BMJ Open

Geographically skewed recruitment and COVID-19 seroprevalence estimates: a cross-sectional serosurveillance study and mathematical modelling analysis

Tyler Brown ,[1,2,3] Pablo Martinez de Salazar Munoz,[2] Abhishek Bhatia,[4] Bridget Bunda,[1] Ellen K Williams,[5] David Bor,[3,6] James S Miller,[7] Amir Mohareb ,[1,3] Julia Thierauf,[3,8] Wenxin Yang,[8] Julian Villalba,[1,3,8] Vivek Naranbai,[3,9] Wilfredo Garcia Beltran,[3,8] Tyler E Miller,[3,8] Doug Kress,[10] Kristen Stelljes,[10] Keith Johnson,[10] Dan Larremore,[11] Jochen Lennerz,[3,8] A John Iafrate,[3,8] Satchit Balsari,[3,4] Caroline Buckee,[2] Yonatan Grad [2,3]

To cite: Brown T, de Salazar Munoz PM, Bhatia A, *et al*. Geographically skewed recruitment and COVID-19 seroprevalence estimates: a cross-sectional serosurveillance study and mathematical modelling analysis. *BMJ Open* 2023;**13**:e061840. doi:10.1136/bmjopen-2022-061840

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-061840).

CB and YG are joint senior authors.

Received 17 February 2022
Accepted 26 January 2023

For numbered affiliations see end of article.

Correspondence to
Tyler Brown;
tsbrown@mgh.harvard.edu

ABSTRACT

Objectives Convenience sampling is an imperfect but important tool for seroprevalence studies. For COVID-19, local geographic variation in cases or vaccination can confound studies that rely on the geographically skewed recruitment inherent to convenience sampling. The objectives of this study were: (1) quantifying how geographically skewed recruitment influences SARS-CoV-2 seroprevalence estimates obtained via convenience sampling and (2) developing new methods that employ Global Positioning System (GPS)-derived foot traffic data to measure and minimise bias and uncertainty due to geographically skewed recruitment.

Design We used data from a local convenience-sampled seroprevalence study to map the geographic distribution of study participants' reported home locations and compared this to the geographic distribution of reported COVID-19 cases across the study catchment area. Using a numerical simulation, we quantified bias and uncertainty in SARS-CoV-2 seroprevalence estimates obtained using different geographically skewed recruitment scenarios. We employed GPS-derived foot traffic data to estimate the geographic distribution of participants for different recruitment locations and used this data to identify recruitment locations that minimise bias and uncertainty in resulting seroprevalence estimates.

Results The geographic distribution of participants in convenience-sampled seroprevalence surveys can be strongly skewed towards individuals living near the study recruitment location. Uncertainty in seroprevalence estimates increased when neighbourhoods with higher disease burden or larger populations were undersampled. Failure to account for undersampling or oversampling across neighbourhoods also resulted in biased seroprevalence estimates. GPS-derived foot traffic data correlated with the geographic distribution of serosurveillance study participants.

Conclusions Local geographic variation in seropositivity is an important concern in SARS-CoV-2 serosurveillance studies that rely on geographically skewed recruitment strategies. Using GPS-derived foot traffic data to select recruitment sites and recording participants' home locations can improve study design and interpretation.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ We conducted a local convenience-sampled seroprevalence study that captured neighbourhood-level data on participants' home locations, allowing us to map the geographic sampling distribution of individuals in our study.
⇒ We used this data, combined with Global Positioning System-estimated business foot traffic data and local public health data on confirmed COVID-19 cases, to inform a mathematical model that examines bias and uncertainty in seroprevalence estimates derived from geographically skewed sampling distributions.
⇒ Limitations of our study include uncertainty in modelling specifications, limitations in existing public health data (including disparities in COVID-19 testing and case detection across demographic and geographic groups), and uncertainty in the generalisability of our findings.

INTRODUCTION

Studies estimating SARS-CoV-2 seroprevalence have been critical to our understanding of the COVID-19 pandemic, particularly when diagnostic testing was limited and the extent of community spread was unknown.[1–4] Randomised representational surveys have proven challenging to conduct on local scales because of cost and logistical considerations. As such, many of these studies recruited participants via convenience sampling.[4–6] Their estimates are subject to multiple sources of bias and uncertainty inherent to non-randomised sampling.[7] However, cost and logistical considerations continue to motivate

the use of convenience sampling in COVID-19 serosurveillance efforts[8] and similar studies may have continued utility in the public health response to the pandemic. Understanding these sources of bias and uncertainty and finding ways to measure and account for them are thus important practical goals.

Geographic variation in sampling intensity is inherent to most kinds of convenience sampling. For example, in venue-based (or 'walk-up') studies, in which participants are recruited from among visitors to a central or highly trafficked location,[4 9] the geographic distribution of participants is expected to be skewed towards individuals living closer to the study location. Similarly, discarded blood samples reflect the catchment area of a given hospital or clinical laboratory, which may strongly constrain the geographic distribution of samples available for analysis.[5]

COVID-19 burden is markedly heterogeneous within cities and between neighbourhoods.[1–3] Vaccination coverage is likewise variable over small geographic areas.[10] SARS-CoV-2 seropositivity is thus expected to vary across even relatively small study areas, including those used in local seroprevalence studies. Given this context, geographic variation in sampling, resulting in oversampling or undersampling from areas with relatively higher or lower underlying seropositivity, may have important ramifications for seroprevalence estimates. This source of bias and uncertainty remains poorly understood and has largely not been addressed in COVID-19 serosurveillance studies.

The objective of our work was to understand and quantitate how geographic variation in sampling intensity influences seroprevalence estimates derived from geographically skewed convenience samples. We focus on COVID-19 seroprevalence estimates obtained via venue-based sampling,[4 9] with direct applications to other forms of convenience sampling.[5 11 12] To do this, we used data from a local seroprevalence study we conducted in Somerville, Massachusetts to map geographic distributions of study participants and examine geographic biases in recruitment. Using a numerical model, we evaluated how geographic variation in sampling intensity influenced bias and uncertainty in seroprevalence estimates. Finally, we assessed the use of Global Positioning System (GPS)-derived foot traffic data, which estimates the home locations for daily visitors to a given location, as a tool for evaluating the expected geographic distribution of participants at candidate study sites. Our results offer an approach to improve the design and interpretation of seroprevalence studies that use venue-based convenience sampling with little if any impact on cost and speed.

METHODS

SARS-CoV-2 seroprevalence study design and participant information

We obtained participant demographic and home location data for 398 asymptomatic adults who underwent SARS-CoV-2 serological testing in Somerville, Massachusetts between 4 June and 9 June 2020 (approximately 6 weeks after the first wave of the COVID-19 epidemic peaked in Massachusetts[13]). Somerville is a diverse, densely populated city in the Greater Boston Metropolitan Area, covering 10.93 km^2 and home to approximately 81 175 residents in 2020, of whom 24.2% were born outside the USA.[14] The study took place after COVID-19 public health restrictions on certain activities had been lifted (including those on childcare facilities and some retail businesses), but when restrictions on indoor dining and other activities remained in place. Recruitment took place outside an essential business in Somerville that was not subject to any ongoing public health restrictions. We used an established Bayesian statistical method to adjust seroprevalence estimates, which incorporates both the data used to calculate the estimate and validation data used to measure serological test performance.[15]

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). The study recorded participant demographic information and self-reported home locations by postal code and electoral ward (Somerville has seven electoral wards, each of which covers approximately 1–2 km^2). This study also collected information on how participants learnt about the study in order to distinguish participants directly recruited on site at the study location from those who learnt about the study from friends, family or social media. We did not advertise or announce enrolment for the study prior to its implementation, with the goal of increasing the proportion of individuals recruited on site at the study location. We used this data to calculate P_j^{direct}, the proportion of all directly recruited participants from Somerville with self-reported home locations in each of the city's seven electoral wards (indexed with the subscript j). We refer to the full set of P_j^{direct} values, $\{P_1^{\text{direct}}, \cdots P_7^{\text{direct}}\}$, as the 'survey participant catchment distribution'. Additional information on study procedures and serological testing is included in the online supplemental material.

Patient and public involvement

We collaborated with local government leaders, public health officials and members of the local business community to design and implement the study.

Public health acute infection data

We obtained data on 916 PCR-confirmed COVID-19 cases with reported home addresses in Somerville (collected from the onset of the epidemic through June 2020) from the Massachusetts Virtual Epidemiologic Network (MAVEN). During the study period, home antigen testing was not yet available and daily counts for new, PCR-confirmed infections (sourced from hospital-based and private-sector laboratories and geographically aggregated by home addresses provided by patients) provided a

relatively complete daily record of individuals testing PCR positive for SARS-CoV-2 infection. To account for potential differences in testing effort, we also obtained from MAVEN the total number of residents in each Somerville electoral ward who were tested for SARS-COV-2 via RT-PCR from the start of the epidemic until the end of the study period in June 2020. Data were anonymised and aggregated by electoral ward prior to analysis. We calculated the cumulative incidence of PCR-confirmed infections by ward (θ_j^{PCR}) and the proportion of all PCR tests with positive results ('PCR positivity').

GPS-estimated business foot traffic

We used GPS-estimated foot traffic data (SafeGraph, *safegraph.com*) to determine the expected geographic distribution of visitors to different study recruitment sites. This data source provides approximate home locations, aggregated at the level of census block groups (CBG), for daily visitors to given locations of interest. CBGs are the second smallest geographic unit used in the USA census and are typically defined to have populations of 300 to 6000 people. We removed CBGs with low visitor counts and reaggregated data points to the level of electoral wards (online supplemental figures 1 and 2). We denoted the proportion of all GPS-estimated visitors to the actual study site who have home locations in electoral ward j as $V_{\mathrm{j}}^{\mathrm{site}}$. To assess an alternative study site, we calculated $V_{\mathrm{j}}^{\mathrm{alt}}$, the proportion of all GPS-estimated visitors to a hypothetical alternative recruitment location with home locations in electoral ward j. Thus, V_1^{site} is the proportion of total GPS-estimated visitors to the actual study site who have home locations in ward 1. We refer to the full set of V_{j} values as the 'GPS-estimated visitor catchment distributions'.

Simulations

We used a simple numerical simulation to examine bias and uncertainty in estimated seroprevalence (θ_{pop}) under different sampling conditions. Using demographic and SARS-CoV-2 acute infection data from Somerville, MA, we generated a simulated population with varying true seropositivity $\theta_{\mathrm{j,k}}$ across subgroups stratified by location j and age group k (where locations are electoral wards in Somerville). We specified the size of each subgroup using local census data[16] and specified the true underlying seropositivity for each age-location subgroup by assuming these values are proportional to the observed cumulative incidence of PCR-confirmed infections for each subgroup. The simulation randomly draws a specified number of individuals from each subgroup (denoted $n_{\mathrm{j,k}}$), calculates weighted population-level seroprevalence (adjusted for serological test performance) and repeats this process 10 000 times to generate distributions of θ_{pop} values. We report W, the width of the 95th percentile interval for each distribution, as an approximate measure of uncertainty for θ_{pop}.

We specified the number of individuals drawn from each subgroup $n_{\mathrm{j,k}}$ using three sampling allocation strategies: (1) optimal allocation, in which the number of individuals sampled from each age-location subgroup is specified to optimally reduce uncertainty in the resulting seroprevalence estimates (this optimal allocation strategy is detailed in the online supplemental material); (2) allocation following the observed survey participant catchment distribution for the actual study site; (3) allocation following the GPS-estimated visitor distribution at the hypothetical alternative study site. Additional details on this numerical model and a diagram of the overall modelling procedure (online supplemental figure 3) are available in the online supplemental material. *R* code for the numerical simulations is available at *https://github.com/svsero/COVID19serosurveillance-Somerville.*

RESULTS

Seroprevalence study results

Among directly recruited participants with home locations in Somerville, estimated prevalence of SARS-CoV-2 spike protein antibodies, corrected for test performance characteristics,[15] was 0.111 (95% credible interval: 0.057 to 0.174); we found no statistically significant differences in estimated seroprevalence across locations (Somerville electoral ward), age or household size (online supplemental table 1).

Study participant catchment distributions and neighbourhood-level variation in COVID-19 cases

We first examined how survey participant catchment distributions align with, or mismatch, the geographic distribution of seropositive individuals in a given study area. We observed that the survey participant catchment distribution in the Somerville serosurveillance study was skewed strongly towards locations near the study site (figure 1A). Among directly recruited participants with home locations in Somerville, 43% (43/100) reported home locations in ward 2 (where the study site was located) compared with 4% in ward 1 and 4% in ward 4. In contrast, the cumulative incidence of PCR-confirmed SARS-CoV-2 infections (θ_j^{PCR}) was approximately threefold higher in electoral ward 1 compared with wards 2 and 6 (figure 1B) and the proportion of SARS-CoV-2 PCR tests with positive results

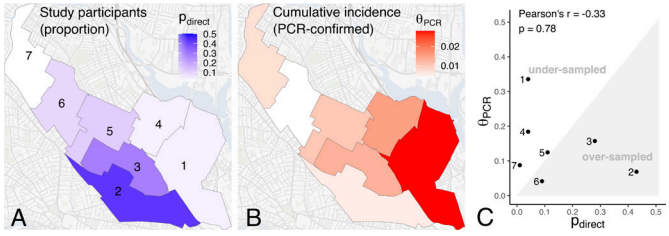

Figure 1 Sample allocation and geographic heterogeneity in proxy measures of epidemic intensity. (A) Survey participant catchment distribution. Wards are shaded by $P_{\mathrm{j}}^{\mathrm{direct}}$, the proportion of all directly recruited participants from each of Somerville wards 1–7; (B) cumulative incidence of prior PCR-confirmed SARS-CoV-2 infections by Ward as of June 8th, 2020 (θ_j^{PCR}); (C) correlation between $P_{\mathrm{j}}^{\mathrm{direct}}$ and θ_j^{PCR}. Significance of the correlation is calculated via permutation testing, as described in the online supplemental material.

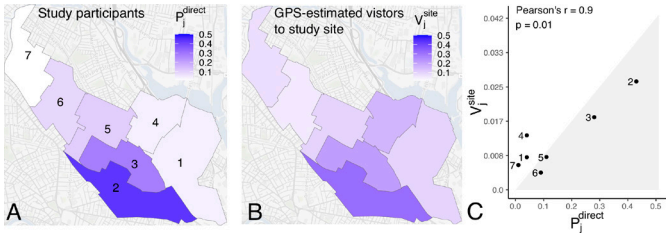

Figure 2 Observed study participant catchment distribution and GPS-estimated visitor catchment distribution for the study site. (A) Study participant catchment distribution for the study site (P_j^{direct}); (B) GPS-estimated visitor catchment distributions for the study site (V_j^{site}); (C) correlation between P_j^{direct} and V_j^{site} . Significance of the correlation is calculated via permutation testing, as described in the online supplemental material.

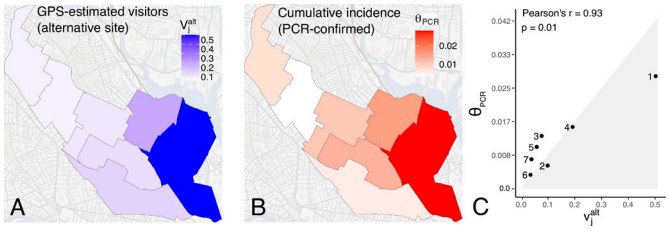

Figure 3 GPS-estimated visitor catchment distribution for the hypothetical alternative study site vs proxy measures of epidemic intensity. (A) GPS-estimated visitor catchment distribution for a hypothetical alternative study site in Somerville ward 1 (V_j^{alt}); (B) cumulative incidence of prior PCR-confirmed SARS-CoV-2 infections by ward (θ_j^{PCR}); (C) correlation between V_j^{alt} and the cumulative incidence of PCR-confirmed SARS-CoV-2 infection by electoral ward (θ_j^{PCR}).

was approximately fivefold higher (online supplemental figure 4A). Both of these proxy measures of epidemic intensity (θ_j^{PCR} and PCR test positivity) are limited by potential biases, some of which are likely to still be present even if PCR testing rates are relatively equal by ward.[17] Nonetheless, these measures suggest substantial heterogeneity in the underlying epidemic intensity, with an apparent higher rate of previously infected individuals (as a proportion of the population) in Somerville wards 1 and 4. Thus, we observed that the venue location chosen for this study, and its associated survey participant catchment distribution, resulted in relative undersampling of wards with expected higher seropositivity and oversampling of those with lower expected seropositivity (figure 1C).

GPS-estimated visitor catchment distributions

Recognising that survey participant catchment distributions can be poorly matched to the underlying geographic distribution of seropositivity, we explored the use of GPS-estimated foot traffic data as a tool for evaluating actual or candidate locations for venue-based sampling. We found that the participant catchment distribution at the actual study site (P_j^{direct}) closely matched its corresponding GPS-estimated visitor catchment distribution, V_j^{site} (Pearson's r = 0.90, p =0.0131, figure 2).

We next evaluated whether choosing an alternative study site could improve the correlation between sample allocation and cumulative incidence of PCR-confirmed infections by ward. We observed that the GPS-estimated visitor catchment distribution at the alternative site (V_j^{alt}) is strongly correlated with ward-level cumulative incidence of PCR-confirmed infections (r=0.93, p=0.0072, figure 3).

Venue location, sampling allocation and uncertainty in SARS-CoV-2 seroprevalence estimates

We used numerical simulation to quantify uncertainty in estimated SARS-CoV-2 seroprevalence under different survey participant catchment distributions. We observed 1.5-fold to twofold higher uncertainty when sampling effort was allocated according to the participant catchment distribution at the study site compared with the alternative site or optimal allocation (figure 4). This observation indicates that choice of recruitment location can result in suboptimal sample allocation and higher uncertainty.

The optimal sampling allocation strategy (which follows from the well-known Neyman allocation,[18] as explained in the online supplemental material) depends on both the size of each subgroup and its underlying seropositivity. Thus, if subgroup sizes are known and differences in subgroup-level seropositivity can be inferred or assumed, allocating more samples to larger subgroups and those with higher expected seropositivity will improve precision for weighted population-level seroprevalence estimates (online

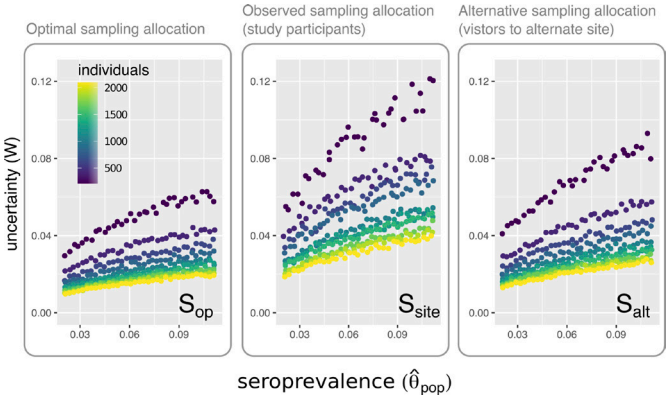

Figure 4 Uncertainty in estimated SARS-CoV-2 seroprevalence obtained using different sample allocation strategies. Left panel: the uncertainty (W, the width of the 95th percentile interval for 10 000 estimated seroprevalence values) vs mean estimated seroprevalence for different values of n (the total number of individuals sampled) when individuals are sampled according to the optimal sample allocation described in the online supplemental material (S_{op}); Centre panel: uncertainty (W) vs mean estimated seroprevalence when n individuals are sampled according to the observed study participant catchment distribution from the Somerville seroprevalence survey (S_{site}); right panel: uncertainty (W) vs mean estimated seroprevalence when n individuals are sampled according to the GPS-estimated catchment distribution at the hypothetical alternative study site (S_{alt}).

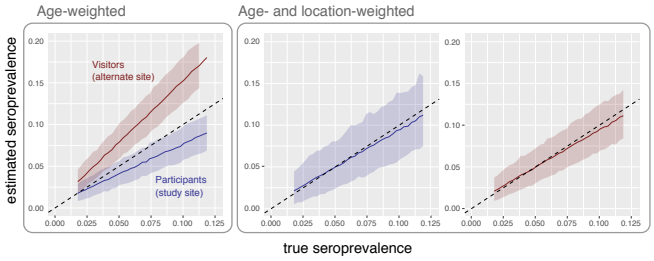

Figure 5 Bias in estimated seroprevalence by sampling strategy and weighting procedures. Left: estimated seroprevalence, weighted only by age subgroups, vs true seroprevalence. Right: estimated seroprevalence, weighted by age-location subgroups, vs true seroprevalence. Blue: sample allocation specified by the observed participant distribution catchment distribution in the Somerville study. Red: sample allocation specified by the catchment distribution of GPS-estimated visitors to the proposed alternate study site. Dotted line indicates where estimated equals true seroprevalence.

supplemental figure 5 and equation 2 in the online supplemental material).[15]

Bias due to unappreciated heterogeneity in seropositivity across geographic subgroups

Biased seroprevalence estimates can result if procedures for generating weighted prevalence estimates do not appropriately account for geographic heterogeneity in sampling and underlying seropositivity. We compared estimated seroprevalence versus true seroprevalence for numerical simulations in which the final seroprevalence estimates were weighted by (1) the sampling probability for each age-location group or (2) by the sampling probability of each age subgroup alone. The first weighting procedure accounts for heterogeneity across age and location subgroups, whereas the second procedure accounts only for heterogeneity across age subgroups. Using the second procedure resulted in overestimation or underestimation of seroprevalence, depending on whether sample allocation enriches for participants from areas with high or low underlying seropositivity, respectively (figure 5).

DISCUSSION

Convenience sampling, despite its inherent limitations, has continued utility in the public health response to the COVID-19 pandemic. Cost and logistical considerations limit the feasibility of randomised structured sampling, particularly in contexts where census data, population rosters or household mapping data are unavailable or unreliable. Certain forms of convenience sampling may be better suited for reaching important subgroups compared with structured approaches: Lower-wage or frontline workers who are at higher risk of SARS-CoV-2 exposure,[19–21] including undocumented workers,[21] may be less likely to participate if recruited using conventional survey outreach methods (eg, mail or phone contact) due to constraints on their time[22–24] and lack of incentives.[22] Convenience sampling at highly visited community locations such as essential businesses may be an attractive alternative to structured sampling in this important population, similar to sampling approaches developed to study so-called hidden populations.[25] Finally, two recent studies have shown that, in some contexts, SARS-CoV-2 seroprevalence studies obtained via convenience sampling closely approximate those obtained via randomised, representational sampling,[26 27] although additional comparisons in other contexts are needed.

Geographic heterogeneity in SARS-CoV-2 epidemic intensity has been a repeatedly observed feature of the pandemic.[1–3 28] This phenomenon poses unique challenges for seroprevalence studies that employ convenience sampling, in which sample allocation across subgroups cannot be prespecified and is geographically non-uniform. This limitation has important implications for bias and uncertainty of resulting seroprevalence estimates and raises questions about potential undersampling or exclusion of important subgroups in venue-based studies. By using a relatively simple mathematical model, we were able to quantify how mismatches between the geographic distribution of cases and the sampling allocation in seroprevalence studies can influence the precision and interpretability of the resulting seroprevalence estimates.

Multiple studies have identified geographic location as a strong surrogate for multiple risk factors associated with severe infection, hospitalisation, and/or death due to COVID-19[28 29] and undersampling in neighbourhoods where these risk factors colocalise can compromise the reliability and interpretability of seroprevalence estimates. Recruiting participants directly from such communities, where rates of COVID-19 related hospitalisation and deaths are often higher, has yielded seroprevalence estimates that are substantially higher than city-level or state-level estimates.[4 9] In the local seroprevalence study examined here, we observed that venue-based sampling resulted in substantial undersampling of areas where proxy measures (cumulative incidence of PCR-confirmed infections and PCR test positivity rates) suggest higher epidemic intensity, indicating that the seroprevalence estimate from our study is likely lower than the true seroprevalence in Somerville. Notably, the areas that were most undersampled in our study strongly overlap neighbourhoods with lower socioeconomic status, larger proportions of non-white residents, lower proportions of English-speaking households (figure 1A, online supplemental figure 4C).

Our work has three practical findings that are applicable to the design, implementation and interpretation of convenience-based seroprevalence studies.

1. Uncertainty in population-level seroprevalence estimates is minimised when sample allocation is proportional to the size and underlying seropositivity of individual subgroups in the population (equation 2 and online supplemental figure 5). (Uncertainty, in addition to accuracy, is an important consideration, given that low precision can obscure differences in estimated seroprevalence between populations, limiting efforts to understand heterogeneity in epidemic intensity or vaccination coverage.)

Practical application of this finding may be limited because of challenges in reliably ascertaining differences in the underlying seroprevalence between subgroups a priori (eg, if access to diagnostic testing for acute infections is limited or disparate across subgroups). However, even in this situation, allocating sampling effort proportional to subgroup sizes alone can substantially reduce uncertainty (online supplemental figure 5). Although sampling allocation in venue-based sampling is inherently stochastic (ie, it is not possible to predict exactly which participants from what locations will come to a given study site) selection of venue locations with the objective of enriching for participants from geographic subgroups with larger populations and/or higher expected seroprevalence can improve sample allocation and help reduce uncertainty. We note that optimal sampling yielded only marginally better precision over the alternative study site, indicating that more carefully chosen recruitment locations can help improve resulting seroprevalence estimates, even if these sites do not yield perfectly optimal sampling allocations.

2. Our work suggests that GPS-estimated foot traffic data may be useful for evaluating and selecting recruitment sites for serosurveillance studies. Validation against other data sources that directly measure the geographic distributions of visitors to locations of interest (eg, aggregated geographic and registration data from COVID-19 mobile testing programmes) can help further evaluate this potentially important data source. This data has several important limitations, including bias associated with differences in mobile device usage among demographic groups and uncertainties in capture and measurement of highly granular mobility patterns.
3. Convenience sampling can produce biased seroprevalence estimates if geographic heterogeneity in underlying subgroup-level seropositivity is not properly accounted for (figure 5). To avoid this problem, studies that employ convenience sampling should collect geographic data on participants' home locations that is granular enough to capture potential geographic heterogeneity in seropositivity within the study area. This information can be used to quantify what would otherwise be an unmeasured source of bias in resulting seroprevalence estimates.

Limitations

Multiple considerations are important for contextualising our work. Importantly, we assumed in our numerical model that the true seropositivity in each age-location subgroup is proportional to its observed cumulative incidence of PCR-confirmed SARS-CoV-2 infections (per local public health data from Somerville, MA). However, wards with higher PCR positivity rates (an indicator of greater epidemic intensity) have relatively similar rates of overall PCR testing per capita (online supplemental figure 4B), indicating that there were gaps in testing effort in areas of Somerville that had more incident infections overall.[17] The assumed true underlying seroprevalence of each age-location group, which is specified using the observed cumulative incidence of PCR-confirmed infection and does 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 (ie, greater PCR testing effort in heavily impacted areas would likely reveal even larger differences in cumulative incidence between wards). This misspecification, and resultant smaller dispersion in assumed true cumulative incidence by ward, is expected to result in more conservative values for the uncertainty in estimated population-level seroprevalence; otherwise, this limitation is not expected to change our primary findings from the numerical model.

Our work addresses a specific issue with convenience and venue-based sampling strategies, but we note that these approaches still involve multiple important limitations. Among other concerns, individuals with disabilities or others who are less likely to leave their homes may be differentially excluded from venue-based sampling. Methods designed to account for participation bias, including those developed for use with time-location sampling,[30] may be applicable here. Collecting information on non-respondents in venue-based sampling—for example, brief demographic surveys collected before recruitment for serological testing—can help measure and account for potential sources of participation bias.

In summary, we have examined how geographic heterogeneity in sample allocation, combined with underlying heterogeneity in geographic distribution of seropositive individuals, can influence seroprevalence estimates derived from convenience and venue-based sampling. Our findings are relevant to studies employing venue-based recruitment and are also applicable to other kinds of convenience sampling, for example, studies using discarded blood specimens from patients within a hospital's geographic catchment area. The methods introduced here can be applied to venue selection for seroprevalence studies in future outbreaks, particularly when GPS-estimated foot traffic is available for analysis.

Author affiliations
[1]Infectious Diseases Division, Massachusetts General Hospital, Boston, Massachusetts, USA
[2]Center for Communicable Disease Dynamics, Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA
[3]Harvard Medical School, Boston, Massachusetts, USA
[4]François-Xavier Bagnoud Center for Health and Human Rights, Harvard University, Boston, Massachusetts, USA
[5]Massachusetts General Hospital, Boston, MA, USA
[6]Department of Medicine, Cambridge Health Alliance, Cambridge, Massachusetts, USA
[7]Global Medicine Program, Massachusetts General Hospital, Boston, Massachusetts, USA
[8]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA
[9]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
[10]City of Somerville, Somerville, Massachusetts, USA
[11]BioFrontiers Institute, University of Colorado Boulder, Boulder, Colorado, USA

Twitter Abhishek Bhatia @abhibhatia08 and Yonatan Grad @yhgrad

Contributors TB designed, planned and implemented the study; conducted data analysis and wrote and revised the manuscript; and is responsible for the overall content of the manuscript as guarantor. YG, CB, AJI and SB designed, planned and implemented the study; supervised data analysis, and wrote and revised the manuscript; PMSM designed, planned and implemented the study and wrote and revised the manuscript. AB, BB, EKW, DK, KS, KJ, JL and DB designed and planned the study; JSM, AM, JT, WY, JV, VN, WGB and TEM implemented the study. DL conducted data analysis and wrote and revised the manuscript.

Funding This work was supported by the Andrew and Corey Morris-Singer Foundation, National Cancer Institute at the National Institutes of Health (U01CA261277) and the National Institute of Allergy and Infectious Diseases at the National Institutes of Health (T32AI007061 to TB and T32AI007433 to AM). This project has been funded in part by contract 200-2016-91779 with the Centers for Disease Control and Prevention.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not required.

Ethics approval This study involves human participants. The seroprevalence 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). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data collected in this study are available upon reasonable request from the authors.

ORCID iDs
Tyler Brown http://orcid.org/0000-0003-2559-2789
Amir Mohareb http://orcid.org/0000-0002-3761-6154
Yonatan Grad http://orcid.org/0000-0001-5646-1314

REFERENCES

1 Feehan AK, Fort D, Garcia-Diaz J, *et al*. Seroprevalence of SARS-cov-2 and infection fatality ratio, Orleans and Jefferson parishes, Louisiana, USA, may 2020. *Emerg Infect Dis* 2020;26:2765–8.
2 Kim SJ, Bostwick W. Social vulnerability and racial inequality in COVID-19 deaths in chicago. *Health Educ Behav* 2020;47:509–13.
3 Kissler SM, Kishore N, Prabhu M, *et al*. Reductions in commuting mobility correlate with geographic differences in SARS-cov-2 prevalence in New York City. *Nat Commun* 2020;11:4674.
4 Rosenberg ES, Tesoriero JM, Rosenthal EM, *et al*. Cumulative incidence and diagnosis of SARS-cov-2 infection in new york. *Ann Epidemiol* 2020;48:23–9.
5 Havers FP, Reed C, Lim T, *et al*. Seroprevalence of antibodies to SARS-cov-2 in 10 sites in the United States, March 23-may 12, 2020. *JAMA Intern Med* 2020;180:1576.
6 Bendavid E, Mulaney B, Sood N, *et al*. COVID-19 antibody seroprevalence in SANTA Clara County, California. *Int J Epidemiol* 2021;50:410–9.
7 Shook-Sa BE, Boyce RM, Aiello AE. Estimation without representation: early severe acute respiratory syndrome coronavirus 2 seroprevalence studies and the path forward. *J Infect Dis* 2020;222:1086–9.
8 Kelly H, Riddell MA, Gidding HF, *et al*. A random cluster survey and a convenience sample give comparable estimates of immunity to vaccine preventable diseases in children of school age in Victoria, Australia. *Vaccine* 2002;20:3130–6.
9 Naranbhai V, Chang CC, Beltran WFG, *et al*. High seroprevalence of anti-SARS-cov-2 antibodies in chelsea, Massachusetts. *J Infect Dis* 2020;222:1955–9.
10 Webb Hooper M, Nápoles AM, Pérez-Stable EJ. No populations left behind: vaccine hesitancy and equitable diffusion of effective COVID-19 vaccines. *J Gen Intern Med* 2021;36:2130–3.
11 Hobbs CV, Drobeniuc J, Kittle T, *et al*. Estimated SARS-cov-2 seroprevalence among persons aged < 18 years-Mississippi, may-september 2020. *MMWR Morb Mortal Wkly Rep* 2021;70:312–5.
12 Sutton M, Cieslak P, Linder M. Notes from the field: seroprevalence estimates of SARS-cov-2 infection in convenience sample-Oregon, may 11-june 15, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:1100–1.
13 Massachusetts State Department of Health. COVID-19 response reporting. n.d. Available: https://www.mass.gov/info-details/covid-19-response-reporting
14 DataUSA:somerville, MA. n.d. Available: https://datausa.io/profile/geo/somerville-ma
15 Larremore DB, Fosdick BK, Bubar KM, *et al*. Estimating SARS-cov-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. *Elife* 2021;10:e64206.
16 City of Somerville, Massachusetts and Cambridge Health Alliance. The wellbeing of somerville report. n.d. Available: https://www.somervillema.gov/sites/default/files/wellbeing-of-somerville-report-2017.pdf
17 Dryden-Peterson S, Velásquez GE, Stopka TJ, *et al*. Disparities in SARS-cov-2 testing in Massachusetts during the COVID-19 pandemic. *JAMA Netw Open* 2021;4:e2037067.
18 Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 1934;97:558.
19 Hawkins D. Social determinants of COVID-19 in Massachusetts, United States: an ecological study. *J Prev Med Public Health* 2020;53:220–7.
20 Baker MG, Peckham TK, Seixas NS. Estimating the burden of United States workers exposed to infection or disease: a key factor in containing risk of COVID-19 infection. *PLOS ONE* 2020;15:e0232452.
21 Feehan AK, Velasco C, Fort D, *et al*. Racial and workplace disparities in seroprevalence of SARS-cov-2, baton rouge, Louisiana, USA. *Emerg Infect Dis* 2021;27:314–7.
22 Hernández MG, Nguyen J, Casanova S, *et al*. Doing no harm and getting it right: guidelines for ethical research with immigrant communities. *New Dir Child Adolesc Dev* 2013;2013:43–60.
23 Corbie-Smith GM. Minority recruitment and participation in health research. *N C Med J* 2004;65:385–7.
24 Keyzer JF, Melnikow J, Kuppermann M, *et al*. Recruitment strategies for minority participation: challenges and cost lessons from the power interview. *Ethn Dis* 2005;15:395–406.
25 Muhib FB, Lin LS, Stueve A, *et al*. A venue-based method for sampling hard-to-reach populations. *Public Health Rep* 2001;116 Suppl 1:216–22.
26 Kugeler KJ, Podewils LJ, Alden NB, *et al*. Assessment of SARS-cov-2 seroprevalence by community survey and residual specimens, Denver, Colorado, july-august 2020. *Public Health Rep* 2022;137:128–36.
27 Bajema KL, Dahlgren FS, Lim TW, *et al*. Comparison of estimated severe acute respiratory syndrome coronavirus 2 seroprevalence through commercial laboratory residual sera testing and a community survey. *Clin Infect Dis* 2021;73:e3120–3.
28 Maroko AR, Nash D, Pavilonis BT. COVID-19 and inequity: a comparative spatial analysis of new York City and Chicago hot spots. *J Urban Health* 2020;97:461–70.
29 López-Gay A, Spijker J, Cole HVS, *et al*. Sociodemographic determinants of intraurban variations in COVID-19 incidence: the case of barcelona. *J Epidemiol Community Health* 2022;76:1–7.
30 Leon L, Jauffret-Roustide M, Le Strat Y. Design-based inference in time-location sampling. *Biostatistics* 2015;16:565–79.