



## Practice of Epidemiology

# Snowball Sampling Study Design for Serosurveys Early in Disease Outbreaks

Lee Kennedy-Shaffer\*, Xueting Qiu, and William P. Hanage

\* Correspondence to Dr. Lee Kennedy-Shaffer, Department of Mathematics and Statistics, Vassar College, 124 Raymond Avenue, Box 226, Poughkeepsie, NY 12604 (e-mail: [lkennedyschaffer@vassar.edu](mailto:lkennedyschaffer@vassar.edu)).

Initially submitted July 8, 2020; accepted for publication April 2, 2021.

Serological surveys can provide evidence of cases that were not previously detected, depict the spectrum of disease severity, and estimate the proportion of asymptomatic infections. To capture these parameters, survey sample sizes may need to be very large, especially when the overall infection rate is still low. Therefore, we propose the use of “snowball sampling” to enrich serological surveys by testing contacts of infected persons identified in the early stages of an outbreak. For future emerging pandemics, this observational study sampling design can answer many key questions, such as estimation of the 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. We provide examples, in the context of the coronavirus disease 2019 (COVID-19) pandemic, of studies and analysis methods that use a snowball sample and perform a simulation study that demonstrates scenarios where snowball sampling can answer these questions more efficiently than other sampling schemes. We hope such study designs can be applied to provide valuable information to slow the present pandemic as it enters its next stage and in early stages of future pandemics.

asymptomatic infection; contact tracing; coronavirus disease 2019; design effect; SARS-CoV-2; serosurvey sampling; transmission chain

Abbreviations: COVID-19, coronavirus disease 2019; ICC, intracluster correlation coefficient; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

There is great interest in the results of serosurveys based on antibodies against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), to indicate the true numbers of people infected so far in the coronavirus disease 2019 (COVID-19) pandemic, the proportion that might be immune in future waves of infection, and the proportion of people infected who experience mild or no symptoms. These and many other parameters important for the COVID-19 public health response were difficult to estimate early in the pandemic. In addition, while people with asymptomatic or presymptomatic cases are known to be capable of transmission, the proportion of infections caused by such cases is not clear. While it appears that SARS-CoV-2 transmission displays overdispersion, a high variance in secondary cases per index case around the population mean basic reproduction number ( $R_0$ ), we do not yet know what characteristics are associated with superspreading events (1–3).

In addition to estimating population immunity levels, serosurveys can be used to determine the proportion of infections that might have been undetected because they either were minimally symptomatic or exhibited symptoms that did not lead to testing. While random testing is common for such surveys, the cumulative incidence will vary greatly depending on the stage of the pandemic, and a very large sample may be necessary to obtain enough cases to capture less common disease presentations. An alternative study design based on “snowball sampling” offers a route for collecting data on the spectrum of clinical severity and transmissibility. This sampling approach is a form of survey sample enrichment for hard-to-reach populations (4, 5). While enriched sampling might typically focus on a marginalized or underrepresented community, snowball sampling in this context enriches for the presence of seropositivity. Similar enrichment can occur in studies focused on high-risk

populations (e.g., health-care workers (6)) or by studying family members of infected persons (7), but these are not as representative of the population and possible transmission routes in a more general setting.

Snowball sampling, or “chain referral sampling,” has a rich history, especially in the sociology research literature (8–10). The general research method involves identifying index individuals and, along with collecting information on them, asking them to refer other persons suitable for the study (8, 10). These named individuals are then recruited into the study. This process may end there or continue for further stages (9). The method is often used for qualitative studies of “hidden” populations, where subjects are difficult to reach or a sufficient sample is unlikely to be obtained from random sampling (5, 10, 11). It has more recently been used in infectious disease settings as a cost-effective way to recruit people into care and into studies (12–14). One study of human immunodeficiency virus infection found it to be a more cost-effective approach and to generate a more representative sample of the spectrum of disease than standard recruitment methods (14). The generalizability of results from snowball sampling has been criticized because the sample is not a true probability sample (11, 15, 16). This can be mitigated, however, by ensuring a random or representative selection of index individuals and by accounting for clustering in the analysis (15–18).

By taking people who are known to have been infected and tracing their contacts in order to identify possible transmission events, we are able to both estimate how many secondary cases were infected by an index case and obtain a larger data set of persons who have been infected, as well as persons who were contacts of the same index individual but not infected. The goal of snowball sampling, as we propose it, is not to determine the amount of population-level immunity, which is best addressed by conventional serosurveys, but to obtain a large number of persons who have been exposed in order to estimate the range of clinical presentations and their relation to transmission. We propose that investigators consider a snowball sampling approach based on contact tracing in order to more fully answer important questions about the clinical presentations of disease and factors associated with transmission in a cost-effective way.

### PROPOSED APPLICATION OF SNOWBALL SAMPLING FOR SARS-COV-2 SEROLOGICAL SURVEYS

We propose to apply the snowball sampling method to enrich survey samples in outbreaks for seropositive individuals. This relies on contact tracing, which is primarily used for outbreak mitigation and to identify linkages between transmission chains (19). Assuming that a serological assay with high specificity and sensitivity for previous SARS-CoV-2 infection is available, the method begins with a sample of persons who were infected. 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. Backward tracing, which incorporates potential infectors of the index case, may also be used. If feasible, virological testing can be used as well to identify contacts

with active infection who may not yet be seropositive. The sample can be enlarged either by adding more index cases and their contacts or by adding additional layers (contacts of contacts who test positive). Enlargement continues until the sample is estimated to have sufficient statistical power to answer the question(s) of interest. A history of symptoms relevant to the question of interest, covering the time period from the earliest possible infection date to the latest possible date yielding seropositivity, is collected for each identified case and contact.

### POTENTIAL APPLICATIONS AND QUESTIONS OF INTEREST

Several types of scientific questions of interest may be answered through the use of this design. For example, using only sampled persons who test positive, one can:

1. Identify the range of clinical presentations for an individual with a positive serology test (seropositive individual).
2. Identify the probability of a given course of disease for a seropositive individual.
3. Identify the association between a contact’s characteristics of interest (e.g., types of contact or personal characteristics) and their course of disease. For instance, some people may be more likely to experience minimally symptomatic infections.

With information on both seropositive and seronegative contacts, researchers can also answer questions about transmission, such as:

4. Identify the association between some characteristics of interest (of either the contact or index individual) and seropositivity among contacts of index individuals to determine which are associated with increased likelihood of transmission.

For example, cases who are asymptomatic (and were undetected by initial contact tracing) may be more likely to transmit because they are unaware of their status.

Table 1 provides examples of these 4 types of questions that are relevant for outbreaks of infectious diseases such as COVID-19, along with analysis methods that are appropriate for the snowball sampling design.

Other questions of interest may focus on the network structure itself, that is, the number of seropositive contacts of each index individual. Much of the snowball sampling literature focuses on these types of questions, and we refer readers there for appropriate analysis methods (8, 9, 15).

### STATISTICAL ANALYSIS

The primary difference in analysis between random sampling (as in standard serosurveys) and snowball sampling is that, in snowball sampling, estimation and inference must account for the clustered nature of the data. The contacts of an individual have a potentially shared exposure (and perhaps other latent shared characteristics) and thus can be viewed as a cluster (20, 21). We will assume for now that index individuals are a random sample of all possible index

**Table 1.** Examples of Scientific Questions of Interest in Early Disease Outbreaks That Can Be Answered Using Snowball Sample Serosurveys

Question Class	Example Question of Interest	Analysis Methods for Snowball Sampling	Sample Size Considerations	References for Analysis and Sample Size Methods
1. Identify the range of clinical presentations in seropositive persons.	Do people infected with COVID-19 experience severe joint pain?	Identify any such presentations among identified index individuals and contacts.	Clustering of clinical presentations by infector	20–22
2. Identify the probability of a given course of disease in seropositive persons.	What proportion of people infected with COVID-19 experience anosmia?	1-stage cluster sampling ratio estimation using a hierarchical model	Adjust binomial proportion estimation methods by appropriate design effect based on hypothesized ICC.	20–25, 35
3. Identify the association between personal characteristics and course of disease among seropositive contacts.	Is diabetes associated with an increased risk of hospitalization among people infected with COVID-19?	Logistic or multinomial logistic regression model, adjusted for clustering by index individual using a hierarchical mixed-effects model or fitted with generalized estimating equations	Sample size calculations for mixed-effects models or generalized estimating equations fits, using a hypothesized ICC	20–26, 31, 39–41
4. Identify association between characteristics of the index individual or contacts and seropositivity among contacts.	Is the age of the index individual associated with an increased risk of transmission to identified contacts?	Cochran-Mantel-Haenszel analysis and conditional logistic regression as used for analysis of matched case-control studies	Matched case-control study sample size methods	18, 25, 27–30, 33, 34

Abbreviations: COVID-19, coronavirus disease 2019; ICC, intracluster correlation coefficient.

individuals from some larger population of interest (e.g., all persons with confirmed infection in a given time range at a given geographic location, workplace, etc.). This could occur if the index individuals are identified through random surveys of the population or routine surveillance with a very sensitive test. This allows inference to proceed, treating the sample as a cluster sample from a larger population of clusters (22).

For questions 1 and 2, analysis can account for this potential correlation by using a hierarchical model, where the probability of a given clinical presentation differs depending on the index individual. A common assumption is that the cluster-specific probabilities are independent and identically distributed according to some distribution with mean  $\pi$  and variance  $\sigma^2$  (23, 24). This is often parameterized using the intracluster correlation coefficient (ICC),  $\rho = \frac{\sigma^2}{\pi(1-\pi)}$ . Details of this approach are given in the Web Appendix (available at <https://doi.org/10.1093/aje/kwab098>).

For question 3, we can use logistic or multinomial logistic regression approaches that account for the clustering of the data (22). Two common approaches are using mixed-effects models and fitting regression models using generalized estimating equations (20, 21, 25). Both approaches allow for the specification of either individual-level (i.e., character-

istics of the contact or the test used) or cluster-level (i.e., characteristics of the index case) covariates of interest in the model. Generalized estimating equations have the advantage of being robust to misspecification of the correlation model (21)—for example, if some of the persons identified by an index individual were actually infected by someone else. For mixed-effects modeling approaches, clustering parameters may not be interpretable if contacts include directly infected persons, persons infected by another source, and potential infectors of the index individual. However, this is equivalent to misspecifying the random-effects distribution, which in general has a minimal effect on estimation and inference (26). Additional details can be found in the Web Appendix.

Finally, for question 4, we are using not only the characteristics of the seropositive contacts but also those of the seronegative contacts. This mimics a case-control study design matched by index individual and thus can be analyzed similarly to other matched case-control studies; methods include Cochran-Mantel-Haenszel analysis stratified by index individual and conditional logistic regression models (25, 27–30). In a study on drug abuse, use of a snowball sample as the basis for a matched case-control study was shown to perform well and avoid selection bias (18).

The above methods use only data from the contacts. For questions 1–3, however, they can be easily extended to include the index individuals themselves and second-order contacts (i.e., contacts of contacts of index individuals). In these cases, multiple random-effects terms or more complex correlation structures may need to be specified. Depending on the question of interest, using more waves in snowball sampling can ensure greater sample diversity (17) but may introduce more difficulty for valid statistical inference (15).

We have so far assumed that the index individuals are a random sample of persons infected with SARS-CoV-2. If, however, the index individuals over- or underrepresent people with certain clinical presentations (e.g., if index individuals are selected among persons with symptomatic COVID-19) and the clinical presentation of an index case is related to the clinical presentation of their infected contact, the methods presented here will not represent the full population (5, 16). To account for this, one can use a systematic sample of index individuals that represents all clinical presentations or factors associated with them, ensuring diversity of index cases (17). We can adjust the methods appropriately by using stratified analyses based on the factors used in sampling the index individuals (21, 22, 31).

### 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 for various questions.

For question 1, we can calculate the required sample size (given an estimate of the number of seropositive contacts per index individual) by finding the number of index individuals that gives a specific probability of observing at least 1 contact with the clinical presentation, assuming a true underlying probability for that presentation. Clustering can be accounted for by specifying the ICC  $\rho$ . For questions 2 and 3, we must inflate the variance (and thus the required sample size) of a standard analysis by an appropriate design effect. This design effect can be estimated by  $DE = 1 + (\bar{m} - 1)\rho$ , where  $\bar{m}$  is the average number of seropositive contacts per index individual (21, 22, 32). More details can be found in the Web Appendix.

Note that for questions 1–3, the outcome of interest is not seropositivity but rather the clinical presentation itself. This probably has a lower ICC than seropositivity. In addition, some infected persons will have been infected by someone other than the index individual, so the hierarchical model may not be correctly specified. Sample size calculations are thus likely to be conservative, as the true correlation will be lower than the hypothesized correlation. For analyses with an individual-level covariate, precise methods for sample size estimation are not available, but the design effect using an appropriately adjusted ICC may be a reasonable approximation (21, 24, 31). More complex sample size formulae can be used if the analysis has multiple levels of clustering or stratification (21, 22, 31).

We can use these formulae to compare the required sample size for a snowball sample with that for a simple random sample. The inflated variance of the design is counterbalanced by the higher percentage of tested individuals who

are seropositive due to the enriched sample from this design. The effective sample size of a snowball sample is the number of identified seropositives divided by the design effect. This can be compared with the number of identified seropositives from a random serosurvey to determine the relative efficiency of the 2 designs. An example of this calculation is given in the Web Appendix.

For question 4, sample size and power calculation methods are available for matched case-control studies (33, 34). No simple comparison exists for the relative efficiency of these matched approaches and approaches based on a random sample.

### EXAMPLE: SIMULATED STUDY

To illustrate the use of this study design, consider a study that aims to identify a certain symptom of the disease and estimate the percentage of infected persons who experience that symptom (akin to questions 1 and 2 above). One design would be a random-sample serosurvey where investigators ask people who test positive whether they have experienced this symptom. The seropositive persons form the sample, and the proportion of these who experienced that symptom can be used as an estimate of the symptom rate. Inference can proceed using standard methods for binomial proportions.

Using snowball sampling, instead, a small number of index individuals who were known to be infected are asked to identify contacts during their potentially infectious period. These contacts are tested and asked whether they have experienced the symptom. The seropositive contacts form the sample. The proportion of these persons who experienced that symptom, corrected to account for clustering, can be used as an estimate of the symptom rate. Inference can proceed using clustered survey sample methods for binomial proportions (22, 35).

To illustrate the potential benefits of snowball sampling, we present results from a simulation study using a susceptible-exposed-infectious-recovered model in a population of 10,000 persons. Each individual in the population has a set of daily contacts, with an average of 20. A basic reproduction number of 2.5 is used with a dispersion parameter  $k = 0.1$ . Both the number of contacts an individual has and their “infectiousness” contribute to the overdispersion of transmission. For simplicity, infected individuals are assumed to be infectious for 8 days beginning with the second day after infection, with an equal probability of infecting a contact on each of those days. On average, 5% of individuals experience the symptom of interest, with additional variation due to a risk factor. The parameters used are described in Web Table 1.

We conduct 250 simulations per parameter combination, varying both the sampling time and the ICC of symptoms among infector and infectee. Note that by “sampling time” here, we are referring to the last time at which an individual could become infected and test positive on the serological assay. So the actual time at which testing occurred may be weeks later; incorporating virological testing as well would reduce the need for this wait time.

In each simulated outbreak, at the designated sampling time, we sample and test 600 persons at random for the

“regular sampling” approach and 30 index individuals, who were infected at least 10 days (i.e., the maximum generation interval) prior to the sampling time, for the “snowball sampling” approach. All of the contacts of the index individuals are then tested, for an average of 600 tested contacts. We also construct a “snowball, error” sample, where each index individual misses 2 of their true contacts and identifies 2 false contacts. For each sample, we estimate the symptom rate and construct a 95% confidence interval using a logit transformation approach, appropriately adjusted for clustering by index individual for the snowball samples (35). For simplicity, we conservatively exclude anyone who is identified as a contact by multiple index individuals, although methods for overlapping clusters could be used instead. R code for replicating this simulation study or applying the sampling methodology to a more specific simulation setting is available on GitHub (36).

First, to see how well the methods identify this symptom, we can compare the percentages of each simulation with at least 1 symptomatic individual in the sample. At the earliest sampling time, when the population prevalence is around 5%, only 77% of the regular samples included at least 1 symptomatic individual, while over 97% of the snowball samples did. Incorporating contact recall error in the snowball samples reduced this percentage by less than 1%.

To compare performance on question 2, estimation of the symptom rate among infected persons, Figure 1 displays the number of infections identified in each sample (Figures 1A and 1B), the median and interquartile range of the estimated symptom rate (Figures 1C and 1D), and the root mean squared error of estimation (Figures 1E and 1F) across all simulations for each method, by sampling time and ICC. All methods provide unbiased estimation of the true symptom percentage at the sampling time. However, the snowball samples, with or without contact error, identify more infections and thus have a smaller interquartile range and a lower root mean squared error than the regular samples. Figure 2 shows the 95% confidence interval width (Figures 2A and 2B) and empirical coverage (Figures 2C and 2D) across all simulations for each method by sampling time and ICC. All methods achieve nominal coverage, but the snowball sample estimates have narrower confidence intervals, indicating higher precision. The differences are less pronounced at later sampling times (where the background prevalence is higher) and when the ICC is higher (where the clustered analysis lowers precision). The highest ICC considered, 0.10, is shown in Figure 3. Simulations with less overdispersion of transmission demonstrate similar results (see Web Figures 1–6).

This simulation demonstrates the potential value of snowball sampling in increasing the precision of the estimated symptom rate among infected persons by enriching the sample for infected individuals. It also demonstrates that minor violations of assumptions, such as some incorrect contact identification, do not negate the benefit of snowball sampling. The approach described here is agnostic to the true infector of a contact, so, for the ICCs studied here, the method is robust to whether the index individual includes their infector as a contact and to whether there are alternative sources of infection for the identified contacts.

This simulation study is limited by its simplicity; more complex models for parameters of interest, contact matrices, and transmission parameters can be incorporated to assess the benefits of snowball sampling in a specific setting, for COVID-19 or another disease. It also does not account for imperfect testing sensitivity and specificity and delays in seroconversion, although these would affect both the regular and snowball samples.

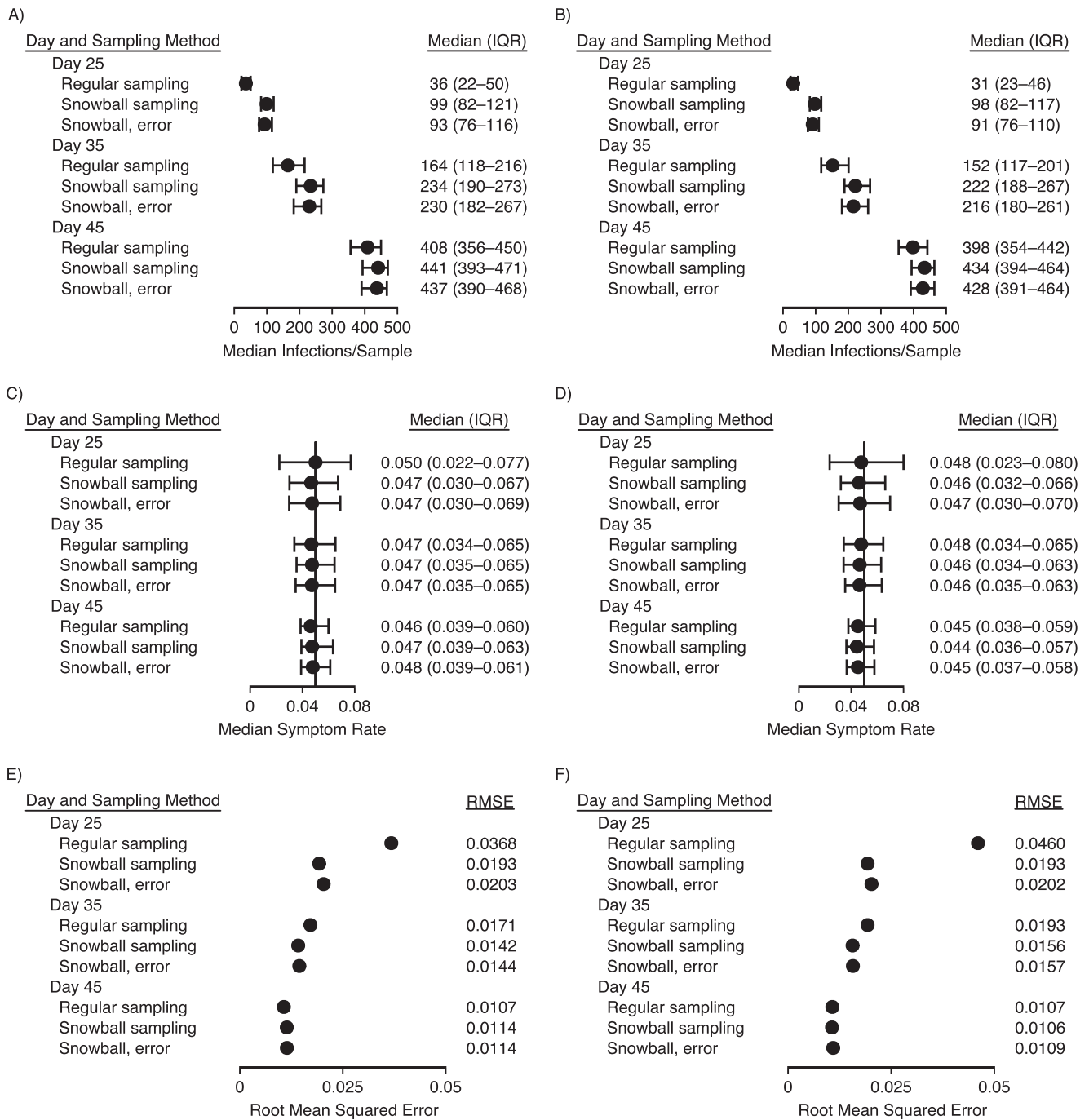
## DISCUSSION

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 those among hard-to-reach individuals who may not have had contact with health-care providers (14). 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 in onward transmission is critical to designing appropriate responses, as existing studies may reflect changing contact patterns due to public awareness of the disease rather than biological patterns of infectiousness (37). 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. As in all studies of secondary attack rates, appropriate definition of contacts is crucial to obtaining unbiased estimates (38).

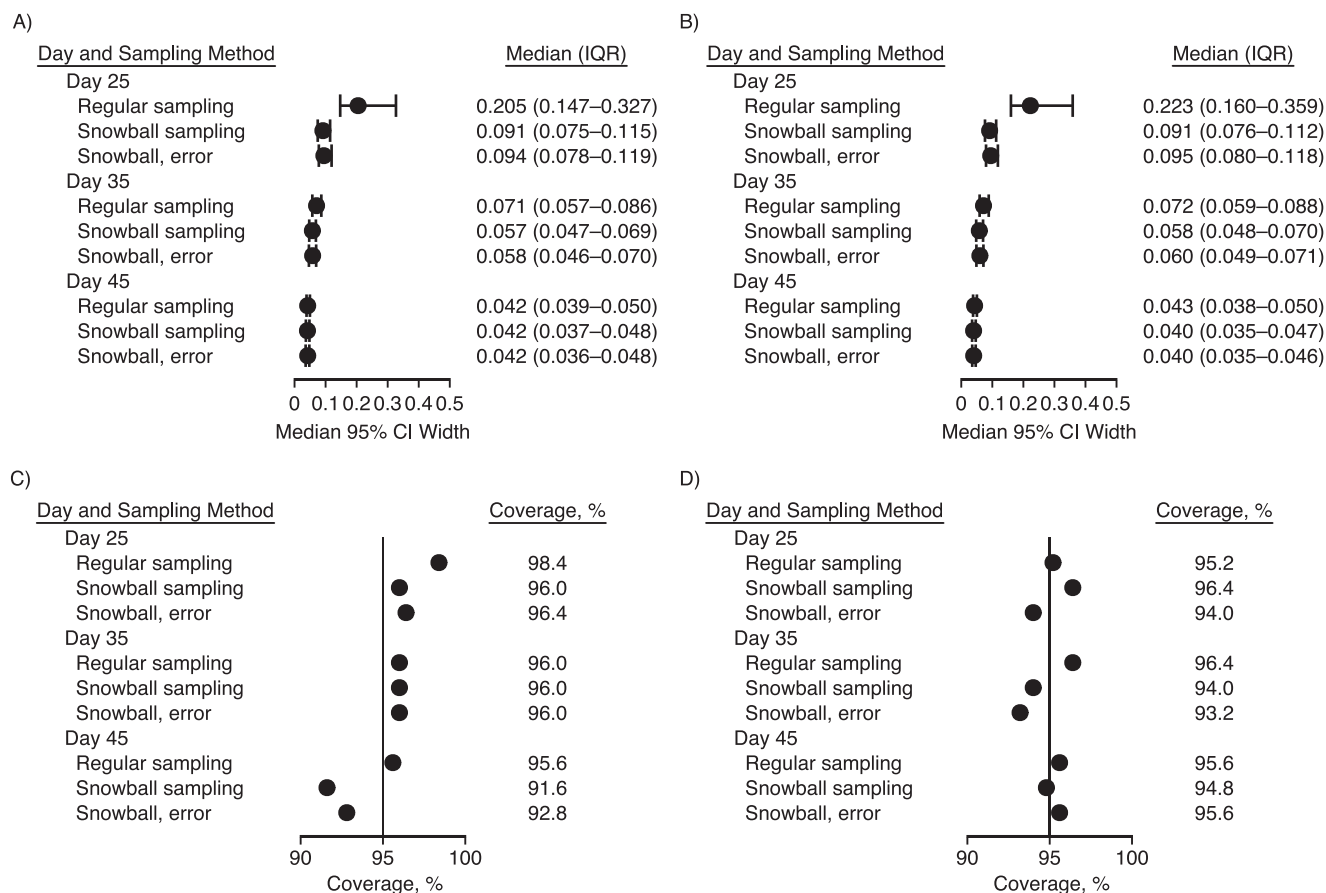
Compared with other enriched designs, this approach is not limited to certain segments of the population and thus provides a more representative sample of clinical presentations and demographic factors. It also provides a larger and more representative sample per index individual than only sampling household members, allowing better statistical power to answer a wider range of questions of interest.

There are settings where the snowball sampling design is not feasible, however, or where it has minimal benefits compared with random sampling. First, it is not suited to estimation of overall population seroprevalence. Second, it is less beneficial in a more mature epidemic, where the higher overall prevalence reduces the efficiency of the snowball sample enrichment. In a later epidemic, nonpharmaceutical interventions may have also reduced the reproduction number, thus reducing the number of seropositive persons per index case. These interventions (e.g., social distancing) may also reduce contacts, however, thus preserving the relative benefit of snowball sampling. Third, if there is substantial misclassification of close contacts or bias in recall by index individuals, this can limit the benefits of the design and even lead to bias. In particular, if index individuals are more likely to recall contacts who did experience symptoms, this approach may lead to overestimation of the rates of severe clinical presentations. Finally, a high proportion of superspreading events will lead to high variance in cluster sizes, reducing the effective size of the snowball sample.

From the practical side, there is a cost of identifying and reaching index individuals and their close contacts above that of identifying and reaching randomly sampled



**Figure 1.** Accuracy and variability of the estimation of the proportion of infected individuals with symptoms from simple random samples, snowball samples, and snowball samples with error in contact identification. The median number of infections per sample (•) and its interquartile range (IQR; bars) (panels A and B), the median estimated symptom rate (proportion of infected individuals who experience symptoms) (•) and its IQR (bars) (panels C and D), and the root mean squared error (RMSE) of the estimated symptom rate (panels E and F) are compared by sampling time (days 25, 35, and 45), the intracluster correlation coefficient (ICC) of infector and infectee symptom status (ICC = 0 in panels A, C, and E and ICC = 0.05 in panels B, D, and F), and sampling method, with the default dispersion parameter  $k = 0.1$ . The underlying probability of being symptomatic given infection is 5% (vertical line in panels C and D). All symptom rates were estimated using the logit transformation; estimates for the 2 snowball samples were adjusted for clustering by the index individual, with contacts named by 2 or more index individuals removed. Results are from 250 simulations per parameter combination.



**Figure 2.** Confidence interval (CI) width and coverage for estimation of the proportion of infected individuals with symptoms from simple random samples, snowball samples, and snowball samples with error in contact identification. The median 95% CI width (•) and its interquartile range (IQR; bars) (panels A and B) and the empirical coverage of the 95% CIs for the symptom rate among infected persons (panels C and D) are compared by sampling time (days 25, 35, and 45), the intracluster correlation coefficient (ICC) of infector and infectee symptom status (ICC = 0 in panels A and C and ICC = 0.05 in panels B and D), and sampling method, with the default dispersion parameter  $k = 0.1$ . The nominal CI coverage is 95% (vertical line in panels C and D). Results are from 250 simulations per parameter combination.

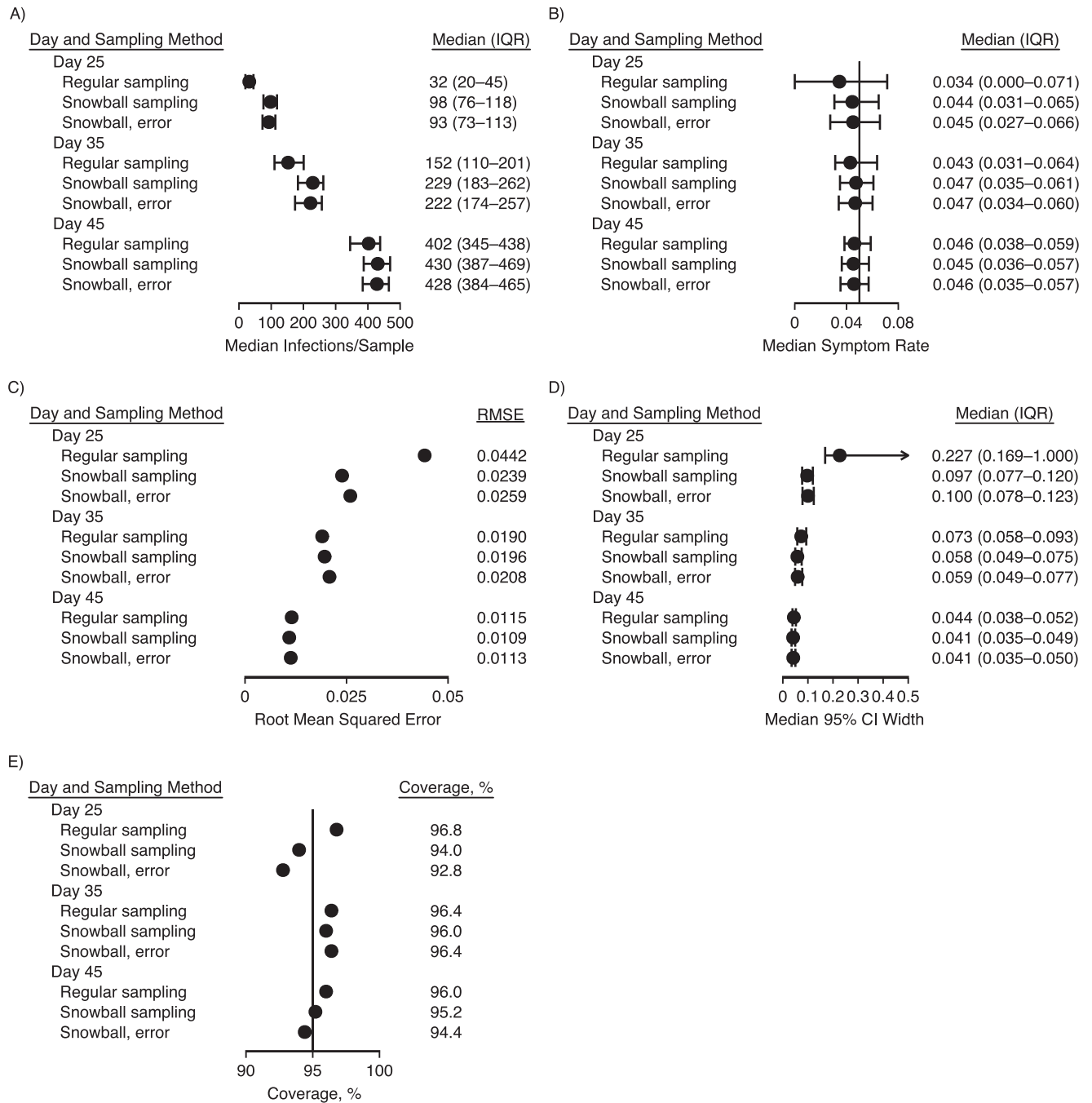
individuals. This may reduce the number of people who can be sampled. This is particularly the case if a true random sample of index individuals is desired so that the results are generalizable, rather than use of a convenience sample of index individuals (5). The use of preexisting contact tracing information from public health authorities will reduce the contact tracing labor required and thus the cost of snowball sampling. It will also reduce contact recall bias by shortening the time delay. In further work, researchers should consider whether contact tracing efforts with virological testing provide enough information to be used as de facto snowball samples for retrospective analysis.

As in all serological surveys, the results depend on the sensitivity and specificity of the assay used. Low sensitivity will lead to the exclusion of cases, reducing the sample size achieved and potentially reducing the representativeness of clinical presentations. Low specificity will lead to the inclusion of noninfected persons as index individuals, resulting in wasted resources, and could also lead to bias in determining the association between characteristics and

the likelihood of infection among contacts. By enriching the sample through snowball sampling, resources that might otherwise be applied to sampling of individuals could be directed toward the use of a more accurate test, mitigating some of these problems. If these test characteristics are known, then analysis methods can be adjusted to account for them.

Since directionality is important in identifying transmission risk factors but cannot be fully established in general in such surveys, evaluations of factors that predict transmission should be interpreted with caution (37). In some cases, limiting the analysis to cases where temporal ordering of infection can be established may be warranted.

Finally, the symptom or risk factor being accurately recalled by the close contacts is also important for obtaining unbiased results. This will depend upon the elapsed time between their possible infection and the conduct of the survey, as well as the specificity of the symptom under investigation. For many of these factors, surveys carried out early in the epidemic will perform better than those



**Figure 3.** Accuracy, variability, and 95% confidence interval (CI) width and coverage for estimation of the proportion of infected individuals with symptoms from simple random samples, snowball samples, and snowball samples with error in contact identification, for an intracluster correlation coefficient (ICC) of 0.10. The median number of infections per sample (•) and its interquartile range (IQR; bars) (A), the median estimated symptom rate (proportion of infected individuals who experience symptoms) (•) and its IQR (bars) (B), the root mean squared error (RMSE) of the estimated symptom rate (C), the median 95% CI width and its IQR (bars) (D), and the empirical coverage of the 95% CIs for the symptom rate (E) are compared by sampling time (days 25, 35, and 45) and sampling method, with the default dispersion parameter  $k = 0.1$  and with  $ICC = 0.10$ . The underlying probability of being symptomatic given infection is 5% (vertical line in panel B), and the nominal CI width is 95% (vertical line in panel E). All symptom rates were estimated using the logit transformation; estimates for the 2 snowball samples were adjusted for clustering by the index individual, with contacts named by 2 or more index individuals removed. Results are from 250 simulations per parameter combination.



done later in the epidemic, so snowball sampling might be an effective approach to ensuring adequate power.

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 persons. It should be considered for use in locations that are still in the early stage of the COVID-19 pandemic, and its properties should be further studied so the method can be improved and used in future infectious disease outbreaks.

## ACKNOWLEDGMENTS

Author affiliations: Department of Mathematics and Statistics, Vassar College, Poughkeepsie, New York, United States (Lee Kennedy-Shaffer); Center for Communicable Disease Dynamics, T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts (Lee Kennedy-Shaffer, Xueting Qiu, William P. Hanage); and Department of Epidemiology, T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts (Lee Kennedy-Shaffer, Xueting Qiu, William P. Hanage).

This work was funded by the Models of Infectious Disease Agent Study (MIDAS), National Institutes of Health, through the National Institute of General Medical Sciences (award U54GM088558). L.K.-S. was also supported by the Morris-Singer Foundation and a subcontract from Carnegie Mellon University under an award from the Centers for Disease Control and Prevention (award U01IP001121).

All simulated data generated for this study, as well as R code with which to reproduce the simulations, analyses, and figures, are available on GitHub (36).

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest: none declared.

## REFERENCES

- Liu Y, Eggo RM, Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet*. 2020; 395(10227):e47.
- Lloyd-Smith JO, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. *Nature*. 2005;438(7066):355–359.
- Endo A, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott S, et al. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res*. 2020;5: Article 67.
- Valerio MA, Rodriguez N, Winkler P, et al. Comparing two sampling methods to engage hard-to-reach communities in research priority setting. *BMC Med Res Methodol*. 2016; 16(1):Article 146.
- Frank O, Snijders T. Estimating the size of hidden populations using snowball sampling. *J Off Stat*. 1994;10(1): 53–67.
- World Health Organization. *Assessment of Risk Factors for Coronavirus Disease 2019 (COVID-19) in Health Workers: Protocol for a Case-Control Study*. (Version 1.0). Geneva, Switzerland: World Health Organization; 2020. [https://www.who.int/publications/i/item/assessment-of-risk-factors-for-coronavirus-disease-2019-\(covid-19\)-in-health-workers-protocol-for-a-case-control-study](https://www.who.int/publications/i/item/assessment-of-risk-factors-for-coronavirus-disease-2019-(covid-19)-in-health-workers-protocol-for-a-case-control-study). Accessed March 27, 2021.
- World Health Organization. *Household Transmission Investigation Protocol for 2019–Novel Coronavirus (COVID-19) Infection*. (Version 2.2). Geneva, Switzerland: World Health Organization; 2020. [https://www.who.int/publications/i/item/household-transmission-investigation-protocol-for-2019-novel-coronavirus-\(2019-ncov\)-infection](https://www.who.int/publications/i/item/household-transmission-investigation-protocol-for-2019-novel-coronavirus-(2019-ncov)-infection). Accessed March 27, 2021.
- Coleman JS. Relational analysis: the study of social organizations with survey methods. *Hum Organ*. 1958;17(4): 28–36.
- Goodman LA. Snowball sampling. *Ann Math Stat*. 1961; 32(1):148–170.
- Biernacki P, Waldorf D. Snowball sampling: problems and techniques of chain referral sampling. *Sociol Methods Res*. 1981;10(2):141–163.
- Kalton G, Anderson DW. Sampling rare populations. *J R Stat Soc Ser A*. 1986;149(1):65–82.
- Kendall C, Kerr LRFS, Gondim RC, et al. An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil. *AIDS Behav*. 2008;12(4 suppl):S97–S104.
- Siddiqui NA, Rabidas VN, Sinha SK, et al. Snowball vs. house-to-house technique for measuring annual incidence of kala-azar in the higher endemic blocks of Bihar, India: a comparison. *PLoS Negl Trop Dis*. 2016;10(9):e0004970.
- Chhim S, Macom J, Pav C, et al. Using risk-tracing snowball approach to increase HIV case detection among high-risk populations in Cambodia: an intervention study. *BMC Infect Dis*. 2017;17(1):Article 689.
- Snijders TAB. Estimation on the basis of snowball samples: how to weight? *Bull Methodol Sociol*. 1992;36(1): 59–70.
- Sedgwick P. Snowball sampling. *BMJ*. 2013;347(2):f7511.
- Kirchherr J, Charles K. Enhancing the sample diversity of snowball samples: recommendations from a research project on anti-dam movements in Southeast Asia. *PLoS One*. 2018; 13(8):e0201710.
- Lopes CS, Rodrigues LC, Sichieri R. The lack of selection bias in a snowball sampled case-control study on drug abuse. *Int J Epidemiol*. 1996;25(6):1267–1270.
- Yong SEF, Anderson DE, Wei WE, et al. Connecting clusters of COVID-19: an epidemiological and serological investigation. *Lancet Infect Dis*. 2020;20(7):809–815.
- 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.
- Hayes RJ, Moulton LH. *Cluster Randomised Trials*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press; 2017.
- Lohr SL. *Sampling: Design and Analysis*. 2nd ed. Boston, MA: Brooks/Cole Publishing Company; 2010.
- Commenges D, Jacqmin H. The intraclass correlation coefficient: distribution-free definition and test. *Biometrics*. 1994;50(2):517–526.

24. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev.* 2009;77(3):378–394.
25. Agresti A. *Categorical Data Analysis*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2003.
26. McCulloch CE, Neuhaus JM. Misspecifying the shape of the random effects distribution: why getting it wrong may not matter. *Stat Sci.* 2011;26:388–402.
27. Pearce N. Analysis of matched case-control studies. *BMJ.* 2016;352:i969.
28. Cochran WG. Some methods for strengthening the common  $\chi^2$  tests. *Biometrics.* 1954;10(4):417–451.
29. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst.* 1959;22(4):719–748.
30. Breslow NE, Day NE, Halvorsen KT, et al. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epidemiol.* 1978;108(4):299–307.
31. Kennedy-Shaffer L, Hughes MD. Sample size estimation for stratified individual and cluster randomized trials with binary outcomes. *Stat Med.* 2020;39(10):1489–1513.
32. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol.* 2015;44(3):1051–1067.
33. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med.* 2002;21(1):35–50.
34. Lachin JM. Sample size evaluation for a multiply matched case-control study using the score test from a conditional logistic (discrete Cox PH) regression model. *Stat Med.* 2008;27(14):2509–2523.
35. Dean N, Pagano M. Evaluating confidence interval methods for binomial proportions in clustered surveys. *J Surv Stat Methodol.* 2015;3(4):484–503.
36. Kennedy-Shaffer L. Snowball-sampling. <https://github.com/leekshaffer/snowball-sampling>. Published March 31, 2021. Accessed March 31, 2021.
37. Park SY, Kim Y-M, Yi S, et al. Coronavirus disease outbreak in call center, South Korea. *Int Conf Emerg Infect Dis.* 2020;26(8):1666–1670.
38. Accorsi E, Qiu X, Rumpler E, et al. How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19. *Eur J Epidemiol.* 2021;36(2):179–196.
39. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. *Biometrics.* 1984;40(4):961–971.
40. Hedeker D. A mixed-effects multinomial logistic regression model. *Stat Med.* 2003;22(9):1433–1446.
41. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;42(1):121–130.