IDSA Infectious Diseases Society of America   **hivma** hiv medicine association   OXFORD

Improving Surveillance Estimates of Coronavirus Disease 2019 (COVID-19) Incidence in the United States

Eli S. Rosenberg[1,2] and Heather M. Bradley[3]

[1]Department of Epidemiology and Biostatistics, University at Albany School of Public Health, State University of New York, Rensselaer, New York, USA, [2]Center for Collaborative HIV Research in Practice and Policy, University at Albany School of Public Health, State University of New York, Rensselaer, New York, USA, and [3]Department of Population Health Sciences, Georgia State University School of Public Health, Atlanta, Georgia, USA

(See the Major Articles by Basavaraju et al on pages e1004–9 and by Reed et al on pages e1010–7).

Keywords. COVID-19; cumulative incidence; surveillance; epidemiology; SARS-CoV-2.

Assessment of the cumulative burden of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection across demographic groups and geography informs disease-control policies, public health prevention efforts, and risk communication to the public. Estimating burden of disease is challenging, however, because reported diagnoses represent a fraction of total infections, as a function of symptomatic status along with patient and provider behaviors, compounded in this year's pandemic by time-varying conditions that impacted test-seeking and receipt. In this issue of *Clinical Infectious Diseases,* Reese et al [1] adapt a multiplier method that has previously been used by the Centers for Disease Control and Prevention (CDC) to convert the reported 2.1% symptomatic US diagnoses through September 2020 into 16.2%, or 53 million, persons estimated to have been infected. This estimate is a staggering assessment of both how many Americans have been touched by this infection and how many remain vulnerable in the months ahead, necessitating ongoing widespread prevention efforts.

To date, 2 study designs have been employed to measure cumulative incidence in US jurisdictions. Studies of clinical laboratory residual serum use new serological testing for SARS-CoV-2 antibodies to understand history of infection and can be combined with external data to estimate rates of diagnosis, hospitalization, and death [2, 3]. These studies use plentiful, available specimens to understand burden of infection across jurisdictions and time but have limited variables for describing cumulative incidence by demographic features and for adjusting estimates for biases resulting from passive sampling of persons attending medical care during a pandemic. Population-based serological studies that collect specimens from participants through combinations of in-person, drive-through, and at-home modalities offer the opportunity for more extensive survey data collection and representative sampling, but come with the disadvantages of cost and threatened validity due to the complexities of recruiting large samples in a time of misinformation and societal closure [4–6]. Data from both designs require adjustment for serological test characteristics, including that a portion (~10%) of infected persons never develop detectable antibodies. Further, waning antibody detection, particularly for persons with milder or absence of symptoms, may complicate antibody positivity interpretation as time progresses, necessitating additional time-dependent corrections by symptomology [7, 8].

Reese et al have developed a third approach, not yet taken for US coronavirus disease 2019 (COVID-19) surveillance, in the use of a multiplier model, an approach that builds on the disease pyramid concept, wherein total infections are estimated from the subset of diagnosed infections that are visible as the pyramid's tip. Similar to models developed for influenza and viral hepatitis surveillance, this process involves applying serial correction factors to an underascertained, and potentially otherwise biased, disease indicator universally available in surveillance data [9, 10]. In the case of COVID-19, the authors begin with mandatorily reported symptomatic COVID-19 cases and apply 4 levels of successive corrections for the probabilities of detection given testing (test sensitivity), test ordering given clinical presentation with symptoms, care seeking given symptoms, and symptom development given infection. They stratify parts of this process by subgroups defined by demography, geography, and time, which facilitates the following: (1) control for heterogeneity in these probabilities across subgroups by varying the correction factors applied and (2) stratification of results to display

the differential burden of disease across subgroups. We discuss below the implications of this approach as performed as well as enhancements that could be afforded by improved surveillance data.

A clear advantage of the multiplier method is because no new data are collected beyond routine surveillance, estimates can be frequently updated, allowing for a more real-time understanding of COVID-19 burden. Given the public health urgency of quantifying COVID-19 infection and hospitalization burden throughout successive phases of the US epidemic since early 2020, and the prior existence of this method and data inputs, we lament that CDC could have far earlier and routinely made such estimates available to inform the public and public health community. As with all models, this approach's robustness may be evaluated by the external validity of its estimates, made challenging by the lack of external estimates of total US infections, and by evaluating the quality of its approach and inputs.

Enhancements in input surveillance data would increase the robustness of this method for estimating COVID-19 burden of disease. To produce optimally informative results, more granular geographic and demographic data are needed for both parameter estimates and to create the strata within which they are applied. The authors apply parameters B (the extent to which symptomatic patients seek care) and C (the proportion of care-seeking patients who are tested for COVID-19) within Department of Health and Human Services (HHS) census regions, but the parameter estimates themselves are not available by region or other characteristics. These parameters likely vary substantially by geography and population characteristics, including race/ethnicity and other social determinants of health.

Race/ethnicity is a particularly important characteristic that should be accounted for in the results, because racial and ethnic minority populations have the highest rates of COVID-19 diagnosis per population and are also geographically

concentrated in the United States [11]. For this to be feasible, more complete race/ethnicity data on case reports are required. Despite many calls for more complete case report data on race/ethnicity, including from CDC, only 52% of reported cases currently have race/ethnicity specified [11]. Varying the parameter estimates applied within racial/ethnic strata would also require care-seeking and testing behaviors data by race/ethnicity. The extent to which participation in voluntary surveillance such as Flu Near You and COVID Near You varies by characteristics of underlying populations by geography is unclear. A national surveillance system that routinely collects information on COVID-19 prevention and testing behaviors, symptoms status, and vaccine readiness from a representative sample, with the ability to produce estimates by state and population characteristics, is needed to optimally parameterize this multiplier method and would serve other important functions. As applied in the present model, use of the same parameters for patient care-seeking and provider testing behaviors across geographic areas may underestimate burden of disease in geographic areas with larger minority populations. This is particularly problematic for future resource-allocation decisions, including for vaccine distribution, which may be made based on estimated infections.

Robust estimates from the method employed by Reese et al also rely on jurisdiction-wide standardization of surveillance practices in 2 areas: de-duplication and merging of case reports and symptoms ascertainment. In terms of de-duplication, insufficient merging of case reports by person will result in overestimation of burden of disease. Further, because of latency between COVID-19 diagnosis and hospitalization, unless hospital tests are merged to the original case data many persons ultimately hospitalized will be missing this information on their case report, leading to underestimation of hospitalization rates. Uniform practices for collecting data on

symptoms are also required. In the present model, asymptomatic cases were excluded from diagnoses, and an asymptomatic fraction was applied to the total number of estimated asymptomatic cases across the United States. The underlying, but unstated, assumption is that missing symptom data reflect asymptomatic infection and the quality/completeness of these data is relatively uniform across jurisdictions and, ultimately, HHS region.

As advancements are made in COVID-19 data quality and surveillance data systems, 1 interim solution for burden-of-disease estimates resulting from a multiplier approach is to stratify the method's steps and estimates by jurisdiction based on completeness and quality of surveillance data. The CDC has previously performed such stratified reporting for human immunodeficiency virus (HIV) clinical outcomes and drug overdose deaths [12, 13]. Data quality related to completeness of race/ethnicity and symptoms status, as well as procedures used to merge patient information across case reports, should be considered [14].

Reese et al have provided us with an important set of COVID-19 burden-of-disease estimates that can continue to be improved over time as the quality and completeness of surveillance data also improve. While their estimate of 53 million infections is distressing, it indicates that the United States is far from achieving herd immunity even with minimal assumptions about waning immunity. Swift and equitable vaccine distribution will be critical to curbing the US epidemic, and continual improvements to burden-of-disease estimates and how they vary by person, time, and place will be needed to optimally allocate resources toward such efforts and monitor success.

Notes

the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Reese H, Iuliano AD, Patel NN, et al. Estimated incidence of COVID-19 illness and hospitalization—United States, February–September, 2020. Clin Infect Dis **2021**; 72(12):e1010–7.
2. Bajema KL, Wiegand RE, Cuffe K, et al. Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. JAMA Intern Med **2020**. doi: 10.1001/jamainternmed.2020.7976.
3. Stadlbauer D, Tan J, Jiang K, et al. Repeated cross-sectional sero-monitoring of SARS-CoV-2 in New York City. Nature **2020**. doi: 10.1038/s41586-020-2912-6.
4. Rosenberg ES, Tesoriero JM, Rosenthal EM, et al. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. Ann Epidemiol **2020**; 48:23–9, e4.
5. Sood N, Simon P, Ebner P, et al. Seroprevalence of SARS-CoV-2–specific antibodies among adults in Los Angeles County, California, on April 10–11, 2020. JAMA **2020**; 323:2425–7.
6. Siegler AJ, Sullivan PS, Sanchez T, et al. Protocol for a national probability survey using home specimen collection methods to assess prevalence and incidence of SARS-CoV-2 infection and antibody response. Ann Epidemiol **2020**; 49:50–60.
7. Fenwick C, Croxatto A, Coste AT, et al. Changes in SARS-CoV-2 spike versus nucleoprotein antibody responses impact the estimates of infections in population-based seroprevalence studies. J Virol **2020**. doi: 10.1128/jvi.01828-20.
8. Self WH, Tenforde MW, Stubblefield WB, et al. Decline in SARS-CoV-2 antibodies after mild infection among frontline health care personnel in a multistate hospital network—12 states, April-August 2020. MMWR Morb Mortal Wkly Rep **2020**; 69:1762–6.
9. Reed C, Angulo FJ, Swerdlow DL, et al. Estimates of the prevalence of pandemic (H1N1) 2009, United States, April-July 2009. Emerg Infect Dis **2009**; 15:2004–7.
10. Klevens RM, Liu S, Roberts H, Jiles RB, Holmberg SD. Estimating acute viral hepatitis infections from nationally reported cases. Am J Public Health **2014**; 104:482–7.
11. Centers for Disease Control and Prevention. CDC COVID data tracker: demographic trends of COVID-19 cases and deaths in the US reported to CDC. Available at: https://covid.cdc.gov/covid-data-tracker/#demographics. Accessed 29 November 2020.
12. Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data: United States and 6 dependent areas, 2018. Vol. 25. **2020**. Available at: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-25-2.pdf. Accessed 29 November 2020.
13. Scholl L, Seth P, Kariisa M, Wilson N, Baldwin G. Drug and opioid-involved overdose deaths—United States, 2013–2017. MMWR Morb Mortal Wkly Rep **2018**; 67:1419–27.
14. Vital Strategies. Tracking COVID-19 in the United States: progress and opportunities. Available at: https://preventepidemics.org/wp-content/uploads/2020/11/Tracking-COVID-in-the-US-Progress-Opportunities.pdf. Accessed 29 November 2020.