medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

COVID-19 Antibody Seroprevalence in Santa Clara County, California

Eran Bendavid[1], Bianca Mulaney[2], Neeraj Sood[3], Soleil Shah[2], Emilia Ling[2], Rebecca Bromley-Dulfano[2], Cara Lai[2], Zoe Weissberg[2], Rodrigo Saavedra-Walker[4], Jim Tedrow[5], Dona Tversky[6], Andrew Bogan[7], Thomas Kupiec[8], Daniel Eichner[9], Ribhav Gupta[10], John P.A. Ioannidis[1,10], Jay Bhattacharya[1]

April 11, 2020

[1] Department of Medicine, Stanford University School of Medicine, Stanford CA
[2] Stanford University School of Medicine, Stanford CA
[3] Sol Price School of Public Policy, University of Southern California, Los Angeles CA
[4] Health Education is Power, Inc., Palo Alto CA
[5] The Compliance Resource Group, Inc., Oklahoma City OK
[6] Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford CA
[7] Bogan Associates, LLC, Palo Alto CA
[8] ARL BioPharma, Inc. , Oklahoma City OK
[9] Sports Medicine Research and Testing Laboratory, Salt Lake City UT
[10] Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford CA

Corresponding author: Eran Bendavid MD, Department of Medicine, Stanford University, Stanford CA 94305. email: ebd@stanford.edu.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Background
Addressing COVID-19 is a pressing health and social concern. To date, many epidemic projections and policies addressing COVID-19 have been designed without seroprevalence data to inform epidemic parameters. We measured the seroprevalence of antibodies to SARS-CoV-2 in Santa Clara County.

Methods
On 4/3-4/4, 2020, we tested county residents for antibodies to SARS-CoV-2 using a lateral flow immunoassay. Participants were recruited using Facebook ads targeting a representative sample of the county by demographic and geographic characteristics. We report the prevalence of antibodies to SARS-CoV-2 in a sample of 3,330 people, adjusting for zip code, sex, and race/ethnicity. We also adjust for test performance characteristics using 3 different estimates: (i) the test manufacturer's data, (ii) a sample of 37 positive and 30 negative controls tested at Stanford, and (iii) a combination of both.

Results
The unadjusted prevalence of antibodies to SARS-CoV-2 in Santa Clara County was 1.5% (exact binomial 95CI 1.11-1.97%), and the population-weighted prevalence was 2.81% (95CI 2.24-3.37%). Under the three scenarios for test performance characteristics, the population prevalence of COVID-19 in Santa Clara ranged from 2.49% (95CI 1.80-3.17%) to 4.16% (2.58-5.70%). These prevalence estimates represent a range between 48,000 and 81,000 people infected in Santa Clara County by early April, 50-85-fold more than the number of confirmed cases.

Conclusions
The population prevalence of SARS-CoV-2 antibodies in Santa Clara County implies that the infection is much more widespread than indicated by the number of confirmed cases. Population prevalence estimates can now be used to calibrate epidemic and mortality projections.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

The first two cases of COVID-19 in Santa Clara County, California were identified in returning travelers on January 31 and on February 1, 2020, and the third case was identified four weeks later on February 27, 2020.[1] In the following month, nearly 1,000 additional cases were identified in Santa Clara County, showing a pattern of rapid case increase reflective of community transmission as well as the scaling up of SARS-CoV-2 viral testing that was common across many communities globally. In some countries, the rapid increase in COVID-19 case counts and hospitalizations has overwhelmed health systems and led to large reductions in social and economic activities. The measures adopted to slow the spread of COVID-19 were justified by projected estimates of health care system capacity and case fatality rate. These projections suggested that, in the absence of strict measures to reduce transmission, the COVID-19 pandemic would overwhelm existing hospital bed and ICU capacity throughout the United States and lead to over 2 million deaths.[2]

Measuring fatality rates and projecting the number of deaths depend on estimates of the total number of infections. To date, in the absence of seroprevalence surveys, estimates of the fatality rate have relied on the number of confirmed cases multiplied by an estimated factor representing unknown or asymptomatic cases to arrive at the number of infections.[3–6] However, the magnitude of that factor is highly uncertain. Because the implications of infection fatality rate and projected deaths are large, the extent of COVID-19 infection under-ascertainment (the multiplier used to arrive from cases to infections) has been a topic of great interest and provided estimates of the number of infections about 1-6-fold higher than the number of cases.[7–10] The extent of infection under-ascertainment has been difficult to assess because of three biasing processes: (i) cases have been diagnosed with PCR-based tests, which do not provide information about resolved infections; (ii) the majority of cases tested early in the course of the epidemic have been acutely ill and highly symptomatic, while most asymptomatic or mildly symptomatic individuals have not been tested; and (iii) PCR-based testing rates have been highly variable across contexts and over time, leading to noisy relationships between the number of cases and infections. If, in the absence of interventions, the epidemic's early doubling time is estimated to be four days[6,11,12], then by February 27th, 2020, when the third case was identified in Santa Clara County, the county may have already had 256 infections.

At the time of this study, Santa Clara County had the largest number of confirmed cases of any county in Northern California (1,094). The county also had several of the earliest known cases of COVID-19 in the state - including one of the first presumed cases of community-acquired disease - making it an especially appropriate location to test a population-level sample for the presence of active and *past* infections.

On April 3rd and 4th, 2020 we conducted a survey of residents of Santa Clara County to measure the seroprevalence of antibodies to SARS-CoV-2 and better approximate the number of infections. Our goal is to provide new and well-measured data for informing epidemic models, projections, and public policy decisions.

Methods

We conducted serologic testing for SARS-CoV-2 antibodies in 3,330 adults and children in Santa Clara County using capillary blood draws and a lateral flow immunoassay. In this section we describe our sampling and recruitment approaches, specimen collection methods, antibody testing procedure, test kit validation, and statistical methods. Our protocol was informed by a World Health Organization protocol for population-level COVID-19 antibody testing.[13] We conducted our study with the cooperation of the

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Santa Clara County Department of Public Health. The IRB at Stanford University approved the study prior to recruitment.

Study Participants and Sample Recruitment

We recruited participants by placing targeted advertisements on Facebook aimed at residents of Santa Clara County. We used Facebook to quickly reach a large number of county residents and because it allows for granular targeting by zip code and sociodemographic characteristics.[14] We used a combination of two targeting strategies: ads aimed at a representative population of the county by zip code, and specially targeted ads to balance our sample for under-represented zip codes. In addition, we capped registrations from overrepresented areas. Individuals who clicked on the advertisement were directed to a survey hosted by the Stanford REDcap platform, which provided information about the study.[15] The survey asked for six data elements: zip code of residence, age, sex, race/ethnicity, underlying co-morbidities, and prior clinical symptoms. Over 24 hours, we registered 3,285 adults, and each adult was allowed to bring one child from the same household with them (889 children registered).

Specimen Collection and Testing Methods

We established drive-through test sites in three locations spaced across Santa Clara County: two county parks in Los Gatos and San Jose, and a church in Mountain View. Only individuals with a participant ID were allowed into the testing area. Verbal informed consent was obtained to minimize participant and staff exposure. With participants in their vehicles, sample collectors in personal protective equipment drew 50-200μL of capillary blood into an EDTA-coated microtainer. Tubes were barcoded and linked with the participant ID. Samples were couriered from the collection sites to a test reading facility with steady lighting and climate conditions. Technicians drew whole blood up to a fill line on the manufacturer's pipette and placed it in the test kit well, followed by a buffer. Test kits were read 12-20 minutes after the buffer was placed. Technicians barcoded tests to match sample barcodes and documented all test results.

Test Kit Performance

The manufacturer's performance characteristics were available prior to the study (using 85 confirmed positive and 371 confirmed negative samples). We conducted additional testing to assess the kit performance using local specimens. We tested the kits using sera from 37 RT-PCR-positive patients at Stanford Hospital that were also IgG and/or IgM-positive on a locally developed ELISA assay. We also tested the kits on 30 pre-COVID samples from Stanford Hospital to derive an independent measure of specificity. Our procedure for using these data is detailed below.

Statistical Analysis

Our estimation of the population prevalence of COVID-19 proceeded in three steps. First, we reported the raw frequencies of positive tests as a proportion of the final sample size. Second, we re-weighted our sample by zip code, sex, and race/ethnicity (non-Hispanic White, Asian, Hispanic, and other). We chose these three adjustors because they contributed to the largest imbalance in our sample, and because including additional adjustors would result in small-N bins. Our weights were the zip-sex-race proportion in Santa Clara County divided by the zip-sex-race proportion in our sample, for each zip-sex-race combination in the county and in the sample.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

$$weight_{zsr} = \frac{\frac{N^{c}_{zsr}}{N^{C}_{total}}}{\frac{N^{S}_{zsr}}{N^{S}_{total}}}$$

Where N^{C} represents county counts, N^{S} represents sample counts, and the subscripts zsr identifies the unique zip-sex-race groups. These weights were then applied to the entire sample. To provide a concrete example, suppose the populations of two zip codes (A and B) include 10,000 men and 10,000 women. Our sample included 250 men and 500 women from zip A, and 750 men and 1500 women from zip B. This is exemplary of the imbalance in our sample. Applying the formula above, we get a weight of 3 for men in zip A, 1.5 for women in zip A, 1 for men in zip B, and 0.5 for women in zip B.

Third, we adjusted the prevalence for test sensitivity and specificity. Because SARS-CoV-2 lateral flow assays are new, we applied three scenarios of test kit sensitivity and specificity. The first scenario uses the manufacturer's validation data (S1). The second scenario uses sensitivity and specificity from a sample of 37 known positive (RT-PCR-positive and IgG or IgM positive on a locally-developed ELISA) and 30 known pre-COVID negatives tested on the kit at Stanford (S2). The third scenario combines the two collections of samples (manufacturer and local sample) as a single pooled sample (S3). We use the delta method to estimate standard errors for the population prevalence, which accounts for sampling error and propagates the uncertainty in the sensitivity and specificity in each scenario. A more detailed version of the formulas we use in our calculations is available in the Appendix to this paper.

Results

The test kit used in this study (Premier Biotech, Minneapolis, MN) was tested in a Stanford laboratory prior to field deployment. Among 37 samples of known PCR-positive COVID-19 patients with positive IgG or IgM detected on a locally-developed ELISA test, 25 were kit-positive. A sample of 30 pre-COVID samples from hip surgery patients were also tested, and all 30 were negative. The manufacturer's test characteristics relied on samples from clinically confirmed COVID-19 patients as positive gold standard and pre-COVID sera for negative gold standard. Among 75 samples of clinically confirmed COVID-19 patients with positive IgG, 75 were kit-positive, and among 85 samples with positive IgM, 78 were kit-positive. Among 371 pre-COVID samples, 369 were negative. Our estimates of sensitivity based on the manufacturer's and locally tested data were 91.8% (using the lower estimate based on IgM, 95 CI 83.8-96.6%) and 67.6% (95 CI 50.2-82.0%), respectively. Similarly, our estimates of specificity are 99.5% (95 CI 98.1-99.9%) and 100% (95 CI 90.5-100%). A combination of both data sources provides us with a combined sensitivity of 80.3% (95 CI 72.1-87.0%) and a specificity of 99.5% (95 CI 98.3-99.9%).

Our study included 3,439 individuals that registered for the study and arrived at testing sites. We excluded observations of individuals who could not be tested (e.g. unable to obtain blood or blood clotted, N=49), whose test results could not be matched to their personal data (e.g. if an incorrect participant ID was recorded onsite, N=30) , who did not reside in Santa Clara County (N=29), and who had invalid test results (no Control band, N=1). This yielded an analytic sample of 3,330 individuals with complete records including survey registration, attendance at a test site for specimen collection, and lab results (Figure 1). The sample distribution meaningfully deviated from that of the Santa Clara County population along several dimensions: sex (63% in sample was female, 50% in county); race (8% of the sample was Hispanic, 26% in the county; 19% of the sample was Asian, 28% in the county); and zip

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

distribution (median participant density per 1,000 population 1.6, IQR 0.9-3.6). Table 1 includes demographic characteristics of our unadjusted sample, population-adjusted sample, and Santa Clara County.[16] Figure 2 shows the geographical zip code distribution of study participants in the county (counts and density per 1,000 population).

The total number of positive cases by either IgG or IgM in our unadjusted sample was 50, a crude prevalence rate of 1.50% (exact binomial 95% CI 1.11-1.97%). After weighting our sample to match Santa Clara County by zip, race, and sex, the prevalence was 2.81% (95% CI 2.24-3.37 without clustering the standard errors for members of the same household, and 1.45-4.16 with clustering). We further improved our estimation using the available data on test kit sensitivity and specificity, using the three scenarios noted above. The estimated prevalence was 2.49% (95CI 1.80%-3.17%) under the S1 scenario, 4.16% (95CI 2.58%-5.70%) under the S2 scenario, and 2.75% (95CI 2.01%-3.49%) under the S3 scenario. Notably, the uncertainty bounds around each of these population prevalence estimates propagates the uncertainty in each of the three component parameters: sample prevalence, test sensitivity, and test specificity.

Discussion

After adjusting for population and test performance characteristics, we estimate that the seroprevalence of antibodies to SARS-CoV-2 in Santa Clara County is between 2.49% and 4.16%, with uncertainty bounds ranging from 1.80% (lower uncertainty bound of the lowest estimate), up to 5.70% (upper uncertainty bound of the highest estimate). Test performance characteristics are the most critical driver of this range, with lower estimates associated with data suggesting the test has a high sensitivity for identifying SARS-CoV-2, and higher estimates resulting from data suggesting over 30% of positive cases are missed by the test.

These results represent the first large-scale community-based prevalence study in a major US county completed during a rapidly changing pandemic, and with newly available test kits. We consider our estimate to represent the best available current evidence, but recognize that new information, especially about the test kit performance, could result in updated estimates. For example, if new estimates indicate test specificity to be less than 97.9%, our SARS-CoV-2 prevalence estimate would change from 2.8% to less than 1%, and the lower uncertainty bound of our estimate would include zero. On the other hand, lower sensitivity, which has been raised as a concern with point-of-care test kits, would imply that the population prevalence would be even higher. New information on test kit performance and population should be incorporated as more testing is done and we plan to revise our estimates accordingly.

The most important implication of these findings is that the number of infections is much greater than the reported number of cases. Our data imply that, by April 1 (three days prior to the end of our survey) between 48,000 and 81,000 people had been infected in Santa Clara County. The reported number of confirmed positive cases in the county on April 1 was 956, 50-85-fold lower than the number of infections predicted by this study.[17] The infection to case ratio, also referred to as an under-ascertainment rate, of at least 50, is meaningfully higher than current estimates.[10,18] This ascertainment rate is a fundamental parameter of many projection and epidemiologic models, and is used as a calibration target for understanding epidemic stage and calculating fatality rates.[19,20] The under-ascertainment for COVID-19 is likely a function of reliance on PCR for case identification which misses convalescent cases, early spread in the absence of systematic testing, and asymptomatic or lightly symptomatic infections that go undetected.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The under-ascertainment of infections is central for better estimation of the fatality rate from COVID-19. Many estimates of fatality rate use a ratio of deaths to lagged cases (because of duration from case confirmation to death), with an infections-to-cases ratio in the 1-5-fold range as an estimate of under-ascertainment.[3,4,21] Our study suggests that adjustments for under-ascertainment may need to be much higher.

We can use our prevalence estimates to approximate the infection fatality rate from COVID-19 in Santa Clara County. As of April 10, 2020, 50 people have died of COVID-19 in the County, with an average increase of 6% daily in the number of deaths. If our estimates of 48,000-81,000 infections represent the cumulative total on April 1, and we project deaths to April 22 (a 3 week lag from time of infection to death[22]), we estimate about 100 deaths in the county. A hundred deaths out of 48,000-81,000 infections corresponds to an infection fatality rate of 0.12-0.2%. If antibodies take longer than 3 days to appear, if the average duration from case identification to death is less than 3 weeks, or if the epidemic wave has peaked and growth in deaths is less than 6% daily, then the infection fatality rate would be lower. These straightforward estimations of infection fatality rate fail to account for age structure and changing treatment approaches to COVID-19. Nevertheless, our prevalence estimates can be used to update existing fatality rates given the large upwards revision of under-ascertainment.

While our prevalence estimates of 2.49% to 4.16% are representative of the situation in Santa Clara County as of April 4, other areas are likely to have different seroprevalence estimates based on effective contact rates in the community, social distancing policies to date, and relative disease progression. Our prevalence estimate also suggests that, at this time, a large fraction of the population remains unexposed in Santa Clara County. Repeated serologic testing in different geographies, spaced a few weeks apart, could establish extent of infection over time.

This study had several limitations. First, our sampling strategy selected for members of Santa Clara County with access to Facebook and a car to attend drive-through testing sites. This resulted in an over-representation of white women between the ages of 19 and 64, and an under-representation of Hispanic and Asian populations, relative to our community. Those imbalances were partly addressed by weighting our sample population by zip code, race, and sex to match the county. We did not account for age imbalance in our sample, and could not ascertain representativeness of SARS-CoV-2 antibodies in homeless populations. Other biases, such as bias favoring individuals in good health capable of attending our testing sites, or bias favoring those with prior COVID-like illnesses seeking antibody confirmation are also possible. The overall effect of such biases is hard to ascertain.

The Premier Biotech serology test used in this study has not been approved by the FDA by the time of the study, and validation studies for this assay are ongoing. We used existing test performance data to establish a range of sensitivity and specificity, including reliable but small-size data sourced at Stanford. Test sensitivity varied between the manufacturer's data and the local data. It is possible that asymptomatic or mildly symptomatic individuals may generate only low-titer antibodies, and that sensitivity may be even lower if there are many such cases.[23] Additional validation of the assays used could improve our estimates and those of ongoing serosurveys.

Several teams worldwide have started testing population samples for SARS CoV-2 antibodies, with preliminary findings consistent with a large under-ascertainment of SARS CoV-2 infections. Reports from the town of Robbio, Italy, where the entire population was tested, suggest at least 10% seropositivity;[24] and data from Gangelt, a highly affected area in Germany,[25] point to 14% seropositivity.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A recent effort to test the town of Telluride, Colorado is underway, and interim results suggest a prevalence just under 2%.[26] Our data from Santa Clara county suggest higher spread of the infection than Telluride but lower than some areas in Europe.

We conclude that based on seroprevalence sampling of a large regional population, the prevalence of SARS-CoV-2 antibodies in Santa Clara County was between 2.49% and 4.16% by early April. While this prevalence may be far smaller than the theoretical final size of the epidemic,[27] it suggests that the number of infections is 50-85-fold larger than the number of cases currently detected in Santa Clara County. These new data should allow for better modeling of this pandemic and its progression under various scenarios of non-pharmaceutical interventions. While our study was limited to Santa Clara County, it demonstrates the feasibility of seroprevalence surveys of population samples now, and in the future, to inform our understanding of this pandemic's progression, project estimates of community vulnerability, and monitor infection fatality rates in different populations over time. It is also an important tool for reducing uncertainty about the state of the epidemic, which may have important public benefits.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Novel Coronavirus Press Archives - Public Health Department - County of Santa Clara [Internet]. [accessed 2020 Apr 12];Available from: https://www.sccgov.org/sites/phd/DiseaseInformation/novel-coronavirus/Pages/archives.aspx

2. Report 12 - The global impact of COVID-19 and strategies for mitigation and suppression [Internet]. Imperial College London. [accessed 2020 Apr 7];Available from: http://www.imperial.ac.uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-12-global-impact-covid-19/

3. Spychalski P, Błażyńska-Spychalska A, Kobiela J. Estimating case fatality rates of COVID-19. The Lancet Infectious Diseases [Internet] 2020 [accessed 2020 Apr 9];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30246-2/abstract

4. Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA [Internet] 2020 [accessed 2020 Apr 9];Available from: https://jamanetwork.com/journals/jama/fullarticle/2763667

5. Li L-Q, Huang T, Wang Y-Q, et al. COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. J Med Virol 2020;

6. Wu JT, Leung K, Bushman M, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nature Medicine [Internet] 2020 [accessed 2020 Apr 8];1–5. Available from: https://www.nature.com/articles/s41591-020-0822-7

7. Kimball A. Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility — King County, Washington, March 2020. MMWR Morb Mortal Wkly Rep [Internet] 2020 [accessed 2020 Apr 10];69. Available from: https://www.cdc.gov/mmwr/volumes/69/wr/mm6913e1.htm

8. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance [Internet] 2020 [accessed 2020 Apr 10];25(10):2000180. Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.10.2000180

9. Nishiura H, Kobayashi T, Suzuki A, et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). Int J Infect Dis 2020;

10. Li R, Pei S, Chen B, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science [Internet] 2020 [accessed 2020 Apr 10];Available from: https://science.sciencemag.org/content/early/2020/03/24/science.abb3221

11. Pellis L, Scarabel F, Stage HB, et al. Challenges in control of Covid-19: short doubling time and long delay to effect of interventions. arXiv:200400117 [q-bio] [Internet] 2020 [accessed 2020 Apr 8];Available from: http://arxiv.org/abs/2004.00117

12. Muniz-Rodriguez K, Chowell G, Cheung C-H, et al. Epidemic doubling time of the COVID-19 epidemic by Chinese province. medRxiv [Internet] 2020;2020.02.05.20020750. Available from: http://medrxiv.org/content/early/2020/02/28/2020.02.05.20020750.abstract

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

13. Population-based age-stratified seroepidemiological investigation protocol for COVID-19 virus infection [Internet]. [accessed 2020 Apr 8];Available from: https://www.who.int/publications-detail/population-based-age-stratified-seroepidemiological-investigation-protocol-for-covid-19-virus-infection

14. Shaver LG, Khawer A, Yi Y, et al. Using Facebook Advertising to Recruit Representative Samples: Feasibility Assessment of a Cross-Sectional Survey. J Med Internet Res 2019;21(8):e14021.

15. REDCap: Resources operated by Stanford Research IT [Internet]. [accessed 2020 Apr 7];Available from: https://med.stanford.edu/researchit/resources.html#citations

16. Bureau UC. American Community Survey (ACS) [Internet]. The United States Census Bureau. [accessed 2020 Apr 10];Available from: https://www.census.gov/programs-surveys/acs

17. County of Santa Clara Public Health Department - Coronavirus (COVID-19) Data Dashboard [Internet]. [accessed 2020 Apr 11];Available from: https://www.sccgov.org/sites/phd/DiseaseInformation/novel-coronavirus/Pages/dashboard.aspx#cases

18. Salomon JA. Defining high-value information for COVID-19 decision-making. medRxiv [Internet] 2020;2020.04.06.20052506. Available from: http://medrxiv.org/content/early/2020/04/08/2020.04.06.20052506.abstract

19. Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of COVID-19 disease. medRxiv [Internet] 2020 [accessed 2020 Apr 10];2020.03.09.20033357. Available from: https://www.medrxiv.org/content/10.1101/2020.03.09.20033357v1

20. IHME | COVID-19 Projections [Internet]. Institute for Health Metrics and Evaluation. [accessed 2020 Apr 10];Available from: https://covid19.healthdata.org/

21. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. The Lancet Infectious Diseases [Internet] 2020 [accessed 2020 Apr 10];Available from: http://www.sciencedirect.com/science/article/pii/S147330992030195X

22. Bhatraju PK, Ghassemieh BJ, Nichols M, et al. Covid-19 in Critically Ill Patients in the Seattle Region — Case Series. New England Journal of Medicine [Internet] 2020 [accessed 2020 Apr 10];0(0):null. Available from: https://www.nejm.org/doi/full/10.1056/NEJMoa2004500

23. Wu F, Wang A, Liu M, et al. Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and their implications. medRxiv [Internet] 2020;2020.03.30.20047365. Available from: http://medrxiv.org/content/early/2020/04/06/2020.03.30.20047365.abstract

24. What the first data of Robbio reveal, the only Italian country to take immunity tests on all citizens [Internet]. Business Insider Italia. 2020 [accessed 2020 Apr 10];Available from: https://it.businessinsider.com/esclusiva-cosa-rivelano-i-primi-test-di-robbio-primo-paese-italiano-a-fare-i-test-sullimmunita-a-tutti-i-cittadini/

25. Blood tests show 14% of people are now immune to covid-19 in one town in Germany [Internet]. MIT Technology Review. [accessed 2020 Apr 10];Available from: https://www.technologyreview.com/2020/04/09/999015/blood-tests-show-15-of-people-are-now-immune-to-covid-19-in-one-town-in-germany/

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

26. San Miguel County suspends, then recommits to its universal coronavirus testing experiment over 3-day span [Internet]. The Colorado Sun. [accessed 2020 Apr 10];Available from: https://coloradosun.com/2020/04/09/san-miguel-county-telluride-universal-coronavirus-testing/

27. Coronavirus (COVID-19): Press Conference with Marc Lipsitch, 03/04/20 [Internet]. News. 2020 [accessed 2020 Apr 10];Available from: https://www.hsph.harvard.edu/news/features/coronavirus-covid-19-press-conference-with-marc-lipsitch-03-04-20/

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figures

Figure 1

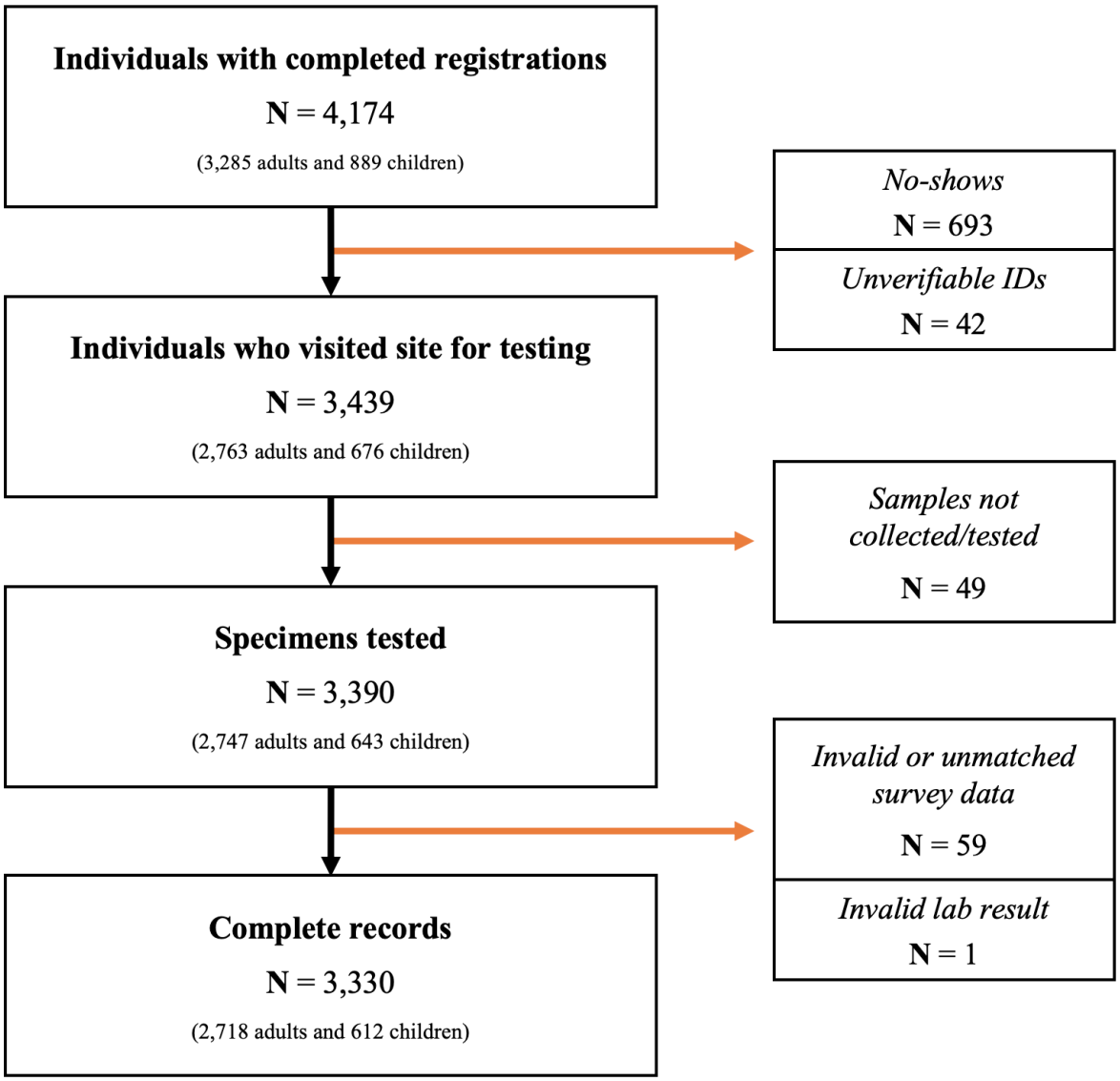
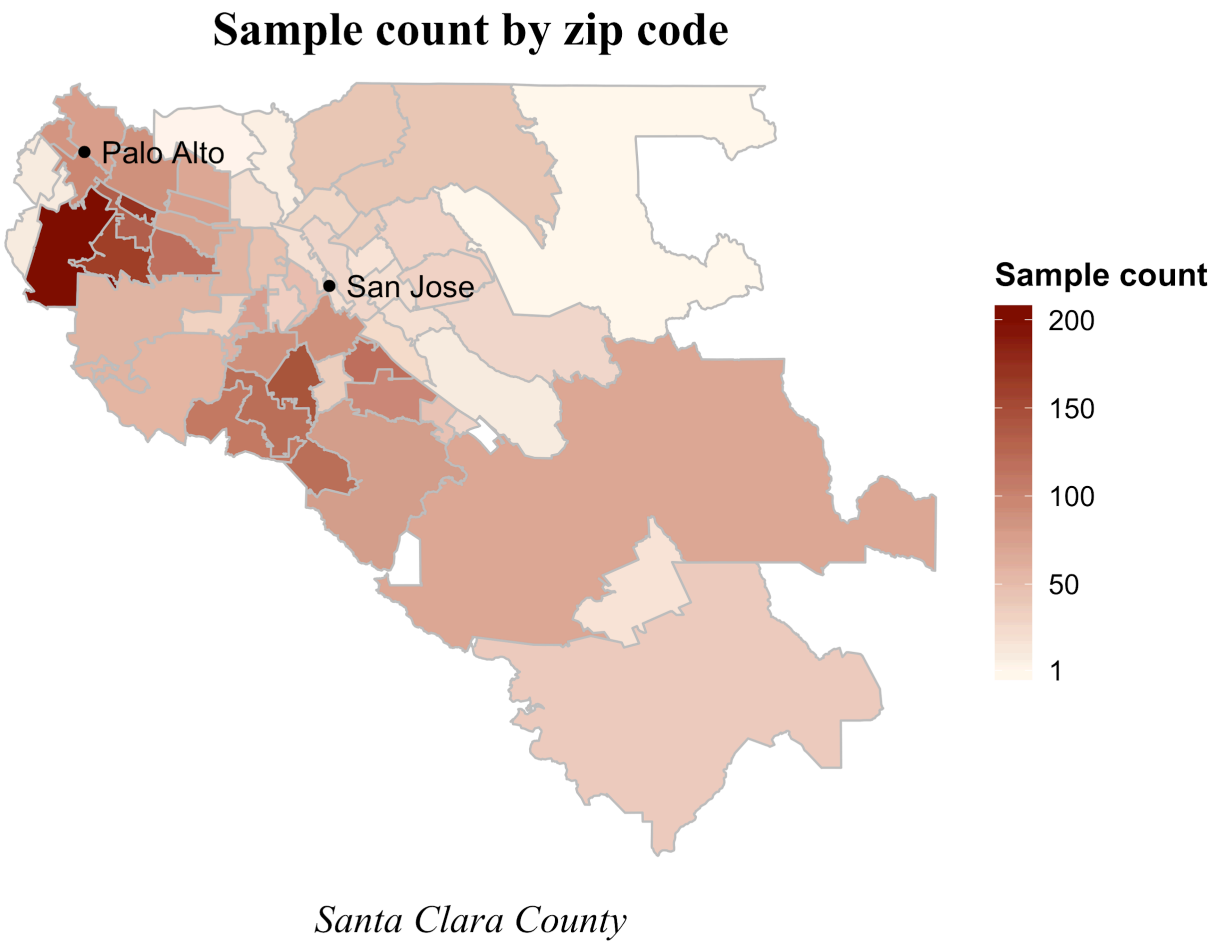


Figure 1: Of 3,439 individuals that arrived for testing, samples could not be obtained or tested (e.g. blood clotted) on 49, and an additional 60 were excluded because of invalid data (e.g. residence outside Santa Clara County), test data that could not be matched to a participant, or invalid test results. The final sample contained 3,330 records.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2



medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

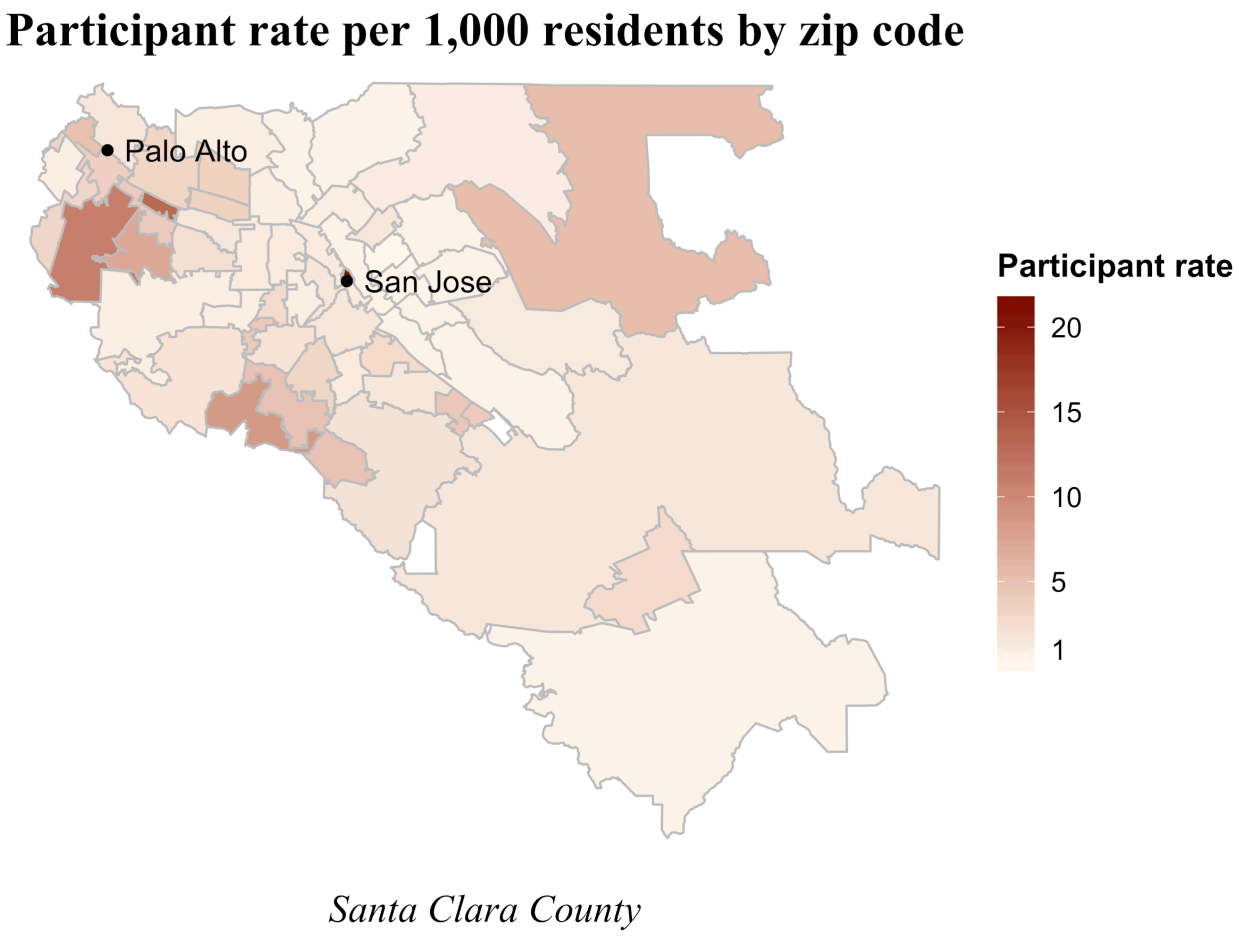


Figure 2: The number of registrations with complete records in our analytic dataset by zip code (panel A), and the participant rate per 1,000 residents in the zip code (panel B).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Tables

Table 1: Sample characteristics, relative to Santa Clara County population estimates from the 2018 American Community Survey

Characteristic		Sample - unadjusted	Sample - adjusted	County
Population (N)		3,330	3,330	1,943,411
Women (%)		63.1	49.7	49.5
Men (%)		36.9	50.3	50.5
Age (%)	0-4	2.1	2.6	6.2
	5-18	16.5	14.5	18.6
	19-64	76.3	78.4	62.3
	⩾65	5.0	4.5	12.9
Race/ethnicity (%)	Non-hispanic white	64.1	35.4	33.1
	Hispanic	8.0	24.9	26.3
	Asian	18.7	28.9	27.8
	Other	9.2	10.8	12.8

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2: Prevalence estimation in Santa Clara County. We report the prevalence and uncertainty bounds of estimates from unadjusted frequency counts, population-adjusted estimates, and population-adjusted + test performance-adjusted estimates. For the population-adjusted + test performance-adjusted estimates, we show estimates using the three test performance scenarios described in the Methods. For each point estimate, we present the method used to estimate the uncertainty bounds. Where noted, we clustered the standard errors for participants that brought a child with them (members of the same household).

Approach		Point estimate (%)	Uncertainty (95% CI)
Unadjusted (%)		50/3,330 = 1.50	1.11-1.97 (binomial exact) 1.07-1.93 (normal approximation, cluster adjusted)
Population-adjusted (only, %)		2.81	2.24-3.37 (normal approximation) 1.45-4.16 (normal approximation, cluster adjusted)
Population & test-performance adjustment	Manufacturer's data	2.49	1.80-3.17 (delta method)
	Local Stanford data	4.16	2.58-5.70 (delta method)
	Manufacturer's data + local data	2.75	2.01-3.49 (delta method)

medRxiv preprint doi: https://doi.org/10.1101/2020.04.14.20062463. The copyright holder for this preprint **(which was not peer-reviewed)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Article Information

Corresponding author
Eran Bendavid, ebd@stanford.edu

Conflict of interest
None. Of note, test kits were purchased from Premier Biotech for this study, and none of the authors have a relationship to the test manufacturer (beyond purchasing the tests).

Funding/Support
We acknowledge many individual donors who generously supported this project with gift awards.

Role of Funders
Our funders had no role in the design and conduct of the study, nor in the decision to prepare and submit the manuscript for publication.

Statistical Appendix

In this appendix, we describe our statistical methods in more detail. Section 1 describes our approach to calculating population weights. Section 2 describes our approach to adjusting our population prevalence estimate for the sensitivity and specificity properties of the LFA test kit we are using. Finally, Section 3 describes our approach to incorporating three separate sources of uncertainty in our prevalence estimates: sampling variability, error in the sensitivity estimate, and error in the specificity estimate.

Population weighting

In all but the unadjusted prevalence results, we reweight our sample to reflect the sex, race/ethnicity, and zip code of residence distribution of Santa Clara County. We derive these weights from the 2018 American Community Survey, from which we derived an estimate of the population of each zip code, as well as the race/ethnicity and sex distribution of county residents. We applied the county-wide sex distribution to each race/ethnicity group in each zip code to estimate the number of people within each zip-race-sex group. For example, zip code 95037 has a total of 51,652 residents, of which 50.6% are female, 49.7% are white, 7.7% are Asian, 33.0% are Hispanic, and 9.5% are other. We applied the female proportion to each race category in the zip code to obtain the number of residents in each zip-race-sex group.

Let $Epop_{zip,race,sex}$ be the number of people in each zip-race-sex cell produced by this calculation. Let $SmpSz_{zip,race,sex}$ be the number of people in our sample population in each cell. We need to up-weight cells that are underrepresented in our population relative to their frequency in Santa Clara County, and down-weight cells that are overrepresented. We can accomplish this by weighting proportional to the following ratio:

$$\tilde{\theta}_{zip,race,sex} = \frac{Epop_{zip,race,sex}}{SmpSz_{zip,race,sex}}$$

We renormalize so that the sum of our weights equals the size of the SCC sample, N. Define

$$S \equiv \sum_{zip,race,sex} \tilde{\theta}_{zip,race,sex}.$$

Our final sample weights are:

$$\theta_{zip,race,sex} = \frac{N}{S}\tilde{\theta}_{zip,race,sex}$$

This is identical to the formula we present in the Methods:

$$weight_{zsr} = \frac{\frac{N^{c}_{zsr}}{N^{C}_{total}}}{\frac{N^{S}_{zsr}}{N^{S}_{total}}}$$

Adjusting the prevalence estimate for test kit accuracy

Our main goal is to derive an estimate of the population prevalence of specific COVID-19 antibody seroconversion in Santa Clara County. However, we observe a noisy signal of antibody presence because the test kit we use has both type 1 and type 2 errors. In this section, we describe our approach to adjusting for these errors. To spare notation, we do not incorporate the sample weighting process we describe

above. Introducing sample weighting would complicate our notation, but would not change the approach. The analytic weights were used in the results shown in Table 2.

Let $\pi = P(COVID +)$ represent the population prevalence of antibodies to COVID-19, and let $q = P(TEST +)$ be the proportion of participants who test positive in our sample (this latter quantity measured using our sample weights). Note: we consider $TEST +$ as any band on the test kit indicating the presence of IgG or IgM antibodies or both.

Let $r = P(TEST + \mid COVID +)$ be the sensitivity of the test and let $s = P(TEST - \mid COVID -)$ be the specificity of the test. Let $z = P(COVID + \mid TEST +)$ be the positive predictive value of the test, and $y = P(COVID + \mid TEST -)$ be the (one minus) the negative predictive value of the test.

By Bayes rule,

$$P(COVID + \mid TEST +) = \frac{P(TEST + \mid COVID +)P(COVID +)}{P(TEST + \mid COVID +)P(COVID +) + P(TEST + \mid COVID -)P(COVID -)}$$

and

$$P(COVID + \mid TEST -) = \frac{P(TEST - \mid COVID +)P(COVID +)}{P(TEST - \mid COVID +)P(COVID +) + P(TEST - \mid COVID -)P(COVID -)}.$$

Rewriting these in our notation, we have:

$$z = \frac{r\pi}{r\pi + (1-s)(1-\pi)}, \text{and}$$

$$y = \frac{(1-r)\pi}{(1-r)\pi + s(1-\pi)}.$$

By the definition of conditional probability, we also have:

$$P(COVID +) = P(COVID + \mid TEST +)P(TEST +) + P(COVID + \mid TEST -)P(TEST -),$$

Or in our notation:

$$\pi = zq + y(1-q).$$

If we plug in our expressions for y and z and simplify, we have a quadratic expression in π.

$$1 = \frac{rq}{r\pi + (1-s)(1-\pi)} + \frac{(1-q)(1-r)}{(1-r)\pi + s(1-\pi)}.$$

We solve for π as a function of the sample prevalence, sensitivity, and specificity:

$$\pi = \frac{q + s - 1}{r + s - 1}.$$

There is one important caveat to this formula: it only holds as long as (one minus) the specificity of the test is higher than the sample prevalence. If it is lower, all the observed positives in the sample could be due to false-positive test results, and we cannot exclude zero prevalence as a possibility. As long as the specificity is high relative to the sample prevalence, this expression allows us to recover population prevalence from sample prevalence, despite using a noisy test.

Delta method approach to variance measurement

In this section, we derive an estimate of the variance of our population prevalence estimate that accounts for the error in the test kit sensitivity and specificity numbers, as well as sampling variation. We retain our notation from the previous section, and start with the formula we derived that relates all our variables of interest:

$$\pi = \frac{q + s - 1}{r + s - 1}.$$

We use a first order Taylor series approximation to derive the delta method approximation for $Var(\pi)$.

$$\pi \approx \left.\frac{\partial \pi}{\partial q}\right|_{(q_0, r_0, s_0)} (q - q_0) + \left.\frac{\partial \pi}{\partial r}\right|_{(q_0, r_0, s_0)} (r - r_0) + \left.\frac{\partial \pi}{\partial s}\right|_{(q_0, r_0, s_0)} (s - s_0).$$

The three relevant derivatives are as follows:

$$\left.\frac{\partial \pi}{\partial q}\right|_{(q_0, r_0, s_0)} = \frac{1}{r_0 + s_0 - 1},$$

$$\left.\frac{\partial \pi}{\partial r}\right|_{(q_0, r_0, s_0)} = \frac{q_0 + s_0 - 1}{(r_0 + s_0 - 1)^2}, \text{and}$$

$$\left.\frac{\partial \pi}{\partial s}\right|_{(q_0, r_0, s_0)} = \frac{r_0 - q_0}{(r_0 + s_0 - 1)^2}.$$

Since q, s, and r are all derived from independent samples:

$$Var(\pi) \approx \left(\frac{1}{r_0 + s_0 - 1}\right)^2 Var(q) + \left(\frac{q_0 + s_0 - 1}{(r_0 + s_0 - 1)^2}\right)^2 Var(r) + \left(\frac{r_0 - q_0}{(r_0 + s_0 - 1)^2}\right)^2 Var(s).$$

Thus, the delta method estimate for $Var(\pi)$ is a weighted sum of three variance terms: sampling variability, sampling error in sensitivity, and sampling error is specificity.

In the main text, we explore three alternative assumptions about the sampling error in sensitivity and specificity. For each of these scenarios, we estimate the variance in sensitivity and specificity by the standard formulas for a binomial outcome: $Var(r) = \hat{r}(1 - \hat{r})$ and $Var(s) = \hat{s}(1 - \hat{s})$, where $\hat{r}$ and $\hat{s}$ are the estimated values of sensitivity and specificity that pertain to each scenario.

- In the first scenario, we estimate these quantities based upon the numbers provided by the manufacturer of the test kits. For sensitivity, the manufacturer reported 78 positive test readings for 85 samples (from Chinese blood samples) known to have specific IgM antibodies to the receptor-binding domain (RBD) spike on the SARS-nCOV2 virus. They reported 75 positive test readings for 75 of the samples with specific IgG antibodies to the same RBD spike. We adopt a conservative estimate of sensitivity equal to $\hat{r} = \frac{78}{85} \approx 91.8\%$. The manufacturer reports specificity based on an experiment using their kit on a sample of 371 known negative blood samples collected from before the epidemic, and 369 were tested negative. This implies a specificity of $\hat{s} = \frac{369}{371} \approx 99.5\%$.
- In the second scenario, we estimate these quantities based on tests run locally at Stanford University. We identified serum from 37 patients who had RT-PCR-confirmed cases of COVID-19 and either IgG or IgM on a locally-developed ELISA assay; of these, 25 tested positive with the test kit, implying a sensitivity of $\hat{r} = \frac{25}{37} \approx 67.6\%$. We also identify serum from 30 patients drawn from before the COVID-19 epidemic, and all 30 tested negative, implying a specificity of

$\hat{s} = \frac{30}{30} = 100\%$. In this case, since the standard formula implies $Var(s) = 0$, we instead estimate $Var(s)$ from a simple transformation of the width of the Clopper-Pearson exact confidence interval.

- In the third scenario, we estimate these quantities by combining the manufacturer tests with our local tests in a simple additive fashion. Under these assumptions, the sensitivity estimate is $\hat{r} = \frac{103}{122} \approx 84.4\%$ and the specificity estimate Is $\hat{s} = \frac{399}{401} \approx 99.5\%$.

To provide one concrete example, we have the following values under the first scenario:

$q_0 = 0.028$ $\quad$ $var(q_0) = 0.027$

$r_0 = 0.918$ $\quad$ $var(r_0) = 0.076$

$s_0 = 0.995$ $\quad$ $var(s_0) = 0.005$

$$\left(\frac{1}{r_0+s_0-1}\right)^2 = 1.202$$

$$\left(\frac{q_0 + s_0 - 1}{(r_0 + s_0 - 1)^2}\right)^2 = 7.423 \times 10^{-4}$$

$$\left(\frac{r_0 - q_0}{(r_0 + s_0 - 1)^2}\right)^2 = 1.143$$

Combining all of those values, we get $Var(\pi) = 0.039$, and $SE(\pi) = 0.0034$

We use this procedure throughout.