Prevalence estimates of COVID-19 by web survey compared to inadequate testing

David N Ku (david.ku@me.gatech.edu)
Georgia Institute of Technology https://orcid.org/0000-0002-6034-8004
Ben Ku
Gwinnett County
Traci Leong
Emory University
Zixiang Liu
Georgia Institute of Technology

Technical advance

Keywords: COVID-19, prevalence, symptoms, sampling, web-based, crowd sourcing

DOI: https://doi.org/10.21203/rs.3.rs-40294/v1

Abstract

Background

Current prevalence of COVID-19 drives many policy decisions, but is hampered by ambiguities in testing and reporting. We propose an alternative method for estimating community prevalence that is inexpensive and timely. We test the Hypothesis that the survey sampling provides a quantitative prevalence that is similar to widespread genomic or serological testing.

Methods

We have built a simple, web-based survey of signs and symptoms for COVID-19 based on six questions. No personally identifiable information is collected to maintain privacy. Sampling can be directed to a population of interest such as a company, or broadcast widely to get geographic sampling. Data reporting can be real-time and plotted onto zipcode maps. Rates of prevalence were calculated from presumed COVID cases and respondents, with confidence intervals based on the Blaker method.

Results

The website was created quickly, and survey results were quantitatively useful after only a few days. Analyzing 3161 cases from CountCOVID.org, we found a community prevalence of 7% in Georgia that was much greater than the reported confirmed cases. Our prevalence estimate of 21% in New York City was similar to the reported 19.6% by surveillance antibody serotesting. Our estimates are validated by five other community surveillance studies using genomic or antibody testing.

Conclusions

Prevalence and incidence of COVID-19 symptoms in the community can be estimated by a crowd-sourced website at considerably less expense than widespread testing.

Background

The media has concentrated on “confirmed cases” as the predominant metric to follow the spread of COVID-19. PCR testing is the standard by which a diagnosis can be made for infection. However, PCR testing may give a highly skewed indication of important epidemiologic metrics of community prevalence and incidence. PCR testing is *not* the only way to measure the spread of COVID-19. Tests have been inadequate due to lack of PCR cartridges and nasal swabs, preference for diagnosis of severe cases, and inability to perform population sampling for surveillance.[1-3] When the number of available tests is small, the test kits are reserved for severe cases in hospitals to give a definitive diagnosis and direct treatment for individual patients. There are many sources of inconsistencies in health care reporting between local

and state public health agencies.[3] In this setting, the number of confirmed cases will be much smaller than the actual number of people infected in the community. The ratio between community prevalence to confirmed cases can be as high as 91x by serological testing.[4] Thus, the reliance on confirmed cases to make decisions on community risk may be dangerous.[3,5,7] There is a strong need to quantify an estimate of incidence of disease in the community that is within an order of magnitude of the true incidence that has not been achieved by current confirmed cases that may be off by 10-90x.

While PCR can be specific, there are reports of false negatives and poor negative predictive value when applied to the general population where the underlying incidence will be low.[6] More testing may become available in the future, but it is equally unrealistic to rely on future testing only for public health decisions. The President has threatened to withhold testing and testing results to reduce the perceived number of cases. Some advocate the testing of an entire population at multiple times to capture the spread of the disease. This would be extremely expensive to test millions of people, then repeat the testing every two weeks to catch an outbreak. The naso-pharyngeal swab is currently the best method for obtaining samples for COVID-19, but is uncomfortable, with almost no one volunteering to give this sample a second time. Some people report this experience as similar to a brain biopsy or that it has given them headaches for 3 days.

Epidemiology has traditionally counted prevalence and incidence by directly interviewing populations.[6] Clinical diagnosis of COVID-19 currently relies primarily on symptoms and signs, especially for mild cases.[5] Traditionally, public health surveillance does not require PCR testing, but does require ongoing, systematic collection of data, defined by the CDC as “Syndromic Surveillance”.[5,25] Statistically, sampling is an efficient way of gathering these population statistics without testing everyone.[8] Some groups have reported using digital means such as a smartphone application or web-based series of directed questions to gather information on symptoms that can be tracked individually.[9-11] Triangulation of estimates of prevalence and incidence by independent means can substantially reduce errors.[7] The independent information may address issues if a state manually changes data[12] or claims that corona virus outbreaks are due to rises in testing.[13]

We hypothesize that a web-based system of data collection from the population can provide a reasonable and useful estimate of prevalence at much lower cost than expensive, widespread laboratory testing. Our metrics are then compared and validated against several independent publications describing community testing.

Methods

An interactive website was developed to survey the community for signs and symptoms of COVID-19. See CountCOVID.org. This minimum viable system was developed as a proof of concept that could be implemented within one week with minimal expenditures. No application was required to be downloaded and no personally identifiable information was collected or extracted. Because no personally identifiable

information was collected, the need for consent was waived by the Institutional Review Board of Georgia Institute of Technology. The website was purposely designed to ask only 6 binary questions to encourage completion (Figure 1). The questionnaire was newly written for this study by one of the authors and the authors retain copyright of the website and print screens. User zip code was requested to assign cases to a geographic area. Users were asked about fever, cough, shortness of breath, loss of smell, difficulty breathing, and previous COVID testing. These symptoms were selected to be sensitive and specific to reports of COVID illness. [1,2] The respondents were instructed to answer any symptom experience starting March 2020. A ThankYou page acknowledged the submission of the survey and provides instantaneous information to the user (Figure 2). Knowing the sample size, we could calculate a prevalence per 100,000 population (php) for a geographic region.[8] Crude rates of prevalence were calculated from presumed COVID cases and respondents in any one county, with confidence intervals based on the Blaker method.[14] No adjustment for age and gender was made due to the lack of personal identification information. For geographic data visualization, the open-source QGIS 3.12 GIS application was used.

Results

The first 3161 cases were collected between April 10 and April 25, 2020. The responses to the six questions were analyzed for COVID symptoms (Figure 1). 9% had fever, 18% had a dry cough, and 4% had lost smell (Figure 2). 7% reported difficulty breathing, while 88% could easily hold their breath at the time of the survey. Only 1% had tested positive for COVID. Using a combination of signs and symptoms led to a metric of Presumed COVID infection in 7% of the sample population. Presumed COVID was defined for this dataset as: Positive COVID test; Fever + Cough; and Loss of Smell with any other symptom such as fever, cough, difficulty breathing or inability to hold breath.

This yielded a prevalence of 7,000 (php) for Georgia(7%) and about 20,000 for New York City(20%). Note that the prevalence values are based on limited sampling, so the 95% confidence levels are given in Table 1. We used data from 1018 cases in Georgia and 103 cases in NYC. The COVID confirmed case counts for reference were obtained from the JHU CSSE COVID-19 data repository.[15]

Table 1 Prevalence and Confidence Intervals (95%) by County (except Boston which is Greater Area)

County	php	Lower 95% CI	Upper 95% CI
Dekalb	6349	4117	9290
Fulton	5255	3675	7264
Cobb	9890	6115	14948
Gwinnett	4545	1997	9565
Bergen	5882	301	28216
Westchester	16667	3045	45563
Bronx	16667	851	59386
NYC	15625	6365	32318
Hudson	25000	1274	75140
King	16667	851	59386
Nassau	29412	12377	54420
Queens	22222	4101	56003
Greater Boston	5260	269	25167

The county prevalence varied throughout the state of Georgia. The four major counties in the Atlanta metropolitan area are shown in Figure 3. At the county level, the prevalence php for Fulton county from CountCOVID.org was 5,255 that appears to be less dense than Cobb county that had the highest prevalence of 9890, despite the fact that Fulton had the most total confirmed cases in Georgia.

Counties in the New York City area are shown in Figure 4. Note that Nassau and Queens display a very high php of up to 29,000, while Manhattan and Westchester were lower at around 16,000 php. Hudson county in New Jersey was high, while the adjoining Bergen county was much lower at 5900.

Comparison with confirmed cases in Georgia indicate that the Presumed COVID cases in the wider community are ~ 40x that of confirmed cases.[16] There were regional differences as Fulton county had Presumed COVID cases that are about 20x higher than the number of confirmed cases of ~300 for this period. Cobb county symptomatic cases were ~40x higher than the reported number of confirmed cases of ~230.

For the month of May, the survey was changed to ask about symptoms in the past two weeks only. The time restriction yields an estimate of incidence rather than prevalence. The data was then analyzed in weekly intervals to estimate incidence as it changed. The incidence for May is given in Table 2.

Table 2

Dates	Incidence in GA (%)
May 3-May 9	1.42
May 10-May 16	0.56
May 17-May 23	1.34
May 24-May 30	Sample too small
May 31-June 6	1.79
June 7-Jun 13	0.94

Discussion

This paper illustrates the use of a web-based system for collecting salient signs and symptoms of COVID-19 in a community. The crowd-sourced information can estimate local prevalence and incidence without the heavy expense of testing of biological samples. The website questionnaire can be completed in 30 seconds. The system relies on self-reporting, not on location tracking to protect the anonymity of the user by only requesting the zip code. Repeated completion of the website can give accurate daily updates. Responses were rapid with the majority of responses coming within a 24-hour period of a request. No serologic or molecular testing was required for the prevalence estimates.

To illustrate the effectiveness of a web-based model such as ours, one can compare our results with other surveillance studies of community testing. Our system reported 21% prevalence in NYC. After the study had accrued over 3000 participants, including over 1300 in New York City, Mayor Cuomo reported antibody testing results that indicated 1/5 of New Yorkers were infected (19.9%).[17,18] Our survey estimate for the Greater Boston area is 5.24% (2.687-9.109, Table 1). Three weeks later, the City of Boston reported a seroprevalence study of 9.9% (6.3% to 13.3%) that varied by zip code on May 18, 2020.[19] Tests in other geographical areas of the USA show a similar magnitude of community prevalence to the 7% we found in Georgia. In Los Angeles County, antibody testing estimated prevalence at 4.06% (2.84-5.6%).[20] Antibody seroprevalence in Santa Clara County showed 2.8% (1.3-4.7%).[21] A recent report of prevalence by large-scale RT-PCR testing in the Baltimore-Washington, DC region shows 16.3% (16.0-16.7%) from March 11 to May 25.[22] The remarkable similarity in prevalence estimates between our survey-based study and the aforementioned testing studies highlights the ability of self-reporting to yield a reasonable determination of COVID19 prevalence.

The prevalence values for community COVID are much greater than confirmed cases. Given the preferential testing of moderate to severe cases which present in the hospital setting, it is likely that the number of confirmed cases greatly underestimates the overall prevalence and incidence of the disease. When the CountCOVID results showed 40x confirmed cases in Georgia, we were initially concerned that this ratio was improbable and too high. Since then, two antibody surveillance studies in California gave

ratios of 43.5x (28-55) for Los Angeles County[20], and 54x (25-91) for Santa Clara County[21]. Thus, despite the striking discrepancy with confirmed cases, our estimate of 40x for Georgia appears to be validated.

The ratio of prevalence values to confirmed counts may fall as the number of widespread testing (php) increases. For example, the ratio for NYC (20x) is about 1/2 of that for Georgia (40x), although NYC has 5x more confirmed cases php. NYC boasts the largest number of tests in the country in May 2020. Although presumed cases and confirmed cases may approach each other with increasing testing, urgency and financial consideration is of utmost importance.

Incidence will track the number of new cases and may provide information on basal levels and outbreaks. For the month of May, we estimate the prevalence in Georgia to be between 0.56% to 1.79% as the state reopened. In comparison, the COVID-19 website estimates symptomatic COVID incidence as between 0.2-0.4% for a slightly different time frame.[23] The order of magnitude is similar and all the values may reflect sampling bias and definition of COVID symptoms. Nonetheless, the similarity in values provides confidence that this method of self-reporting is scientifically reasonable.

Criteria for presumed COVID may need to be refined as we learn more about this disease. We did not ask for symptoms of diarrhea or “COVID toes” in the initial survey. Each question is subject to False Positives and False Negatives. For instance, cough was quite common in the community, especially given the temporal association with an uptick in hay fever symptoms during the spring. Thus, cough by itself at 17% was not specific and would have many False Positives. Fever was at 9%, but can be caused by a plethora of illness. However, the intersection of fever and cough yielded about 5%. We included in the definition of Presumed COVID all positively tested individuals and loss of smell plus at least one other symptom. These additional categories yielded a small subset of the total cases of the final estimate of Presumed COVID prevalence of 7%. This algorithm of signs and symptoms mirrors current clinical judgement. We do note that the survey by the Imperial College with predominantely users from the UK yielded a much greater percentage of loss of smell.[24] It is not known whether this is a difference in patient population or survey technique. The selection of other criteria can be made post-hoc, but our current criteria yielded extraordinarily similar results to serotesting. Given time, the analysis can be back-calibrated with selective surveillance testing to correct for errors or biases. However, the application of natural intelligence instead of artificial intelligence may be good enough.

Prevalence is an important parameter to assess for determining the effectiveness of social distancing, testing, and herd immunity. The proposed method of web sampling is rapid and inexpensive. Given the current financial crisis which has resulted from this pandemic, economic burden can be minimized in the quantification of disease burden by using web sampling. In contrast, serologic or PCR testing is so much more expensive. PCR testing of 1000 people would be approximately $1 million. The web-based survey of 1000 people is estimated to cost approximately $100. Because one can sample quickly and often, a sudden increase in symptoms on CountCOVID.org may provide advance warning of an outbreak.

This survey, and all sampled studies have bias. Bias is possible if the sample size is small or skewed by the population completing the survey. It would be essential to widely encourage the population to participate in a web-based survey such as ours. The current results are likely biased to adult faculty and staff from Georgia Tech who are employed, instead of the general population. Unfortunately, the comparison Confirmed cases is also subject to large bias since they are mostly directed at severe cases that can access high quality health care, and not a sampling of the wider community. Conversely, sampling may be purposely restricted to quantitatively assess the baseline and trends for selected populations such as the elderly or particular neighborhoods. There are other electronic based systems that do a sampling based on signs and symptoms.[13-14] We applaud all of these systems and encourage them to report on their findings for academic comparison and collaboration.

Conclusions

We describe an inexpensive, crowd-sourced system to rapidly obtain prevalence in the community that does not rely on testing. The website may be directed to a particular population, such as a large office building or an at-risk population, to identify an outbreak faster than following confirmed cases or COVID-19 deaths.

Abbreviations

COVID-19: COrona VIrus Disease of 2019.

PCR: Polymerase Chain Reaction

Declarations

Ethics approval and consent to participate

As we are not collecting any personally identifiable data, no Human Subjects consent is required. The IRB of Georgia Institute of Technology reviewed this study and waived the need for consent.

Availability of data and materials

The datasets analysed during the current study are available from the corresponding author on reasonable request.

Authors Contributions

DK designed the study, analyzed the data and wrote the manuscript. BK wrote the software code for the website. TL provided statistical analysis of the results. ZL displayed the prevalence and incidence on state maps. All authors reviewed and approved the manuscript.

Acknowledgements

The Nebo Agency, LLC provided the new webpage design for CountCOVID.org

Competing Interests

The authors declare that they have no competing interests.

Consent for Publication

No consent is required as there is no individual person's data.

Funding

Funding for the design of the study and collection, analysis, and interpretation of the data and in writing the manuscript are provided by the Larry P. Huang Chair Professorship held by David Ku.

Maps: All maps are generated by us, not taken from another source.

References

1. Wu Z, McGoogan JM. Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China. JAMA 2020;323; 13, 1239. ***doi:1001/jama.2020.2648***
2. DelRio C. COVID-19 – New Insights on a Rapidly Changing Epidemic. *JAMA*.2020;323(14):1339-1340. doi:10.1001/jama.2020.3072
3. Silver N. Coronavirus Case Counts are Meaningless*. https://fivethirtyeight.com/features/coronavirus-case-counts-are-meaningless/
4. USC-LA County Study: Early Results of Antibody Testing Suggest Number of COVID-19 Infections Far Exceeds Number of Confirmed Cases in Los Angeles County. http://publichealth.lacounty.gov/phcommon/public/media/mediapubhpdetail.cfm?prid=2328&fbclid=IwAR2odNQwAchsTdMsTPW9Q4YfqLPzSreSn16imtSsKzV-eWyGm6wfK5ecv-A
5. Rosenthal E. The Real Tragedy of Not Having Enough COVID-19 Tests. NYT April 6, 2020. https://www.nytimes.com/2020/04/06/opinion/coronavirus-testing.html?searchResultPosition=1
6. Introduction to Public Health Surveillance – CDC Public Health 101 Series. https://www.cdc.gov/publichealth101/surveillance.html
7. Ioannidis JPA. All science should inform policy and regulation. PLoS Med 15(5): e1002576. May 3, 2018. https://doi.org/10.1371/journal. pmed.1002576
8. National Institutes of Health/ Statistics. https://www.nimh.nih.gov/health/statistics/what-is-prevalence.shtml
9. Drew DA, et al. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. Science 10.1126/science.abc0473 (2020)
10. Stanford site: https://med.stanford.edu/covid19/covid-counter.html
11. Georgia Institute of Technology site: https://countcovid.org

12. Gabbatt A. Scientist produces own Florida Covid-19 count after being fired by state. The Guardian June 15, 2020.
13. Rogers K, Martin J. Pence Misleadingly Blames Coronavirus Spikes on Rise in Testing, New York Times, June 15, 2020.
14. Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions, *Canadian Journal of Statistics*28 (4), 783–798
15. JHU Website (https://github.com/CSSEGISandData/COVID-19)
16. Georgia COVID-19 Statistics, Atlanta Regional Commission website: https://atlantaregional.shinyapps.io/COVID19/
17. Goodman JD, Rothfield M. 1 in 5 New Yorkers May Have Had Covid-19, Antibody Tests Suggest. New York Times. April 23, 2020 https://www.nytimes.com/2020/04/23/nyregion/coronavirus-antibodies-test-ny.html?searchResultPosition=1
18. NYC Health COVID-19: Data website: https://www1.nyc.gov/site/doh/covid/covid-19-data.page
19. City of Boston Public Health Commission. *Results Released for antibody and COVID-19 Testing of Boston Residents,* May 18, 2020.
20. Sood N, et al. Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. 2020;323(23):2425-2427. doi:10.1001/jama.2020.8279
21. Bendavid E, Mulaney B, Sood N et al. COVID-19 Antibody Seroprevalence in Santa Clara County, California. April 30, 2020. https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v2
22. Martinez D, et al. SARS-CoV-2 Positive Rate for Latinos in the Baltimore-Washington, DC Region. June 18, 2020 Published Online: June 18, 2020. doi:10.1001/jama.2020.11374
23. COVID Symptom Study site: https://covid.joinzoe.com/us
24. Menni C, Valdes AM, Freidin MB. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nature medicine. https://doi.org/10.1038/s41591-020-0916-2.
25. Immergluck LC, et al. Geographic surveillance of community associated MRSA infections in children using electronic health record data. BMC Infectious Diseases (2019) 19:170. https://doi.org/10.1186/s12879-019-3682-3

Figures

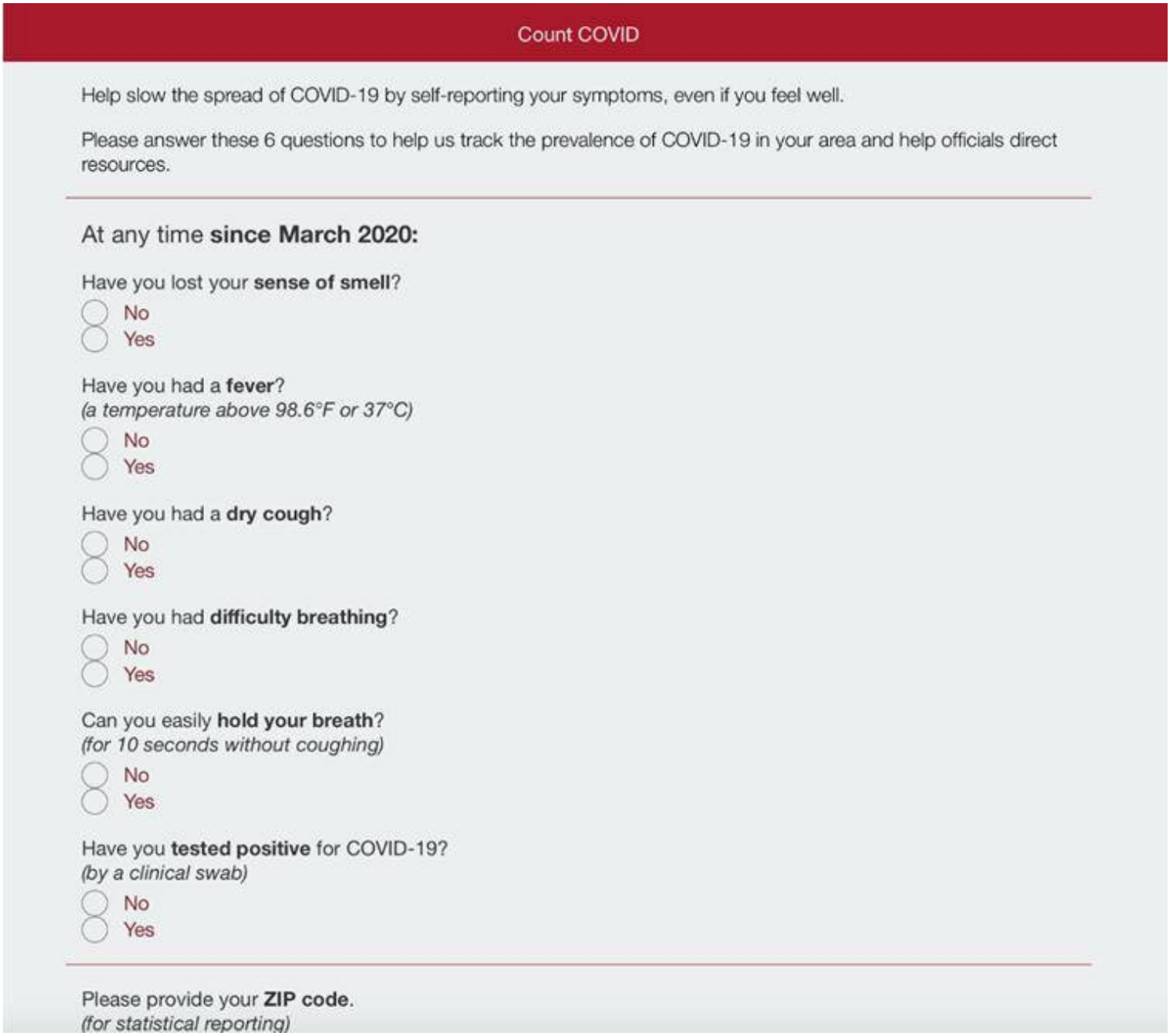

Figure 1

First page of website with questions for survey on one screen

Figure 2

Thank-you page shown after submission of data showing aggregate results in realtime.

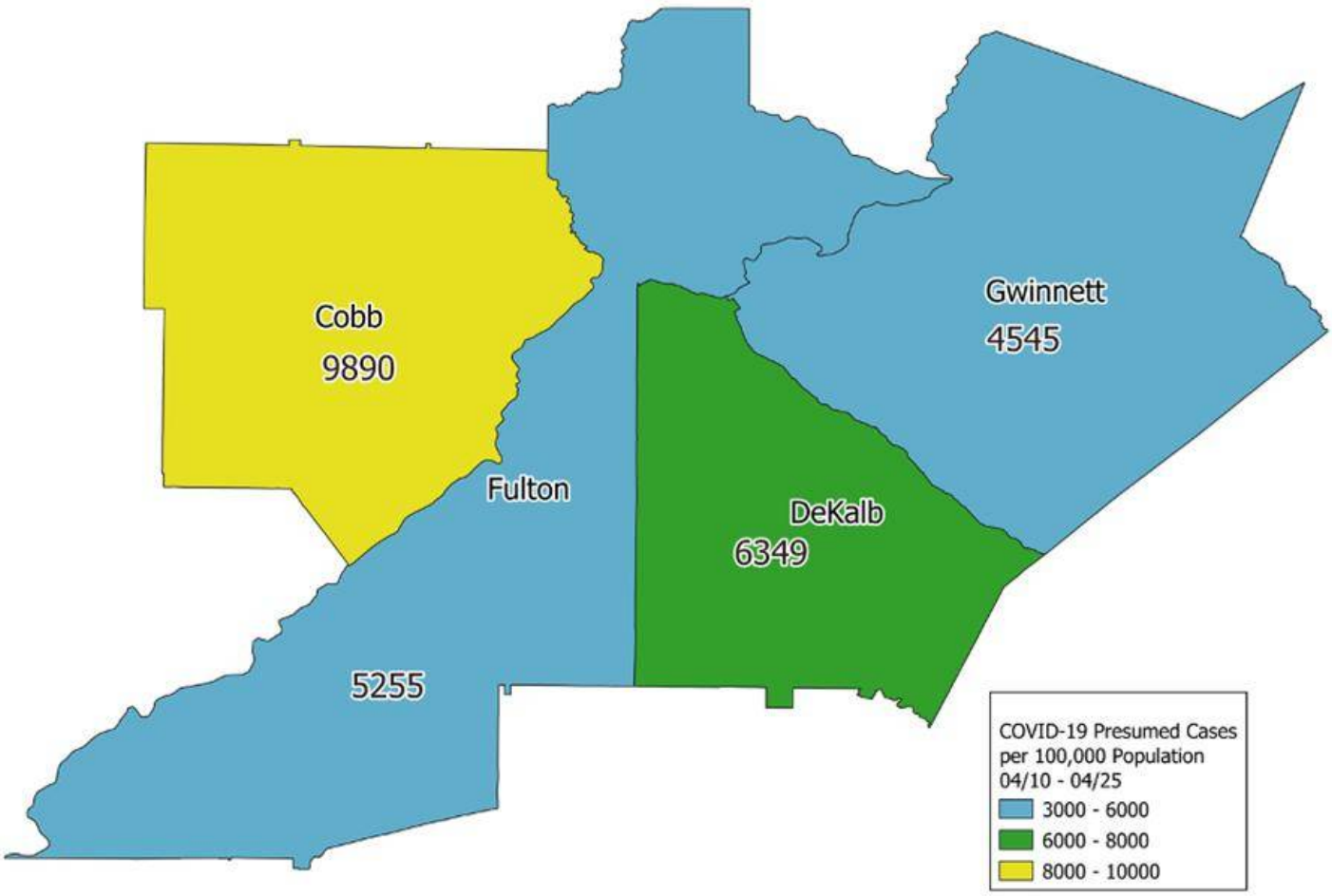

Figure 3

Example map of prevalence per hundred thousand population for four Atlanta area counties. Fulton county surprisingly had less prevalence as it has the most confirmed cases, while Cobb county had more prevalence.

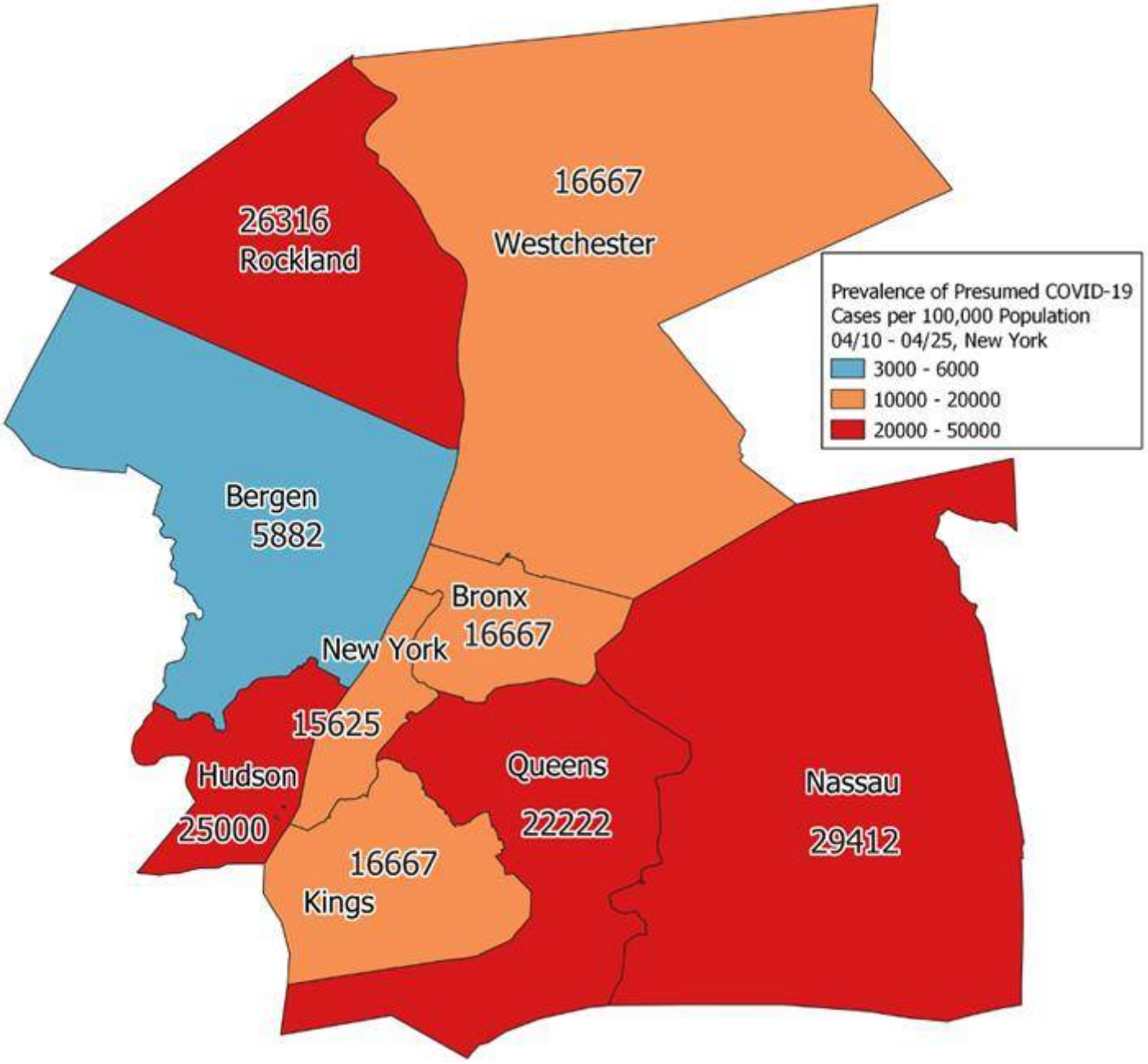

Figure 4

Example map of prevalence per hundred thousand population for eight counties near New York City. The heterogeneity is evident and may have utility in deciding issues of public health response.