thank the National Institutes of Health (NIH) programs Models of Infectious Disease Agent Study (MIDAS) for support through the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM088558. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. LKS is also supported by the Morris-Singer Foundation.

CONFLICT OF INTERESTS

All authors have no conflict of interests to disclose.

REFERENCES

[1] Liu Y, Eggo RM, and Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. *The Lancet*. 2020;395:e47. https://doi.org/10.1016/s0140-6736(20)30462-1.

[2] Valerio MA, Rodriguez N, Winkler P, et al. Comparing two sampling methods to engage hard-to-reach communities in research priority setting. *BMC Medical Research Methodology*. 2016;16(1):146.

[3] Yong SEF, Anderson DE, Wei WE, et al. Connecting clusters of COVID-19: an epidemiological and serological investigation. *The Lancet Infectious Diseases*. 2020;20:809–815. https://doi.org/10.1016/S1473-3099(20)30273-5.

[4] Ebola ça suffit ring vaccination trial consortium. The ring vaccination trial: a novel cluster randomised trial design to evaluate vaccine efficacy and effectiveness during outbreaks, with special reference to Ebola. *BMJ*. 2015;351:h3740.

[5] Hayes RJ, Moulton LH. *Cluster Randomised Trials*. 2nd edn. Boca Raton, FL: Chapman and Hall/CRC; 2017.

[6] Lohr SL. *Sampling: Design and Analysis*. 2nd edn. Boston, MA: Brooks/Cole; 2010.

[7] Sood N, Simon P, Ebner P. Seroprevalence of SARS-CoV-2-specific antibodies among adults in Los Angeles County, California, on April 10–11, 2020. *JAMA*. 2020;323(23):2425–2427. https://doi.org/10.1001/jama.2020.8279.

[8] Pollán M, Pérez-Gómez B, Pastor-Barriuso R, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *The Lancet*. 2020. https://doi.org/10.1016/S0140-6736(20)31483-5.

[9] Okell LC, Verity R, Watson OJ, et al. Have deaths from COVID-19 in Europe plateaued due to herd immunity? *The Lancet.* 2020;395:e110–e111.

[10] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B.* 2007;274:599–604.

[11] Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, UK: Oxford University Press; 1991.

[12] Long Q-X, Liu B-Z, Deng H-J, et al. Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nature Medicine.* 2020;26:845–848. https://doi.org/10.1038/s41591-020-0897-1.

[13] Park SY, Kim Y-M, Yi S, et al. Coronavirus disease outbreak in call center, South Korea. *Emerging Infectious Diseases Journal* (of the CDC). 2020;26(8). Accessed April 28, 2020. https://doi.org/10.3201/eid2608.201274.

[14] Lloyd-Smith JO, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. *Nature*. 2005;438(17):355–359.