COMMENTARY

Monitoring epidemics: Lessons from measuring population prevalence of the coronavirus

Samuel J. Clark[a,1] and Abigail Norris Turner[b]

For the United States, data available from the Centers for Disease Control and Prevention (CDC) on 22 January 2021 (1–3) describe at least 442,000 additional deaths beyond what was expected in 2020 (Fig. 1). The bulk—roughly 336,000—can be attributed directly to COVID-19, and many of the remainder are related to the general disruption wrought by the pandemic. For a sense of scale, there were 291,000 American battle deaths in World War II (4). Adding to the catastrophic excess deaths, many of the hundreds of thousands of people who have survived COVID-19 require months to recover and suffer ongoing disabilities, and everyone is affected by myriad disruptions to daily life. The cumulative human suffering related to COVID-19 is staggering.

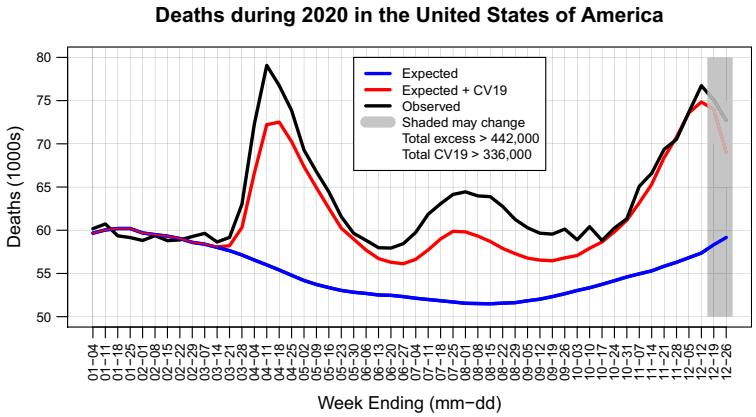

Fig. 1. All-cause deaths in the United States during 2020. Source: CDC (1–3).

Monitoring and Measuring Epidemics

Quantifying the extent and spread of an epidemic is necessary, both to understand how it works and to design and monitor interventions. This is difficult; measures must relate to the whole population instead of individual people, and they must cover reasonable time periods to describe change. The first step is knowledge of how many people are susceptible to infection, how many are infected and likely to be contagious, and at what rate new infections are appearing. In epidemiology, "prevalence" is the fraction of a population currently infected, and "incidence" is the fraction of susceptible people infected in a unit of time. Prevalence tells us the size of the infected group and, in some circumstances, gives us information about the size of the susceptible group. Incidence describes the rate of spread.

Incidence is hard to measure. Uninfected people must be followed through time and tested repeatedly to identify how many become infected and when the infections take place. Accumulating enough new infections to allow reliable, accurate measurement requires long periods of observation or observing large numbers of uninfected people. For these reasons, measures of incidence are less common.

It is usually possible to measure prevalence. The key challenge is to ensure that the indicator of prevalence truly describes the whole population, not just a convenient subgroup. Necessary data describe the size of the population and the disease status of each member. Often-used data sources include the census for population size and administrative records, facility records (e.g., hospital or clinic), and sample surveys for disease status. In a rapidly evolving epidemic, or in a situation when people may fear going to a health-care provider, administrative and facility-based records are inadequate because many potentially infected people will not appear at a facility.

Measuring Prevalence of the Coronavirus

At the time of this writing there are few published, population-representative COVID-19 prevalence studies. In a recent review Franceschi et al. list 37 (5). Two in North America are the state of Indiana (6) and the state of Connecticut (7) in the United States. In addition to these, the state of Ohio Department of Health released results from a prevalence study

[a]Department of Sociology, The Ohio State University, Columbus, OH 43210; and [b]Division of Infectious Diseases, College of Medicine, The Ohio State University, Columbus, OH 43210

Author contributions: S.J.C. and A.N.T. wrote the paper.

The authors declare no competing interest.

See companion article, "Bayesian estimation of SARS-CoV-2 prevalence in Indiana by random testing," 10.1073/pnas.2013906118.

[1]To whom correspondence may be addressed. Email: work@samclark.net.

Published February 24, 2021.

conducted in that state during July 2020 (8, 9). Two important challenges affected many of these studies.

Test Kits. Because the coronavirus that causes COVID-19 was new when these studies were conducted, reliable, well-characterized molecular test kits were sometimes unavailable or in short supply, and serological tests for antibodies raised against the virus generally performed poorly (10). Depending on the exact test kits used, this could produce inaccurate results that affect estimates of prevalence. Not knowing exactly how a test behaves requires an analysis strategy that allows for a wider variety of outcomes and produces less-certain estimates.

Selective Nonresponse. In some prevalence studies—most notably the three state-level prevalence studies in the United States (6–9)—few of those selected to participate agreed to do so, resulting in potentially consequential levels of nonparticipation. Most prevalence studies in large populations use sample survey methods. These methods have four requirements: 1) an accurate list of the whole population, the "frame"; 2) a process to select a sample that represents the whole population; 3) cooperation from many of those selected; and, crucially, 4) no systematic differences between those who participate and those who do not. In most circumstances 1 and 2 are not issues, while 3 and 4 can cause problems. The idea driving the sample survey method is that it is possible to select a comparatively small group of subjects who do not differ in systematic ways from the population; then characteristics of the sample will—on average—not deviate in systematic ways from the corresponding characteristics of the population, and measures derived from the sample can be used to describe the whole population. This correspondence is guaranteed by using a random process to select the sample, ensuring that—on average—the sample is similar to the population. The price for this efficiency is uncertainty resulting from the randomization process and the small size of the sample. When subjects selected to be in the sample refuse to participate they introduce a systematic difference between themselves and everyone else: nonparticipation. In addition to this, nonparticipators may be systematically different from participators and the population in other ways, including factors measured directly by the study—but only from those who participate. Without information from the nonparticipators this cannot be resolved. Nonparticipation can be a source of consequential bias, especially if nonparticipation is associated with study outcomes, either directly or not.

Indiana COVID-19 Prevalence Study

In the first of its kind in the United States, Menachemi et al. published a prevalence estimate of COVID-19 for the state of Indiana based on a well-designed sample survey conducted in April 2020 (6). The survey used a randomized, stratified sampling approach to construct a sample representing the whole state and the 10 Indiana State Department of Health preparedness districts. Of 15,588 people selected into the sample, 3,625 (23%) consented to participate. Participants were tested for the presence of the coronavirus and antibodies raised against the virus. Prevalence of ever-infected for the state was estimated to be 2.79% (95% CI: 2.02 to 3.70%) corresponding to 187,802 people having ever been infected among the 6.7 million residents of Indiana.

The analytical strategy ignored the extra uncertainty associated with poorly characterized tests and existing information about the prevalence of the coronavirus in subgroups defined by ethnicity, race, and age. Because the analysis was conducted in a sequence of separate steps, the final estimates may not have reflected the combined uncertainty inherent in each step. The study did not address possible selective nonresponse (see *Selective Nonresponse*).

In PNAS Yiannoutsos et al. present an updated analysis of the Indiana survey data that addresses all of those issues except selective nonresponse related to prevalence (11). The updated estimate for ever-infected is 3.58% (95% CI: 3.03 to 4.18%), or 241,044 ever-infected state residents—an increase of 0.79% in the point estimate and narrowing of the uncertainty interval from a range of 1.68 to 1.15%. Their paper presents an elegant Bayesian estimation procedure that incorporates existing knowledge of both the performance characteristics of the tests used in the survey and differences in prevalence among subgroups. The nonresponse rate was not equal across categories defined by ethnicity, race, and age. It is known that prevalence differs among those groups, so the differential nonresponse rates could create bias in the overall estimate. Because the composition of Indiana's population along those dimensions is known, it was possible to adjust the sample to match the correct composition. A post-stratification simulation approach was used to accomplish this. The authors discuss several compelling reasons why they think selective nonresponse related to prevalence is not likely to have a consequential effect on their estimate, but it remains an important unresolved issue.

Low Participation and Possible Selective Nonresponse in Sample Surveys of COVID-19 Prevalence

The other two state-level prevalence studies in the United States, Connecticut and Ohio (7–9), also report low participation rates: 7% for Connecticut and 18.5% for Ohio. We were part of the team that conducted the study in Ohio, and although we did not collect data to characterize nonresponders compared to responders, the field team reported a number of factors possibly related to nonresponse: 1) where a respondent falls on the sociopolitical spectrum, with liberal attitudes associated with participation and conservative attitudes associated with refusal; 2) sex/gender; 3) age; 4) history of previous testing; and, possibly, 5) overall health. In addition to these, issues related to mode of recruitment may affect participation: 1) modality (phone, mail, internet, etc.), 2) language spoken, 3) medical literacy (fear of the test result), and 4) employment status (not at home when the survey team visits). If any of these factors are also associated with coronavirus-related behavior and/or COVID-19 disease status, then this level of nonresponse may have consequential effects on estimates of prevalence. To explore this in the Ohio study, we conducted a sensitivity analysis to assess the effect of differences in prevalence between responders and nonresponders. We concluded that although selective nonresponse could have a large effect on our estimate, even in the worst-case scenario the substantive interpretation of the results would be unchanged—namely that prevalence was low. Neither of the existing approaches to selective nonresponse—ignoring it or sensitivity analysis—is satisfactory, and the potential for complex selective nonresponse with consequential effects on prevalence estimates is important.

Monitoring Future Epidemics

It is certain that we will experience new epidemics similar to COVID-19, and at least in the initial stages of the new epidemic people will likely avoid going to medical facilities. Activities to prepare to monitor those future epidemics fall into two categories: 1) addressing the serious challenges faced by recent coronavirus prevalence surveys and 2) developing and implementing an efficient, ongoing epidemic monitoring system that can exist in a mostly dormant state most of the time but is ready to be rapidly activated when a new epidemic emerges.

Addressing Issues with Sample Survey Prevalence Methods. The most important challenges revealed by the coronavirus prevalence surveys are 1) low participation that is potentially related to prevalence and 2) lack of standardized, well-described, fully integrated statistical methods to handle 1) low participation rates that result in damaged samples for which traditional sample weights and estimation methods cannot be used, 2) results from (potentially multiple) poorly characterized tests, and 3) very low levels of positives that may invalidate the assumptions required by commonly used frequentist estimation methods (9). It is possible for moderate levels of selective nonresponse and/or high rates of nonresponse to introduce enough bias to make estimates useless. For this reason, understanding and addressing nonresponse is extremely important and urgent. Nonresponse is complex and may be related in important ways to larger issues beyond the control of a survey team: generalized mistrust in the government or "elites," sociopolitical ideologies, or something else. Understanding this and developing strategies to address it will require participation and cooperation among several disciplines and professions—possibly including, but not limited to, sociology, political science, psychology, linguistics, English, nursing, medicine, marketing, and both pre- and post-secondary education. Both the Indiana and Ohio coronavirus prevalence teams are making progress on the statistical methodology issues (9, 11).

Creating an Ongoing Epidemic Monitoring Capability. When a future epidemic arrives, we want to have a full-coverage*, reliable, responsive epidemic monitoring system in place and ready to start working immediately. To stay feasible and responsive the system must be built around the sample survey idea with a longitudinal component to allow measurement through time and estimates of incidence. The system must incorporate solutions to nonresponse that successfully rebuild community trust and may include ongoing community outreach to maintain that trust. It will be critical to ensure that underrepresented and underserved subgroups are included. In addition to ensuring equity, those groups are likely to be among the first to be affected, and because of their overrepresentation in essential-work professions they may also be important nodes in transmission networks. All of this will require ongoing cooperation between academia and government agencies and the creation and maintenance of basic infrastructure, minimally 1) a continuously maintained sampling frame; 2) access to essential human capital including epidemiologists, biostatisticians, expert study coordinators, fieldwork supervisors, interviewers, nurses/phlebotomists, ethicists, and others; 3) predefined human resource procedures necessary to rapidly hire the field team and other expertise; and 4) predefined standard operating procedures (SOP) for study design, data management, fieldwork, data cleaning and preparation for analysis, analysis, dissemination of data and results†, and termination of the study. The new statistical methods described just above must be thoroughly tested, validated, disseminated, and "routinized" so that they can guide the design of the study and analysis of the data, without requiring time to rethink and invent de novo methods. The overall plan, SOP, and new methods must be published in appropriate outlets. To be continuously ready, a small number of key personnel must be employed permanently and conduct "fire drill" practice activities from time to time to ensure that they and a collection of potential temporary workers are ready to react quickly to a new epidemic. This all sounds expensive, and it will be, but compared to the colossal potential for loss of life, productive years of life, and economic activity, even a very large expense is a negligible and easily justifiable price to pay. The COVID-19 epidemic has demonstrated this decisively.

Acknowledgments

Jon Wakefield, David Kline, and Zehang Li provided helpful comments during the preparation of this article.

*"Full coverage" means that indicators describing the epidemic relate to the whole population and not a selective or convenient subgroup, for example people who appear at medical facilities.

†To encourage trust and ensure maximum utility of the overall effort we strongly recommend that data be shared and updated frequently and that all code/software used to manipulate and analyze the data are also shared and updated whenever changes are made.

1 CDC, Weekly counts of deaths by state and select causes, 2019-2020 (2021). https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-State-and-Select-Causes/muzy-jte6. Accessed 18 January 2021.
2 CDC, Excess deaths associated with COVID-19 (2021). https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm. Accessed 18 January 2021.
3 CDC, Provisional COVID-19 death counts by week ending date and state (2021). https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab. Accessed 18 January 2021.
4 Department of Veterans Affairs, America's wars (2021). https://www.va.gov/opa/publications/factsheets/fs_americas_wars.pdf. Accessed 17 January 2021.
5 V. B. Franceschi *et al.*, Population-based prevalence surveys during the COVID-19 pandemic: A systematic review. *Rev. Med. Virol.*, e2200 (2020).
6 N. Menachemi *et al.*, Population point prevalence of SARS-CoV-2 infection based on a statewide random sample—Indiana, April 25–29, 2020. *Morb. Mortal. Wkly. Rep.* **69**, 960 (2020).
7 S. Mahajan *et al.*, Seroprevalence of SARS-CoV-2-specific IgG antibodies among adults living in Connecticut: Post-infection prevalence (PIP) study. *Am. J. Med.* **9343**, 30909-8 (2020).
8 M. DeWine, Ohio governor Mike DeWine - COVID-19 update — October 1, 2020. https://youtu.be/oVCSOlyJ16k Start 21:00. Accessed 18 January 2021.
9 D. Kline, Z. Li, Y. Chu, J. Wakefield, S. J. Clark, Estimating seroprevalence of SARS-CoV-2 coronavirus in Ohio: A Bayesian multilevel regression and poststratification approach with multiple imperfect diagnostic tests. arXiv [Preprint] (2020). https://arxiv.org/abs/2011.09033 (Accessed 18 January 2021).
10 M. L. Bastos *et al.*, Diagnostic accuracy of serological tests for COVID-19: Systematic review and meta-analysis. *BMJ* **370**, m2516 (2020).
11 C. T. Yiannoutsos, P. K. Halverson, N. Menachemi, Bayesian estimation of SARS-CoV-2 prevalence in Indiana by random testing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2013906118 (2021).