Early Release Estimates for SARS-CoV-2 Prevalence and Antibody Response Interim Weighting for Probability-Based Sample Surveys

Authors

*Heather Bradley, Assistant Professor, Georgia State University School of Public Health

*Mansour Fahimi, Chief Data Scientist, Marketing Systems Group

Travis Sanchez, Research Associate Professor, Emory University Rollins School of Public Health

Ben Lopman, Professor, Emory University Rollins School of Public Health

Martin Frankel, Professor Emeritus, Baruch College, City University of New York

Colleen F. Kelley, Associate Professor, Emory University School of Medicine

Richard Rothenberg, Professor, Georgia State University School of Public Health

Aaron J Siegler, Associate Professor, Emory University Rollins School of Public Health

Patrick S Sullivan, Professor, Emory University Rollins School of Public Health

(*Co-first authors)

Funding: NIAID 3R01AI143875-02S1, Woodruff Foundation

Abstract

Many months into the SARS-CoV-2 pandemic, basic epidemiologic parameters describing burden of disease are lacking. To reduce selection bias in current burden of disease estimates derived from diagnostic testing data or serologic testing in convenience samples, we are conducting a national probability-based sample SARS-CoV-2 serosurvey. Sampling from a national address-based frame and using mailed recruitment materials and test kits will allow us to estimate national prevalence of SARS-CoV-2 infection and antibodies, overall and by demographic, behavioral, and clinical characteristics. Data will be weighted for unequal selection probabilities and non-response and will be adjusted to population benchmarks. Due to the urgent need for these estimates, expedited interim weighting of serosurvey responses will be undertaken to produce early release estimates, which will be published on the study website, COVIDVu.org. Here, we describe a process for computing interim survey weights and guidelines for release of interim estimates.

Keywords

SARS-CoV-2, serosurvey, population-based survey

Introduction

SARS-CoV-2 is responsible for more than 190,000 deaths in the U.S. to date,[1] yet 7 months into the pandemic, much remains unknown about how many individuals have been infected or their demographic characteristics. Diagnostic testing has been fraught with implementation challenges. Reported cases are largely a reflection of testing among those who suspect they may have been infected, so case reports are likely to severely undercount mildly symptomatic or asymptomatic infections, or infections among people unwilling or unable to be tested. Population-based screening

strategies are urgently needed to understand population-level prevalence of SARS-CoV-2 infection and antibody response. To that end, SARS-CoV-2 serosurveys, which pair PCR and/or antibody screening with demographic and/or behavioral data collection, have recently been launched in Europe[2-4] and the U.S.[5-10] Serosurveys conducted or launched in the U.S. to date use are designed to produce seroprevalence estimates for individual counties or states[5, 6, 9, 11], or use convenience samples that may not be generalizable to the underlying population.[8, 10] To fill an on-going need for nationally representative estimates of SARS-CoV-2 disease prevalence, incidence, and antibody response, we are conducting a national probability-based sample serosurvey.[12] Using baseline serosurvey data, we will estimate national prevalence of SARS-CoV-2 infection and antibodies, overall and by demographic, behavioral, and clinical characteristics.

Probability-based sample surveys are needed for estimation of population-level prevalence that is robust to selection bias. Participants are selected at random with known and nonzero probabilities of selection to allow computation of extrapolation factors (weights) for inferential purposes. Weighting processes include computation of design weights to reflect selection probabilities of sampled units, as well as a series of adjustments to compensate for differential nonresponse and under-coverage.[13] Generally, weighting of survey data is undertaken after all survey responses have been collected to allow a full treatment of observed nonresponse patterns after all responses have been received. Due to the complexity of these analytic procedures, weighted estimates from most such surveys are often released months after data collection is completed (see, for example[14-16]). Because of the urgent need for population-based estimates of SARS-CoV-2 prevalence and antibody response, here we describe an expedited interim weighting procedure of serosurvey responses that we will use to produce early release estimates, which will be published on the study website, COVIDVu.org.

Methods

For this study, our overall target is a national sample of 4,000 U.S. adults completing study procedures, and an additional 3,584 adults residing in seven states of interest (CA, FL, GA, NY, IL, TX, WA). Sampling procedures have been previously described in detail.[12] Briefly, households will be selected from an address-based sampling frame created by Marketing Systems Group from the latest Delivery Sequence File of the U.S. Postal Service.[17] Recruitment materials and kits for self-collecting SARS-CoV-2 testing specimens will be mailed to households, and adults will be sampled for participation within households based on household enumeration. Generally, one adult per household will self-collect specimens for PCR and antibody testing and complete a survey, but full households will be included in 10% of randomly selected, participating households. Surveys will be completed online, and specimens will be returned through the mail for lab testing. These procedures will be repeated three months later with persons participating at baseline for incidence estimation. Primary study outcomes will be prevalence and incidence of SARS-CoV-2 infection and antibodies.[12]

Computation of survey weights

Weighting processes usually entail four major steps. In the first step, design weights are computed to reflect selection probabilities of households and, in the case of the present survey, subsampling of adults in sample households. In the second step, design weights are adjusted to correct for nonresponse observed during the survey administration. In the third step, nonresponse-adjusted design weights are adjusted against population benchmarks so that the final weights conform to the target population distributions with respect to a set of demographic characteristics. For general population surveys these characteristics include gender, age, race/ethnicity, education, household income, region, and metropolitan status. For adjustment to population benchmarks, an iterative procedure commonly known as raking is used so that respondents' distributions can be adjusted to multiple benchmarks

simultaneously.[18] Finally, weights are examined, and if necessary, trimmed at both ends of the distribution to avoid extreme weights that can result in unstable estimates.

As is typically done in other weighting processes, missing demographic data for variables used to weight our survey data will be imputed prior to weight computation, although based on our previous work using web-based surveys, we expect minimal missing data for such data.[19, 20] We will use a hot-deck imputation procedure to replace missing values, using observed values from respondents with non-missing data for a given element ("donors") who are deemed to be otherwise demographically similar to respondents for whom the data element is missing.[21]

Computation of interim survey weights

Given the urgent need to produce expedited estimates, we will employ an interim weighting methodology that is an abbreviated version of the standard steps previously described in two notable ways. First, the nonresponse adjustment (Step 2) will be skipped and postponed for the final weighting process when all respondents and nonrespondents have been identified. Second, due to smaller sample sizes available for interim weighting, some of the weighting variables may be collapsed into coarser categories. For example, we may use four categories of education level in final weighting but collapse data into two categories of education level for interim weighting. This need for parsimony may also require replacement of multivariate raking benchmarks with their corresponding marginal distributions, or averages.

Interim weighting will seek to balance potential bias reduction against variance inflation, which is an inevitable consequence of weighting. To accomplish this, we will ensure that (1) there are enough respondents to "carry" the weights to avoid an unstable scenario when a few respondents with extreme weights can heavily influence the resulting estimates and (2) the impact of weighting vis-à-vis the resulting unequal weighting effect is kept to a minimum to avoid undue loss of precision due to excessive weighting.

Following the above guidelines, the first set of interim weights will be produced after 1,000 surveys have been completed and accompanying specimens have been returned. We will require at least 1,000 respondents for interim national estimates and 200 respondents for sub-group estimates (e.g. by age group, race, or state-specific for over-sampled states). As responses accumulate, more robust sub-group estimates will be made possible by increasing the granularity of the weighting adjustments. Having adequate precision for interim estimates using these sample size guidelines assumes, on average, 1% prevalence of SARS-CoV-2 virus and 3% prevalence of SARS-CoV-2 antibodies.[5, 7, 10, 11] If observed prevalence for these outcomes are higher overall or in sub-groups, we may reduce the minimum sample size requirements. After adequate sample size is reached for computation of the first set of interim weights and estimation of prevalence of SARS-CoV-2 infection and antibodies, interim weighting and outcome estimation and dissemination will be conducted periodically until the serosurvey is complete.

Presentation of final estimates

When the serosurvey is complete and final weights have been computed, we will use the resulting design effect (Deff) as a surrogate to assess the impact of unequal weighting effect to set guidelines for adequate stability of estimates to be presented. The Deff is a commonly used metric to measure the efficiency of a weighting methodology to capture the impact of unequal weighting across respondents. While application of weights tends to improve the representation of survey respondents, and hence reduces bias in survey estimates, this gain comes at a precision cost because weighting increases variance of survey estimates. The inflation due to weighting can be approximated by the following formula, in which W_i represents the final weight of the i^{th} respondent[22] [23]:

$$Deff = 1 + CV_W^2 = 1 + \frac{\sum_i \frac{(W_i - \bar{W})^2}{n-1}}{\bar{W}^2}$$

Results

Recruitment for this study began in July, 2020. We anticipate interim findings on prevalence of SARS-CoV-2 infection and antibodies will be available on COVIDVu.org by November, 2020.

Discussion

Population-based estimates of national SARS-CoV-2 infection and antibody prevalence are critical for improving our understanding of burden of COVID-19 disease. Expediting such estimates through the use of interim weighting will allow data from our on-going probability-based sample survey to inform prevention and control measures during a time when they are acutely needed. While interim estimates may diverge somewhat from final estimates due to evolving data availability, the emergency nature of the SARS-CoV-2 pandemic requires that precision of key estimates be balanced against timeliness. Both interim and final estimates will be publicly available on COVIDVu.org, where they will be interpreted and visualized for use by researchers, policy makers, clinicians, and public health program administrators. Estimates of SARS-CoV-2 infection and antibody prevalence, overall and in sub-groups, will provide an empirical foundation that will allow other surveillance data sources to be contextualized and used more robustly.

References

1. Centers for Disease Control and Prevention. CDC COVID Data Tracker 2020 [August 19, 2020]. Available from: https://www.cdc.gov/covid-data-tracker/.

2. Pollan M, Perez-Gomez B, Pastor-Barriuso R, Oteo J, Hernan MA, Perez-Olmeda M, Sanmartin JL, Fernandez-Garcia A, Cruz I, Fernandez de Larrea N, Molina M, Rodriguez-Cabrera F, Martin M, Merino-Amador P, Leon Paniagua J, Munoz-Montalvo JF, Blanco F, Yotti R, Group E-CS. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. Lancet. 2020. doi: 10.1016/S0140-6736(20)31483-5. PubMed PMID: 32645347; PMCID: PMC7336131.

3. Stringhini S, Wisniak A, Piumatti G, Azman AS, Lauer SA, Baysson H, De Ridder D, Petrovic D, Schrempft S, Marcus K, Yerly S, Arm Vernez I, Keiser O, Hurst S, Posfay-Barbe KM, Trono D, Pittet D, Getaz L, Chappuis F, Eckerle I, Vuilleumier N, Meyer B, Flahault A, Kaiser L, Guessous I. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. Lancet. 2020;396(10247):313-9. doi: 10.1016/S0140-6736(20)31304-0. PubMed PMID: 32534626; PMCID: PMC7289564.

4. Ward H, Atchison CJ, Whitaker M, Anslie KEC, Elliot J, Okell LC, Redd R, Ashby D, Donnelly CA, Barclay W, Darzi A, Cooke G, Riley S, Elliot P. Antibody prevalence for SARS-CoV-2 in England following first peak of the pandemic: REACT2 study in 100,000 adults. MedRxiv. 2020. doi: https://doi.org/10.1101/2020.08.12.20173690.

5. Biggs HM, Harris JB, Breakwell L, Dahlgren FS, Abedi GR, Szablewski CM, Drobeniuc J, Bustamante ND, Almendares O, Schnall AH, Gilani Z, Smith T, Gieraltowski L, Johnson JA, Bajema KL, McDavid K, Schafer IJ, Sullivan V, Punkova L, Tejada-Strop A, Amiling R, Mattison CP, Cortese MM, Ford SE, Paxton LA, Drenzek C, Tate JE, Team CDCFS. Estimated Community Seroprevalence of SARS-CoV-2 Antibodies - Two Georgia Counties, April 28-May 3, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(29):965-70. doi:

10.15585/mmwr.mm6929e2. PubMed PMID: 32701941; PMCID: PMC7377817 Journal Editors form for disclosure of potential conflicts of interest. No potential conflicts of interest were disclosed.

6. Feehan AK, Fort D, Garcia-Diaz J, Price-Haywood E, Velasco C, Sapp E, Pevey D, Seoane L. Seroprevalence of SARS-CoV-2 and Infection Fatality Ratio, Orleans and Jefferson Parishes, Louisiana, USA, May 2020. Emerg Infect Dis. 2020;26(11). doi: 10.3201/eid2611.203029. PubMed PMID: 32731911.

7. Havers FP, Reed C, Lim T, Montgomery JM, Klena JD, Hall AJ, Fry AM, Cannon DL, Chiang CF, Gibbons A, Krapiunaya I, Morales-Betoulle M, Roguski K, Rasheed MAU, Freeman B, Lester S, Mills L, Carroll DS, Owen SM, Johnson JA, Semenova V, Blackmore C, Blog D, Chai SJ, Dunn A, Hand J, Jain S, Lindquist S, Lynfield R, Pritchard S, Sokol T, Sosa L, Turabelidze G, Watkins SM, Wiesman J, Williams RW, Yendell S, Schiffer J, Thornburg NJ. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. JAMA Intern Med. 2020. doi: 10.1001/jamainternmed.2020.4130. PubMed PMID: 32692365.

8. Robertson M, Kulkami S, Berry A, Mirzayi C, Maroko A, Zimba R, Westmoreland D, Grov C, Parcesepe A, Waldron L, Nash D. A national prospective cohort study of SARS/COV2 pandemic outcomes in the U.S.: The CHASING COVID Cohort. MedRxiv. 2020. Epub May 4, 2020. doi: https://doi.org/10.1101/2020.04.28.20080630.

9. Sood N, Simon P, Ebner P, Eichner D, Reynolds J, Bendavid E, Bhattacharya J. Seroprevalence of SARS-CoV-2-Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. JAMA. 2020. doi: 10.1001/jama.2020.8279. PubMed PMID: 32421144; PMCID: PMC7235907.

10. Rosenberg ES, Tesoriero J, Rosenthal EM, Chung R, Barranco MA, Styer LM, Parker MM, Leung SJ, Morne J, Greene D, Holtgrave DR, Hoefer D, Kumar J, Udo T, Hutton B, Zucker HA. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. Annals of Epidemiology. 2020;48:23 - 9. doi: https://doi.org/10.1101/2020.05.25.20113050; PMCID: PMC7297691.

11. Menachemi N, Yiannoutsos CT, Dixon BE, Duszynski TJ, Fadel WF, Wools-Kaloustian KK, Unruh Needleman N, Box K, Caine V, Norwood C, Weaver L, Halverson PK. Population Point Prevalence of SARS-CoV-2 Infection Based on a Statewide Random Sample - Indiana, April 25-29, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(29):960-4. doi: 10.15585/mmwr.mm6929e1. PubMed PMID: 32701938; PMCID: PMC7377824 Journal Editors form for disclosure of potential conflicts of interest. Nir Menachemi reports a grant from State of Indiana which funded this study. Virginia Caine reports that she is a member of the MMWR Editorial Board. Brian E. Dixon and William F. Fadel report grants from the Indiana State Department of Health. Paul K. Halverson reports a grant from the State of Indiana. No other potential conflicts of interest were disclosed.

12. Siegler AJ, Sullivan PS, Sanchez T, Lopman B, Fahimi M, Sailey C, Frankel M, Kelley CF, Rothenberg R, Bradley H. Protocol for a national probability survey using home specimen collection methods to assess prevalence and incidence of SARS-CoV-2 infection and antibody response. Annals of Epidemiology. 2020;In press. Epub August 10, 2020. PubMed PMID: 32791199

13. Fahimi M, Barlas F, Thomas R, Buttermore N. Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. Survey Practice 2015;8(5).

14. Ashman JJ, Schappert SM, Santo L. Emergency Department Visits Among Adults Aged 60 and Over: United States, 2014-2017. NCHS Data Brief. 2020(367):1-8. PubMed PMID: 32600519.

15. Centers for Disease Control and Prevention. Behavioral and Clinical Characteristics of Persons with Diagnosed HIV Infection—Medical Monitoring Project, United States, 2018 Cycle (June 2018–May 2019). 2020.

16. Hales CM, Martin CB, Gu Q. Prevalence of Prescription Pain Medication Use Among Adults: United States, 2015-2018. NCHS Data Brief. 2020(369):1-8. PubMed PMID: 32600518.

17. Fahimi M, Kulp D. Address-Based Sampling – Alternatives for Surveys That Require Representative Samples of Households. Quirk's Marketing Research Review. 2009.

18. Levy PS, Lemeshow S. Sampling of Populations: Methods and Applications, 4th Edition. Hoboken, NJ: Wiley; 2008.

19. Siegler AJ, Brock JB, Hurt CB, Ahlschlager L, Dominguez K, Kelley CF, Jenness SM, Wilde G, Jameson SB, Bailey-Herring G, Mena LA. An Electronic Pre-Exposure Prophylaxis Initiation and Maintenance Home Care System for Nonurban Young Men Who Have Sex With Men: Protocol for a Randomized Controlled Trial. JMIR Res Protoc. 2019;8(6):e13982. doi: 10.2196/13982. PubMed PMID: 31199326; PMCID: PMC6592500.

20. Sullivan PS, Sanchez TH, Zlotorzynska M, Chandler CJ, Sineath RC, Kahle E, Tregear S. National trends in HIV pre-exposure prophylaxis awareness, willingness and use among United States men who have sex with men recruited online, 2013 through 2017. J Int AIDS Soc. 2020;23(3):e25461. doi: 10.1002/jia2.25461. PubMed PMID: 32153119; PMCID: PMC7062633.

21. Stavseth MR, Clausen T, Roislien J. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. SAGE Open Med. 2019;7:2050312118822912. doi: 10.1177/2050312118822912. PubMed PMID: 30671242; PMCID: PMC6329020.

22. Kish L. Survey Sampling. Hoboken, NJ: Wiley; 1965.

23. Cochran WG. Sampling Techniques, 3rd Edition. Hoboken, NJ: Wiley; 1977.