GPS-estimated foot traffic data and venue selection for COVID-19 serosurveillance studies

Tyler S. Brown[1,2], Pablo Martinez de Salazar Munoz[1], Abhishek Bhatia[1], Bridget Bunda[2], Ellen K. Williams , David Bor[3], James S. Miller[2], Amir Mohareb[2], Vivek Naranbai[2], Wilfredo Garcia Beltran[2], Tyler E. Miller[2], Julia Thierauf[2], Wenxin Yang[2], Doug Kress[4], Kristen Stelljes[4], Keith Johnson[4], Daniel B. Larremore[5], Jochen Lennerz[2], A. John Iafrate[2], Satchit Balsari[1], Caroline O. Buckee[1,*], and Yonatan H. Grad[1,*]

[1]Harvard T.H. Chan School of Public Health
[2]Massachusetts General Hospital
[3]Cambridge Health Alliance
[4]Somerville Board of Health
[5]University of Colorado, Boulder
[*]Co-senior authors

February 3, 2021

1 Abstract

Tracking the dynamics and spread of COVID-19 is critical to mounting an effective response to the pandemic. In the absence of randomized representative serological surveys, many SARS-CoV-2 serosurveillance studies have relied on convenience sampling to estimate cumulative incidence. One common approach is to recruit at frequently visited community locations ("venue-based" sampling), but the sources of bias and uncertainty associated with this strategy are still poorly understood. Here, we used data from a venue-based community serosurveillance study, GPS-estimated foot traffic data, and data on confirmed COVID-19 cases to report an estimate of cumulative incidence in Somerville, Massachusetts, and a methodological strategy to quantify and reduce uncertainty in serology-based cumulative incidence estimates obtained via convenience sampling. The mismatch between the geographic distribution of participants' home locations (the "participant catchment distribution") and the geographic distribution of infections is an important determinant of uncertainty in venue-based and other convenience sampling strategies. We found that uncertainty in cumulative incidence estimates can vary by a factor of two depending how well the participant catchment distribution matches the known or expected geographic distribution of prior infections. GPS-estimated business foot traffic data provides an important proxy measure for the participant catchment area and can be used to select venue locations that minimize uncertainty in cumulative incidence.

2 Introduction

Tracking the COVID-19 pandemic is a critical public health priority. Given the lack of comprehensive diagnostic testing, cumulative incidence of COVID-19 can instead be estimated using serological surveys that detect anti-SARS-CoV-2 antibodies, which indicate past SARS-CoV-2 infection. While randomized representational population surveys provide the basis for the most accurate cumulative incidence estimates, the exigencies of the ongoing public health emergency and the urgent need for information to inform decision-making has prompted many studies to employ "convenience", or non-probability, sampling [1, 2, 3].

The scientific and public health value of COVID-19 cumulative incidence estimates derived from convenience samples is controversial [4]. All forms of convenience sampling have inherent biases, but depending on

the question of interest and information needed to answer it, convenience sampling may nevertheless provide important, actionable information for public health decision-makers. Understanding the nature and extent of biases in cumulative incidence estimates from convenience samples, and finding analytical approaches to ameliorate them, is therefore an important goal.

Recent studies have detailed how sample size and serological test performance are expected to influence uncertainty in SARS-CoV-2 serosurveillance studies [5]. If substantial heterogeneity in cumulative incidence exists across demographic groups, the use of demographic-informed sampling, in which the number of tests performed in a given demographic group is weighted by its corresponding proportion of the total population, can substantially decrease uncertainty compared to uniform or demographically-naïve sampling strategies [5]. By extension, in situations where substantial geographic heterogeneity in cumulative incidence is likely, sampling strategies informed by available data on the geographic distribution of infections (for example, data on reported, laboratory-confirmed infections by geographic area) could help reduce uncertainty in serology-based estimates of cumulative incidence. These approaches may be particularly useful in situations where cost or logistical considerations limit the total number of serological tests that can be performed.

In this study, we examined the geographic and demographic biases that can influence venue-based and other SARS-CoV-2 serosurveillance studies that employ convenience sampling. To do this, we jointly analyzed data from a venue-based serosurveillance study in Somerville, Massachusetts, data from aggregated GPS-based data on visitors to the study location, and data from neighborhood-level mapping of PCR-confirmed SARS-CoV-2 infections in the community. We found that geographic heterogeneity in underlying cumulative incidence is an important determinant of uncertainty in venue-based studies, specifically if the catchment distribution (the home locations of visitors to a given location) of the study venue results in under-sampling of locations or populations with higher underlying cumulative incidence of infection. In addition, we report serological estimates for the cumulative incidence of SARS-CoV-2 infection in Somerville, informed by our analysis of geographic heterogeneity as an important source of uncertainty for venue-based sampling. We propose that GPS-estimated business foot traffic data, when available, can help predict the catchment distribution of a given venue, and this can inform selection of study venues that minimize uncertainty from geographic heterogeneity in infections.

3 Methods

3.1 Study design and recruitment

We tested the SARS-CoV-2 serological status of asymptomatic adults at a temporary study site near an essential business in Somerville, MA over 4 days, June 4th,5th, 8th, and 9th (Thursday, Friday, Monday, and Tuesday), 2020. Study protocols and site logistics were developed by a collaborative working group, including researchers from Massachusetts General Hospital and the Harvard T.H. Chan School of Public Health, the City of Somerville, and representatives from local businesses located near the study site. The study was designated minimal risk human subjects research and approved by institutional review boards at Massachusetts General Hospital and the Harvard T.H. Chan School of Public Health (Protocol number: 2020P001081). Additional details on recruitment, study procedures, and serological testing are included in the supplemental information.

3.2 Business foot traffic and census block group visitor data

We defined the catchment area of our study location as the distribution of home locations for individuals visiting the business located at the study site. We predicted this catchment area using commercially-available third-party foot traffic data [6]. This data source uses mobile phone-associated GPS information to estimate home and work locations at the level of census block group (CBG) for visitors to designated points-of-interest (POI), such as businesses, and specific CBGs. Available data for visitors to the business at our study location was relatively sparse compared to data for visitors to the CBG in which it is located (n= 518 versus n= 1570 recorded visitors, respectively). Thus, we used CBG-level visitor data for June 2020 as the primary data source for visits in our analysis.

We re-aggregated CBG data by electoral ward to match survey data on study participants' home locations using the following procedure: (1) A map of the CBG-level visitor data (where the number of visitors from

each CBG is stored as an attribute of its corresponding polygon on the map) was rasterized, using the assumption that visitor counts are uniformly distributed across each polygon; (2) The rasterized polygons were reapportioned to the alternative geometry (wards) using the "zonal statistics" function in QGIS [7]. For privacy purposes, the third-party data service does not report home location information for CBGs with < 2 visitors and CBGs with < 4 visitors are all reported as having exactly 4 visitors. Given the potential uncertainty this introduces for wards with low visitor counts, we excluded wards contributing < 5 reported visitors and those representing less than 1% of total visitor traffic to the study site CBG (Figures S1 and S2).

3.3 Public health acute infection data

We obtained data on 916 PCR-confirmed COVID-19 cases with documented home addresses in Somerville, counted from the onset of the epidemic until June 8th, 2020, from the Massachusetts Virtual Epidemiologic Network (MAVEN). COVID-19 cases are reported to MAVEN by state and local health agencies, and cases are designated as "confirmed" if they have a positive result for SARS-CoV-2 RNA detection using a molecular amplification detection test that has been approved or authorized by the FDA. Data were anonymized and aggregated by electoral ward prior to analysis. The total number of individuals tested for SARS-COV-2 via RT-PCR was also obtained from MAVEN.

3.4 Modeling uncertainty in serology-based cumulative incidence estimates

We developed a stochastic model to analyze how sampling strategy and location influence uncertainty in serology-based cumulative incidence estimates. In brief, this model randomly draws participants from a simple synthetic population, stratified by age and location, according to a sampling strategy, which specifies the number of participants (in this case, equal to the number of serological tests performed) in each age-location group. The "true" underlying cumulative incidence in each age-location group is specified using local and state-level data on confirmed COVID-19 cases [8], and this proportion is used to inform the probability with which an individual with "true" prior infection is sampled via a given sampling strategy. The number of true positives and false positives, drawn from the number of individuals with "true" prior infection using the performance characteristics of the serological test, is used to generate population-level cumulative incidence estimates for each instance of the simulation. We repeat this process 1,000 times and report variance across a range of values for n (the total number of tests performed) and m (representing a multiplicative factor by which "true" incident cases exceed detected cases). Additional details are provided in the Supplemental Information. *R* code used to estimate uncertainty by sampling strategy is available at `https://github.com/svsero/COVID19serosurveillance-Somerville`.

4 Results

4.1 Participant demographic characteristics and recruitment pathways

We recruited a total of 408 participants, of whom 10 were excluded from serological testing due to reported symptoms consistent with active SARS-CoV-2 infection (Table S1). Of the participants (n=228), 57% reported being recruited directly on-site (received a recruitment flier from study staff while entering or exiting the adjacent business location), 33% reported learning about the study from a friend or relative ("word of mouth"), and 7% reported learning about the study via internet or social media.

The demographic characteristics of the study participants are broadly consistent with demographic data from Somerville and other Boston-area communities [9, 10]. Individuals aged 20-29 were the largest age group in the study population, consistent with the large population of young adults in Somervillle [9]. The majority of participants (81%, n=221) reported speaking English as their primary language at home; 4% (n=16) and 6% (n=25) reported Spanish and Portuguese, respectively, as their primary household languages. The majority of participants reported living in households with $\leq$3 occupants (Table S1).

4.2 Geographic distribution of study participants' home locations

Of the 228 participants recruited on-site, 102 reported living in Somerville; 78 in the neighboring city of Cambridge, MA; 28 in other locations within 10 km of Somerville; and 20 in more distant locations in Massachusetts. Within Somerville, the geographic distribution of participants' reported home locations closely matched the distribution of GPS-estimated home locations for visitors to the study venue (Pearson's $r = 0.8956$, $p = 0.0131$, Figure 1A-B and Figure 2A).

We next evaluated how well the geographic distribution of participants in our study matched observed cumulative incidence of reported PCR-confirmed COVID-19 cases by ward. The cumulative incidence of PCR-confirmed SARS-CoV-2 infections reported to the Somerville Board of Health and the proportion of SARS-CoV-2 RT-PCR tests with positive results were heterogeneous across electoral wards in Somerville, and both highest in Ward 1 (Figure 1G and H, respectively). The total number of reported SARS-CoV-2 PCR tests conducted in each ward as a proportion of the respective ward population was relatively even across Somerville (Figure 1I). The home location distribution of directly-recruited participants (p_{direct}, Figure 2B) and the home location distribution of all study participants (p_{all}, Figure S3) were not correlated with cumulative incidence of PCR-confirmed cases by ward (r = -0.33, p = 0.78 and r = -0.45, p=0.88, respectively). Wards with higher cumulative incidence of PCR-confirmed cases (Wards 1 and 4 in East Somerville) were under-sampled in our study. The predicted catchment distribution of the study location was also poorly correlated with both cumulative incidence of reported infections (r= -0.1126, p = 0.5413, Figure 2C) and test positivity rates (Figure 1).

Lastly, we evaluated whether choosing an alternative study site could improve the correlation between sampling intensity (number of participants recruited for serological testing by ward) and cumulative incidence of PCR-confirmed infections by ward. To do this, we generated the distribution of GPS-estimated home locations for visitors to an alternative recruitment site, specifically a different essential business located in Somerville Ward 1 (the ward with highest observed cumulative incidence of PCR-confirmed infections). Notably, this alternative site home location distribution was strongly correlated with ward-level cumulative incidence of PCR-confirmed infections (r=0.9299, p=0.0072, Figure 2D).

Examining the distributions of home locations for study participants and visitors, we observed clear decay with distance for the home location distributions of all study participants and directly-recruited study participants, and for the GPS-estimated home location distributions for the actual and alternative study sites (p_{all} and p_{direct}, v_{site}, and v_{alt} in Figures S1 and S2). This distribution was slightly more long-tailed for GPS-estimated home location distributions (v_{site} and v_{alt}), indicating that the mobile data-derived catchment distribution captured more visitors from more distant locations when compared to the geographic distribution of study participants' reported home locations, but these more distant locations contributed only a small proportion of the estimated total visitors to each location. Excluding wards that contribute <1% of the total visitor counts constrained the GPS-estimated catchment distributions to local geographic areas around the study site.

4.3 Uncertainty in serology-based cumulative incidence estimates under different sampling strategies and venue location choice

Using the stochastic model described in the Methods, we evaluated relative levels of uncertainty in serology-based estimates of COVID-19 cumulative incidence (as measured by the width of the 95% confidence interval) for different sampling strategies. Sampling guided by observed geographic heterogeneity in PCR-confirmed incident cases ($\mathbf{S_C}$, described in the Supplemental Methods) yielded overall levels of uncertainty across m and n values that were similar to uniform sampling, $\mathbf{S_U}$ (Figure 3). Estimated uncertainty for venue-based sampling strategies, which we modeled here using both the observed distribution of participants' home locations in our study ($\mathbf{S_P}$) and GPS-estimated catchment distributions for an alternate study location ($\mathbf{S_V}$), was strongly influenced by the choice of venue and its corresponding geographic catchment distribution. For $\mathbf{S_P}$, uncertainty was high across all values of m and n, and approximately two-fold greater than uncertainty estimated for other sampling strategies (Figure 3C), indicating that under-sampling in locations with higher cumulative incidence and relatively over-sampling in areas with lower cumulative incidence (Figure 2) resulted in relatively higher levels of uncertainty, even with larger sample sizes. $\mathbf{S_V}$, in which the number of tests performed in each ward corresponds to the catchment distribution of visitors to an alternative study site in

Somerville Ward 1, yielded overall levels of uncertainty that were comparable to $\mathbf{S_C}$ and $\mathbf{S_U}$. We observed additional but smaller reductions in uncertainty when each sampling strategy was adjusted for the population age distribution (Figures S4 and S5), although modeled uncertainty remained high when the geographic distribution of participants was mismatched versus the underlying cumulative incidence of infection ($\mathbf{S_{PA}}$ in Figure S4).

4.4 COVID-19 symptom histories, self or provider diagnosis, case contacts, and risk activities

Just over half of participants (51%; n=201) reported having any of 15 symptoms of COVID-19 in the 12 weeks prior to the study, whereas only 9% (n=37) reported more specific symptoms of fever plus cough over the same time period (Table S1). A total of 86 (22%) of participants endorsed a self-diagnosis of COVID-19 (answered yes when asked "Do you think you have or have had COVID-19 infection?") and 15 participants (4%) reported previously receiving a suspected or PCR-confirmed COVID-19 diagnosis from a medical provider. Thirty percent of participants (n=120) reported working outside their homes in the week prior to the study; 37% (n=148) reported visiting friends or family outside of their household and 11% (n=45) reported using public transportation over the same time period.

4.5 Serology-estimated cumulative incidence of prior SARS-CoV-2 infection

Estimated cumulative incidence of SARS-CoV-2 infection, based on LFA serology results and corrected for test performance characteristics, was 0.113 (95% credible interval: 0.081-0.148) for the entire group of study participants (Table S2) and 0.130 (95%CI: 0.087-0.178) among participants directly recruited on-site (Table S3). We found no observable differences in estimated cumulative incidence across location (Somerville electoral ward), age, household size, or reported prior symptoms (either for participants reporting cough and fever in the last 12 weeks or for participants reporting any of 15 possible symptoms over the same time period). Cumulative incidence was significantly higher among the small number of participants who reported speaking Spanish as their primary household language (n=16, cumulative incidence: 0.440, 95%CI: 0.231-0.656) compared to those from English- (n=321) or Portuguese-speaking (n=25) households (Tables S2 and S4); this finding was not observed when the analysis was restricted to the smaller number of directly-recruited participants (Tables S3 and S5). COVID-19 self-diagnosis was a significant predictor of LFA seropositivity in both the entire dataset of all participants (OR 4.39, 95%CI: 1.15-15.09, Table S4) and the dataset including only directly recruited participants (OR 3.20, 95%CI 1.18-8.26, Table S5).

5 Discussion

SARS-CoV-2 serological surveillance data collected via convenience sampling will have continued roles in describing COVID-19 epidemiology, including: in rapid local assessment of cumulative incidence; in recruiting participants who may be difficult to reach via structured sampling; and in retrospective efforts aimed at understanding how the epidemic unfolded across communities. Serosurveillance studies based on structured probability samples [11, 12, 13, 14] will provide more rigorous and reliable estimates of SARS-CoV-2 cumulative incidence than convenience sampling [1, 2, 3], but such studies are not always feasible. Here, we performed a venue-based serosurveillance study in Somerville, MA and showed how accounting for sampling intensity by geographic location can improve uncertainty in COVID-19 incidence estimates from a convenience sample. Within this small sample, with important limitations discussed below, we observed 11-13% cumulative incidence of prior COVID-19 infection (adjusted for test performance).

Certain forms of convenience sampling may be better suited than structured surveys for reaching epidemiologically important demographics during the COVID-19 pandemic. For example, lower-wage or frontline workers who are at higher risk of SARS-CoV-2 exposure [15, 16, 14] may be less likely to participate if recruited using conventional survey outreach methods (e.g., mail or phone contact) due to constraints on their time [17, 18] and lack of incentives [17]. Convenience sampling at highly-visited community locations, including essential business, may be an attractive alternative to structured sampling in this important population, similar to venue-based sampling approaches developed to study so-called "hidden populations" [19].

Ultimately, the design of SARS-CoV-2 serosurveillance studies must balance these potential benefits against the biases and limitations on generalizability that are inherent to convenience sampling.

Geographic heterogeneity in SARS-CoV-2 cumulative incidence within cities has been repeatedly observed [13, 20, 21], with areas of lower socioeconomic status hit hardest by COVID-19. In our study, conducted in an area with clear geographic differences in the cumulative incidence of PCR-confirmed SARS-CoV-2 infections between wards, convenience sampling resulted in an under-sampling of the communities with the highest incidence of cases and an over-sampling of low incidence areas. The uncertainty of cumulative incidence estimates based on a modeled sampling strategy that closely approximates the sample in our venue-based study was approximately two-fold higher than the estimated uncertainty for other sampling strategies that more closely match the underlying geographic distribution of prior infections. We found that adjusting sampling weights for demographics decreased uncertainty, but this reduction was smaller than the reductions observed when using sampling strategies that more closely matched the geographic distribution of prior infections.

The observation that uncertainty in serology-based cumulative incidence estimates can be influenced by the underlying geographic distribution of previously-infected individuals followed from recent work [5], in which the optimal variance-reducing number of participants n_i for a given demographic group i was proportional to the group's relative size in the population d_i and its true cumulative incidence of infection θ_i. It follows that both d_i and θ_i can refer to not only demographic groups, but also to groups stratified by both demographics and location, as was done in the stochastic simulations here.

These observations have four important applications to design and interpretation of SARS-CoV-2 serosurveillance studies. First, serosurveillance studies employing convenience sampling can benefit from data on participants' home locations. The spatial resolution of this data should be sufficient to capture underlying geographic heterogeneity in cumulative incidence, while still ensuring that participants remain non-identifiable.

Second, information on sampling intensity by geographic unit, and corresponding public health data on cumulative incidence of reported or confirmed SARS-CoV-2, should be included when reporting results from convenience samples ([2]). This information is important for understanding serosurveillance studies in context, particularly for evaluating the reliability and limitations of serology-based cumulative incidence estimates in areas with known or suspected heterogeneity in the geographic distribution of prior infections.

Third, in contexts where case detection and testing efforts are relatively uniform, such that locations with relatively higher and lower cumulative incidence can be identified reliably, weighted sampling based on the expected underlying distribution of prior infections by location and demographic group can provide a useful strategy for reducing uncertainty in serology-based cumulative incidence estimates. Conversely, this strategy is likely not suitable for situations in which case detection effort is non-uniform across locations, such that underlying relative distribution of prior infections cannot be reliably ascertained from available public health data.

Fourth, for venue-based studies, GPS-estimated foot traffic data can provide important information to guide the selection of the study venue. The predicted catchment distribution for visitors to our study venue correlated closely with the distribution of participants' home locations, but the strength of this correlation, and the utility of GPS-estimated foot traffic for this application, is likely variable across locations. Moreover, the utility of GPS-estimated foot traffic data is expected to be strongly influenced by superimposed participation biases (as described below). At a minimum, this data source can be useful to rule out candidate study venues with predicted catchment areas that are obviously mismatched to the anticipated underlying geographic distribution of prior infections.

Multiple considerations are important for contextualizing our findings and the above recommendations. The GPS-estimated foot traffic data used in our study have several important limitations. Identification of visitors and their home locations in this data source can be biased by differences in mobile device usage between demographic groups, potentially under-sampling visitors from important populations or oversampling others. In addition, this data can be sparse and thus more subject to stochastic variation when only small numbers of users are captured in either the point of interest or home location. Different forms of participation bias are expected to skew the geographic distribution of study participants away from GPS-estimated catchment distributions, potentially making it difficult in practice to capture geographically-representative samples via venue-based sampling. Our study was subject to at least one form of participation bias, "information-seeking" or volunteer bias by individuals with prior COVID-19 self-diagnosis; this source of bias was likely enabled or compounded by the fact that all participants received the results of LFA serological testing, creat-

ing an incentive for information-seeking participants (e.g., those who believe they previously had COVID-19 or those with a prior episode of unexplained illness). Studies in which participants do not receive their serology results, if acceptable from an ethical and community involvement perspective, may be advisable to reduce this type of bias.

The relatively equal testing intensity (number of PCR tests conducted per resident) across Somerville, combined with non-uniform rates of test positivity by ward (proportion of all PCR tests with a positive result), are indicative of gap in testing effort in more highly impacted areas of the city. Higher test positivity, an indicator of higher epidemic intensity, is not appropriately matched with higher testing intensity.[22] Thus, our approach to modeling uncertainty in serology-based cumulative incidence estimates likely mis-specifies the assumed true distribution of infections in each age-location group ($\mathbf{\Theta}$ in Figure 3). The assumed true underlying cumulative incidences in $\mathbf{\Theta}$, which are specified using the observed number of positive PCR results in each ward (and do not account for the testing gap described above), are less dispersed across age-location groups than what would be expected if PCR testing effort better matched epidemic intensity by ward (per test positivity). However, this mis-specification, and resultant smaller dispersion in assumed true cumulative incidence by ward, is expected to result in more conservative estimates for uncertainty by sampling strategy.

Although we observed substantially higher serology-based cumulative incidence among individuals from Spanish-speaking households, compared to English- and Portuguese-speaking households, this finding was based on small sample sizes, both for the overall study and for the number of participants with Spanish as their household language. Larger studies, including those powered and stratified to examine social determinants of SARS-CoV-2 exposure [13, 14], are needed to advance our understanding of how race, ethnicity, location, economic status, occupation, and mobility intersect to drive the deep health disparities observed during the epidemic.

In summary, we evaluated the seroprevalence of COVID-19 through a convenience sample at a venue in Somerville, MA and showed how accounting for geographic heterogeneity in COVID-19 cumulative incidence can improve serosurveillance estimates. Our findings are relevant to studies employing venue-based recruitment and are also applicable to other kinds of convenience sampling, for example, studies using a hospital's discarded blood specimens from patients drawn from the hospital's geographic catchment area. We argue that GPS-estimated foot traffic data provides useful information for evaluating candidate study sites for venue-based studies, with certain limitations related to bias and uncertainty in this data source. We underscore the importance of collecting data on participants' home locations for understanding the resulting population-level serological data, with ultimate goal of improving the reliability and interpretability of SARS-CoV-2 serosurveillance studies.

6 Acknowledgements

We would like to acknowledge support from Tia Hira, the Infectious Diseases Division and Department of Pathology at Massachusetts General Hospital, City of Somerville Health and Human Services and SomerStat, the Somerville business community, and the study staff and volunteers. This work was supported by the Morris-Singer Foundation and NIH T32AI007061 (TSB).

References

[1] Eli S. Rosenberg, James M. Tesoriero, Elizabeth M. Rosenthal, Rakkoo Chung, Meredith A. Barranco, Linda M. Styer, Monica M. Parker, Shu-Yin John Leung, Johanne E. Morne, Danielle Greene, David R. Holtgrave, Dina Hoefer, Jessica Kumar, Tomoko Udo, Brad Hutton, and Howard A. Zucker. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. *Annals of Epidemiology*, 48:23 – 29.e4, 2020.

[2] Fiona P. Havers, Carrie Reed, Travis Lim, Joel M. Montgomery, John D. Klena, Aron J. Hall, Alicia M. Fry, Deborah L. Cannon, Cheng-Feng Chiang, Aridth Gibbons, and et al. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. *JAMA Internal Medicine*, Jul 2020.

[3] Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra-Walker, James Tedrow, Dona Tversky, Andrew Bogan, Thomas Kupiec, Daniel Eichner, Ribhav Gupta, John Ioannidis, and Jay Bhattacharya. COVID-19 Antibody Seroprevalence in Santa Clara County, California. *medRxiv*, 2020.

[4] Bonnie E Shook-Sa, Ross M Boyce, and Allison E Aiello. Estimation Without Representation: Early Severe Acute Respiratory Syndrome Coronavirus 2 Seroprevalence Studies and the Path Forward. *The Journal of Infectious Diseases*, 222(7):1086–1089, Jul 2020.

[5] Daniel B. Larremore, Bailey K. Fosdick, Kate M. Bubar, Sam Zhang, Stephen M. Kissler, C. Jessica E. Metcalf, Caroline Buckee, and Yonatan Grad. Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. *medRxiv*, 2020.

[6] SafeGraph. SafeGraph Neighborhood Patterns. 2020. `https://www.safegraph.com/`.

[7] QGIS Geographic Information System. Open source geospatial foundation project. 2020. `https://qgis.org`.

[8] Massachusetts State Department of Health. COVID-19 Response Reporting. `https://www.mass.gov/info-details/covid-19-response-reporting`.

[9] City of Somerville, Massachusetts and Cambridge Health Alliance. The Wellbeing of Somerville Report. `https://www.somervillema.gov/`.

[10] City of Cambridge, Massachusetts. City of Cambridge Neighborhood Statistical Profile. `https://www.cambridgema.gov/`.

[11] Holly M. Biggs, Jennifer B. Harris, Lucy Breakwell, F. Scott Dahlgren, Glen R. Abedi, Christine M. Szablewski, Jan Drobeniuc, Nirma D. Bustamante, Olivia Almendares, Amy H. Schnall, and et al. Estimated Community Seroprevalence of SARS-CoV-2 Antibodies — Two Georgia Counties, April 28–May 3, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69(29):965–970, Jul 2020.

[12] Nir Menachemi, Constantin T. Yiannoutsos, Brian E. Dixon, Thomas J. Duszynski, William F. Fadel, Kara K. Wools-Kaloustian, Nadia Unruh Needleman, Kristina Box, Virginia Caine, Connor Norwood, and et al. Population Point Prevalence of SARS-CoV-2 Infection Based on a Statewide Random Sample — Indiana, April 25–29, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69(29):960–964, Jul 2020.

[13] Amy K. Feehan, Daniel Fort, Julia Garcia-Diaz, Eboni Price-Haywood, Cruz Velasco, Eric Sapp, Dawn Pevey, and Leonardo Seoane. Seroprevalence of SARS-CoV-2 and Infection Fatality Ratio, Orleans and Jefferson Parishes, Louisiana, USA, May 2020. *Emerging Infectious Diseases*, 26(11), Nov 2020.

[14] Amy K. Feehan, Cruz Velasco, Daniel Fort, Jeffrey H. Burton, Eboni Price-Haywood, Peter T. Katzmarzyk, Julia Garcia-Diaz, and Leonardo Seoane. Racial and workplace disparities in seroprevalence of SARS-CoV-2 in Baton Rouge, Louisiana, July 15-31, 2020. *medRxiv*, 2020.

[15] Devan Hawkins. Social Determinants of COVID-19 in Massachusetts, United States: An Ecological Study. *Journal of Preventive Medicine and Public Health*, 53(4):220–227, Jul 2020.

[16] Marissa G. Baker, Trevor K. Peckham, and Noah S. Seixas. Estimating the burden of United States workers exposed to infection or disease: A key factor in containing risk of COVID-19 infection. *PLOS ONE*, 15(4):e0232452, Apr 2020.

[17] D.M. Corbie-Smith. Minority recruitment and participation in health research. *North Carolina medical journal*, 2004.

[18] J. F. Keyzer, J. Melnikow, M. Kuppermann, S. Birch, C. Kuenneth, J. Nuovo, R. Azari, D. Oto-Kent, and M. Rooney. Recruitment strategies for minority participation: challenges and cost lessons from the POWER interview. *Ethn Dis*, 15(3):395–406, 2005.

[19] Farzana B. Muhib, Lillian S. Lin, Ann Stueve, Robin L. Miller, Wesley L. Ford, Wayne D. Johnson, and Philip J. Smith. A Venue-Based Method for Sampling Hard-to-Reach Populations. *Public Health Reports*, 116:216–222, Jan 2001.

[20] Sage J. Kim and Wendy Bostwick. Social Vulnerability and Racial Inequality in COVID-19 Deaths in Chicago. *Health Education Behavior*, 47(4):509–513, May 2020.

[21] S. Kissler, N. Kishore, M. Prabhu, D. Goffman, Y. Beilin, R. Landau, C. Gyamfi-Bannerman, B.T. Bateman, D. Katz, J. Gal, A. Bianco, J. Stone, D. Larremore, C.O. Buckee, and Y.H. Grad. Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City. May 2020. `https://dash.harvard.edu/handle/1/42665370`.

[22] Scott Dryden-Peterson, Gustavo E. Velásquez, Thomas J. Stopka, Sonya Davey, Shahin Lockman, and Bisola Ojikutu. Sars-cov-2 testing disparities in massachusetts. *medRxiv*, 2020.

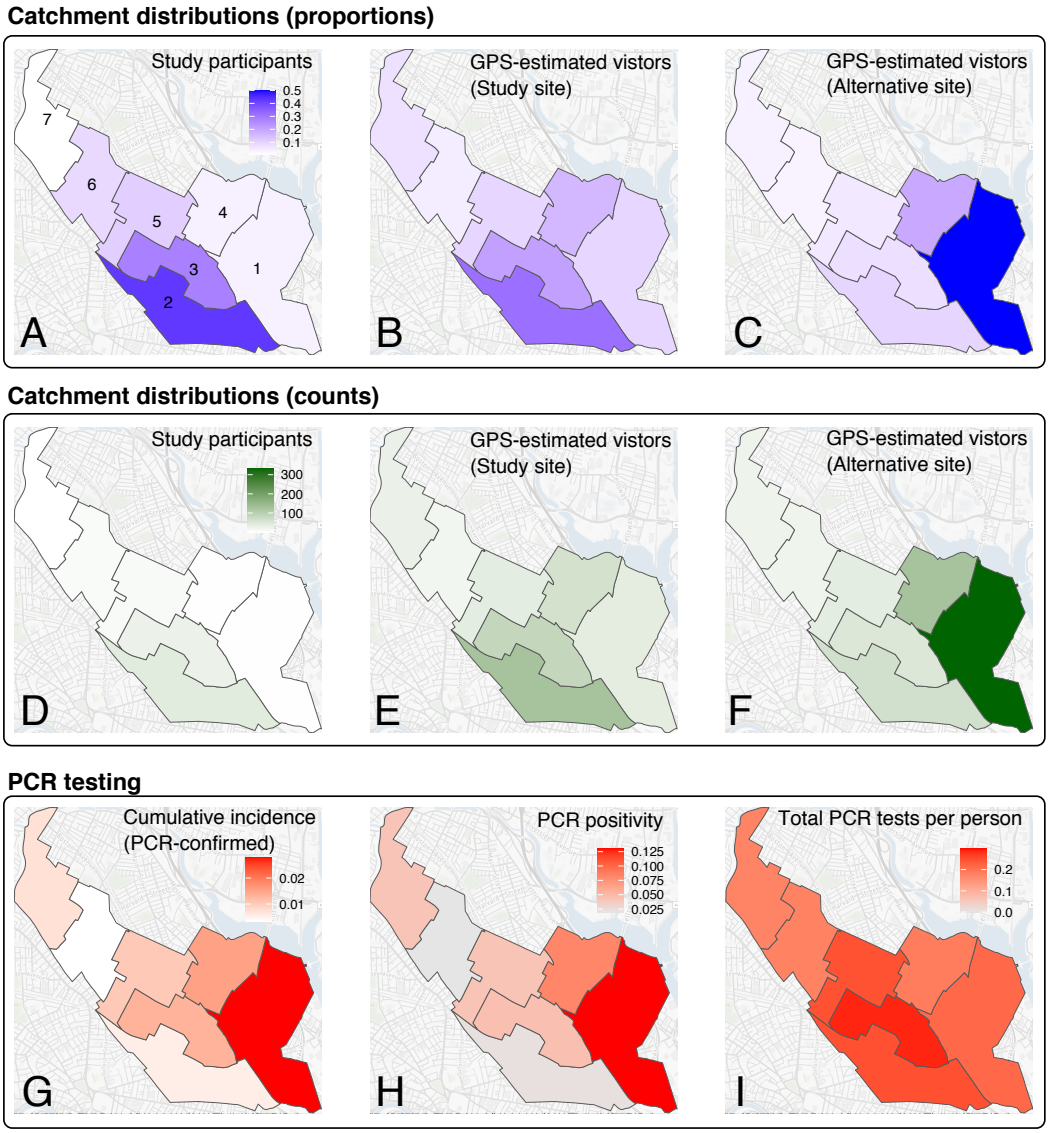

Figure 1: Catchment distributions for participants and GPS-estimated visitors, reported COVID-19 disease activity, and laboratory confirmed cases for Somerville, Massachusetts. (A) Reported home locations for directly-recruited study participants by Somerville electoral ward, reported as a proportion of all participants reporting home locations in Somerville (p_{direct}). Wards are labeled by number. (B) GPS-estimated home locations for daily visitors to the study locations by ward, reported as the proportion of all visitors with GPS-estimated home locations in Somerville (v_{site}). (C) GPS-estimated home locations for daily visitors to an alternate hypothetical study site by ward, reported as the proportion of all visitors to the alternate site with GPS-estimated home locations in Somerville (v_{alt}). (D) Number of directly recruited participants by Somerville electoral ward. (E) Number of GPS-estimated visitors to the study site, aggregated by estimated home ward location. (F) Number of GPS-estimated visitors to the alternate, hypothetical study site, aggregated by estimated home ward location. (G) Cumulative incidence of confirmed SARS-CoV2 infections by ward (as of June 8th,2020), calculated as the number of unique positive SARS-CoV-2 PCR tests divided by the total population of each ward. (H) Proportion of all SARS-CoV-2 PCR tests with positive results, aggregated by ward (I) SARS-CoV-2 PCR testing effort by ward, reported as the total number of unique tests conducted in each ward divided by total ward population.

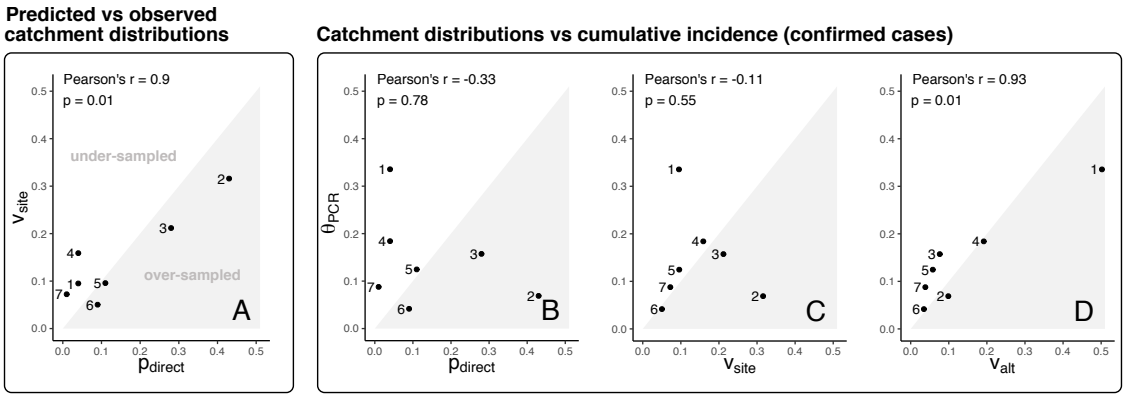

Figure 2: Comparison between observed catchment distribution of study participants (p_{direct}), GPS-estimated catchment distributions for visitors to the study site and an alternate hypothetical study site (v_{site} and v_{alt}), and observed incidence of PCR-confirmed SARS-COV-2 infections reported to the Somerville Board of Health. Points are annotated by Somerville ward number. Grey and white regions of each plot demarcate wards that are relatively over- and under-sampled with respect to observed cumulative incidence of confirmed infections (B-D) or the GPS-estimated proportion of visitors from a given ward (A). P values are estimated by comparing the observed r value versus a null distribution obtained by permutation of ward assignment for each variable (n=10,000 permutations).

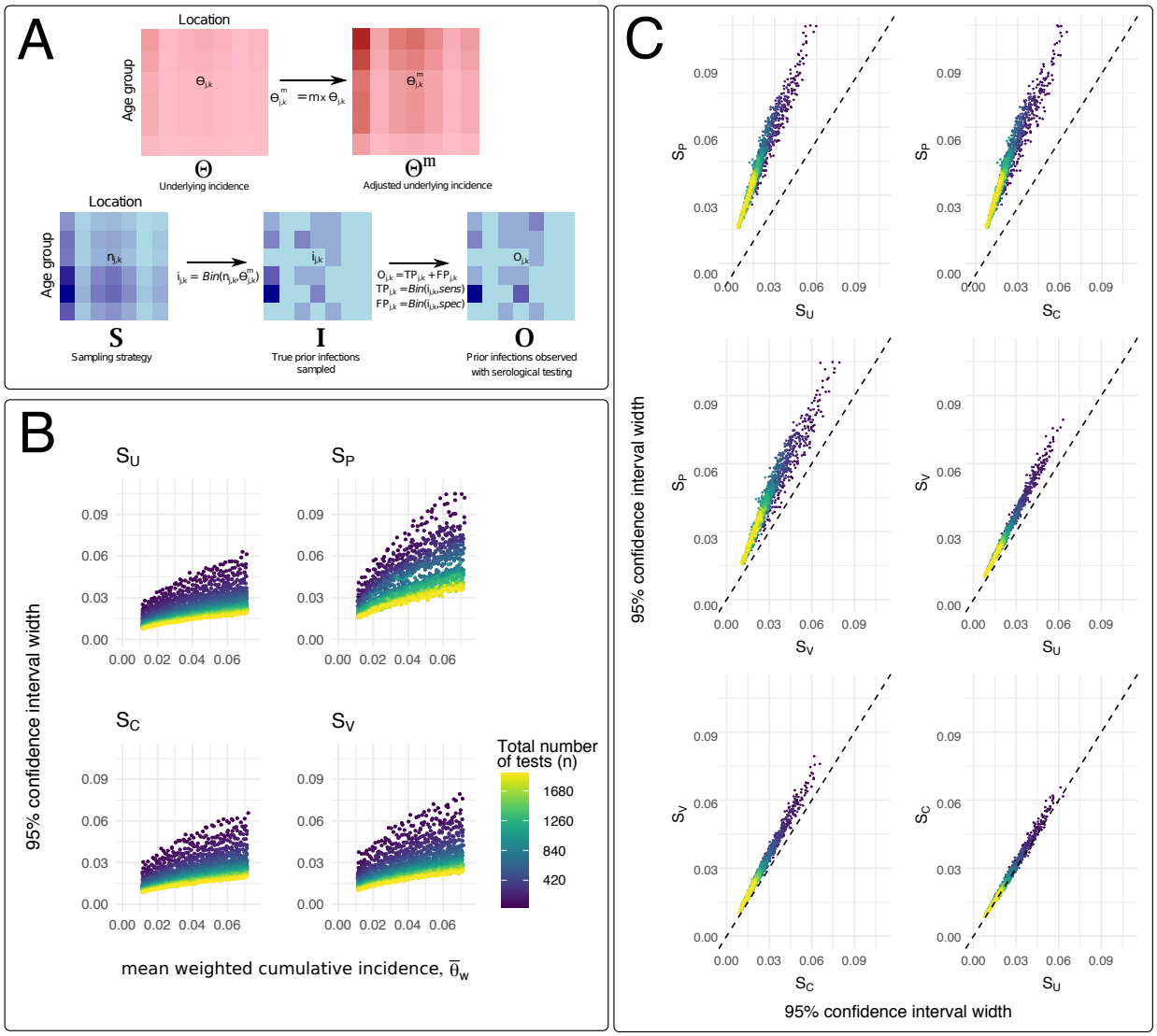

Figure 3: Sampling strategy and uncertainty in SARS-CoV-2 serosurveillance studies. (A) Top: Model assumptions for assumed underlying incidence by age-location group, where entries of matrix $\mathbf{\Theta}$, $\theta_{j,k}$, are the estimated cumulative incidence of detected, confirmed cases (per PCR-confirmed cases reported to the City of Somerville) for individuals in location j and age group k. Different underlying epidemic sizes are given by $\mathbf{\Theta^m}$, where $\theta^m_{j,k} = m \times \theta_{j,k}$ and m is a multiplier approximating the factor by which true infections exceed detected, PCR-confirmed infections. Bottom: Procedures for estimating uncertainty in serology-based cumulative incidence using different sampling strategies, $\mathbf{S}$ (as described in Methods and Supplemental Information). The procedure in (A) is repeated 1,000 times for each value of m and n. (B) Uncertainty (measured as the width of the 95% confidence interval) versus mean value for population-weighted cumulative incidence, for different numbers of total tests (n) and relative epidemic sizes (m). Each point represents the results from 1000 stochastic simulation and is colored according to the total number of tests specified for each simulation. $\mathbf{S_U}$: uniform sampling across age-location groups; $\mathbf{S_C}$ sampling weighted by the observed cumulative incidence of PCR-confirmed infection by ward; $\mathbf{S_P}$: sampling weighted by the observed number of study participants by ward; $\mathbf{S_V}$: sampling weighted by the GPS-estimated catchment distribution of visitors to an alternative study location in Somerville Ward 1. (C) Relative uncertainty, as measured by the width of the 95% confidence interval for $\bar{\theta}_w$, compared between sampling strategies. Dashed line at intercept = 0 and slope = 1 demarcates where uncertainty for the two sampling strategies under consideration are equal to each other. Each point represents the results from 1000 stochastic simulation and is colored according to the total number of tests specified for each simulation.