Article

Burden and characteristics of COVID-19 in the United States during 2020

https://doi.org/10.1038/s41586-021-03914-4

Received: 15 February 2021

Accepted: 13 August 2021

Published online: 26 August 2021

Check for updates

Pei Sen[1]✉, Teresa K. Yamana[1], Sasikiran Kandula[1], Marta Galanti[1] & Jeffrey Shaman[1]✉

The COVID-19 pandemic disrupted health systems and economies throughout the world during 2020 and was particularly devastating for the United States, which experienced the highest numbers of reported cases and deaths during 2020[1–3]. Many of the epidemiological features responsible for observed rates of morbidity and mortality have been reported[4–8]; however, the overall burden and characteristics of COVID-19 in the United States have not been comprehensively quantified. Here we use a data-driven model-inference approach to simulate the pandemic at county-scale in the United States during 2020 and estimate critical, time-varying epidemiological properties underpinning the dynamics of the virus. The pandemic in the United States during 2020 was characterized by national ascertainment rates that increased from 11.3% (95% credible interval (CI): 8.3–15.9%) in March to 24.5% (18.6–32.3%) during December. Population susceptibility at the end of the year was 69.0% (63.6–75.4%), indicating that about one third of the US population had been infected. Community infectious rates, the percentage of people harbouring a contagious infection, increased above 0.8% (0.6–1.0%) before the end of the year, and were as high as 2.4% in some major metropolitan areas. By contrast, the infection fatality rate fell to 0.3% by year's end.

During 2020, the United States documented more COVID-19 cases and deaths than any other country in the world[1]. The first US COVID-19 case was identified in Washington state on 20 January 2020[2]. Over the course of the year, three pandemic waves took place: (1) a spring outbreak in select, mostly urban areas following the introduction of the virus to the United States; (2) a summer wave that predominantly affected the southern half of the country; and (3) an autumn–winter wave that remained pervasive until the spring of 2021. To understand the transmission of the virus and better control its progression in the future, it is vital that the epidemiological features that have supported these outbreaks are quantified and analysed in both space and time.

Here we use a county-resolved metapopulation model to simulate the transmission of SARS-CoV-2 within and between the 3,142 counties of the United States. The model depicts both documented and undocumented infections and is coupled with an iterative Bayesian inference algorithm—the ensemble adjustment Kalman filter—which assimilates observations of daily cases in each county, as well as population movement between counties[9,10] (Supplementary Information). The Bayesian inference supports a fitting of the model to case observations and estimation of unobserved state variables (for example, population susceptibility within a county) and system parameters (for example, the ascertainment rate in each county). Synthetic tests indicate that the inference approach can recover key time-varying parameters across a diversity of simulation scenarios (Extended Data Fig. 1). The model fitting to observed case data captures the three waves of the outbreak as manifest at national scales (Fig. 1a), as well as in major metropolitan areas and at county scales (Extended Data Fig. 2). These inference results are robust to parameter settings and model configurations (Extended Data Figs. 3, 4, Supplementary Information).

To further validate the fitting, we compared model estimates of cumulative infections to findings from US Centers for Disease Control and Prevention (CDC) seroprevalence surveys conducted at site and state levels[3]. The seroprevalence data, which provide an out-of-sample corroboration of the model fitting, were adjusted for the waning of antibody levels following adaptive immune response[11,12] (Extended Data Fig. 5, Supplementary Information). Model estimates of cumulative infected percentages are well aligned with adjusted seroprevalence estimates from the CDC 10-site survey across sites and through time (Pearson's $r = 0.97$, mean absolute error (MAE) = 1.31%) (Fig. 1b) and are similarly well matched to adjusted estimates at the state level (Extended Data Fig. 6). In addition, the seroprevalence generated using the estimated daily infections adjusted for seroreversion also matches the observed seroprevalence, and the results are robust to assumed use of a lower-sensitivity seroassay (Extended Data Fig. 6).

A critical feature of SARS-CoV-2 is its ability to infect and transmit largely from individuals who have not been diagnosed with the virus[4]. The model structure and fitting enable estimation of the ascertainment rate, the percentage of infections confirmed diagnostically, at county scales. The national population-weighted ascertainment rate averaged for all of 2020 was 21.8% (95% CI: 15.9–30.3%), similar to an estimate derived from surveys on healthcare-seeking behaviours[13]. This national ascertainment rate increased from 11.3% (8.3–15.9%) during March 2020 to 24.5% (18.6–32.3%) during December 2020 (Fig. 1c). The increase through time is a likely by-product of increasing testing capacity, a relaxation of initial restrictions on test usage, and increasing

[1]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA. ✉e-mail: sp3449@cumc.columbia.edu; jls106@cumc.columbia.edu

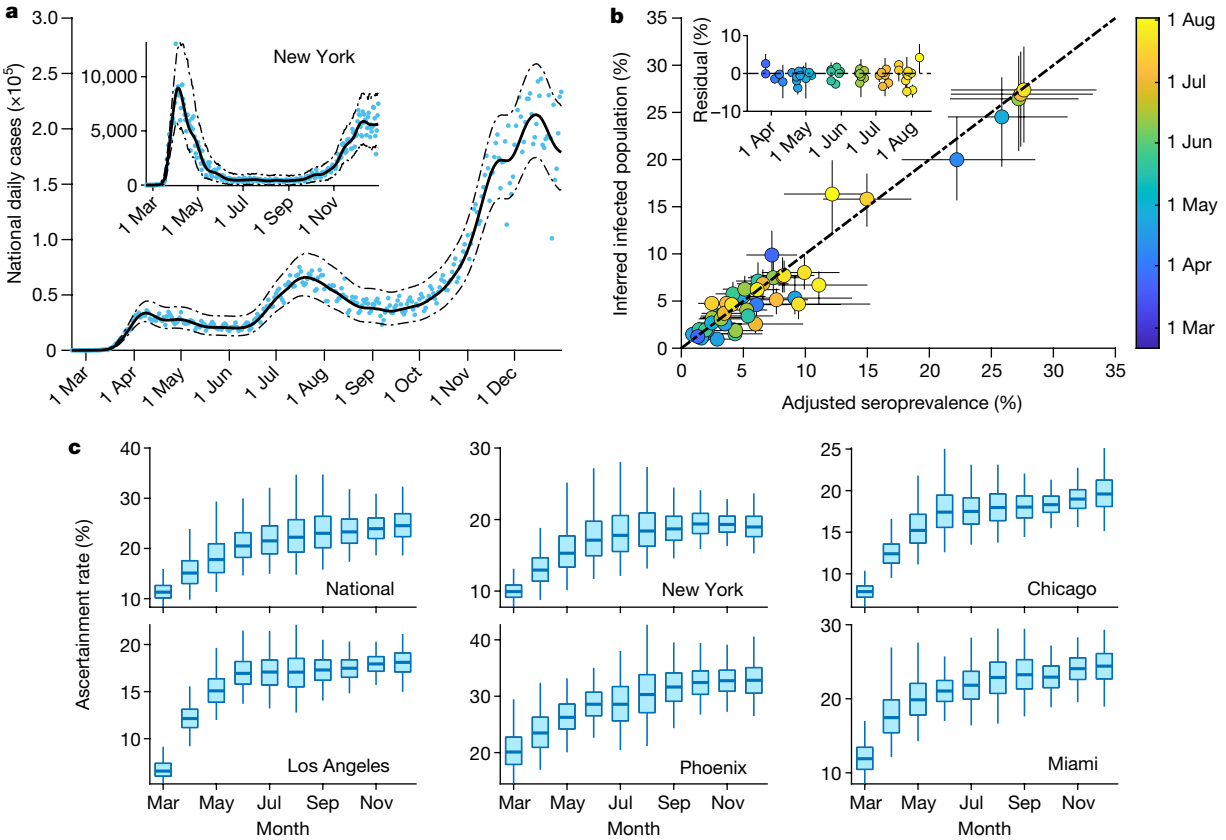


Fig. 1 | Model calibration and ascertainment rate. a, Model fitting to daily case numbers (blue dots) in the United States and the New York metropolitan area (inset). Solid and dashed lines show the median estimate and 95% CIs, respectively. **b**, Comparison between inferred percentage cumulative infections and seroprevalence in ten locations adjusted for antibody waning. The inset shows residuals of inference (inferred percentage of infected population minus adjusted seroprevalence). Centres and whiskers show medians and 95% CIs, and colour indicates the sample collection date in each location. Distributions are obtained from $n = 100$ ensemble members. Details on the serological survey are provided in Supplementary Information. **c**, Distributions of estimated ascertainment rate in the United States and five metropolitan areas. The centre line shows the median, box bounds represent 25th and 75th percentiles, and whiskers show 2.5th and 97.5th percentiles. Monthly posterior estimates are presented for March to December 2020. Distributions are obtained from $n = 100$ ensemble members.

recognition, concern and care-seeking among the public. We additionally focus on five metropolitan areas in the United States. Small differences in the ascertainment rate manifest across these areas—in particular, ascertainment rates for Phoenix and Miami were higher than the national average for much of the year, whereas those for New York City, Chicago and Los Angeles were consistently below the national average.

At the national level, three pandemic waves were evident during spring, summer and autumn–winter (Fig. 1a); however, the structure differs among the five focus metropolitan areas, with New York and Chicago experiencing strong spring and autumn–winter waves but little activity during summer, Los Angeles and Phoenix undergoing summer and autumn–winter waves, and Miami experiencing all three waves (Extended Data Fig. 2). Los Angeles County, the largest county in the United States, with a population of more than 10 million people, was particularly severely affected during autumn–winter. The differences in virus activity produced different cumulative infection numbers through time (Fig. 2a). Population susceptibility at the end of the year was 69.0% (63.6–75.4%) for the United States, and among the focal metropolitan areas it ranged from 47.6% (37.2–54.8%) in Los Angeles to 73.2% (68.3–77.8%) in Phoenix. Although there is variability among counties, a substantial portion of the US population (69.0%) had not been infected by the end of 2020; however, pockets of lower population susceptibility, which are evident in the southwest and southeast on

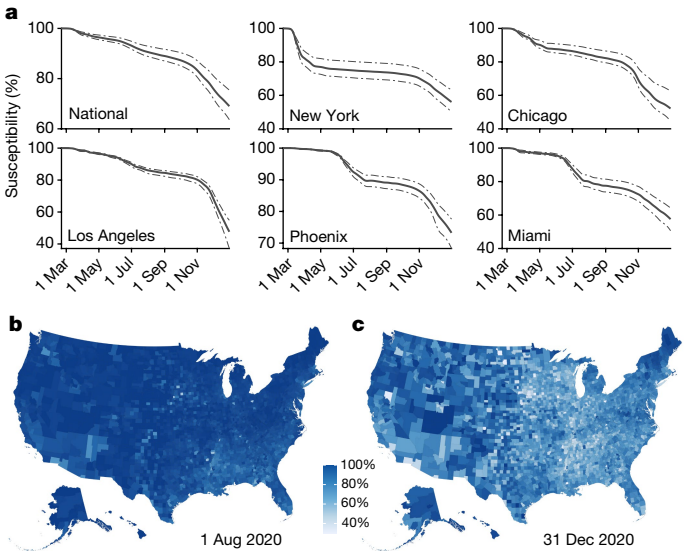


Fig. 2 | Estimates of population susceptibility. a, Estimated evolution of susceptibility to COVID-19 in the United States and five metropolitan areas. Solid lines show median and the area between the dashed lines is the 95% CI. **b**, **c**, Estimated susceptibility in 3,142 US counties on 1 August (**b**) and 31 December (**c**) 2020. Colour shows median estimate.

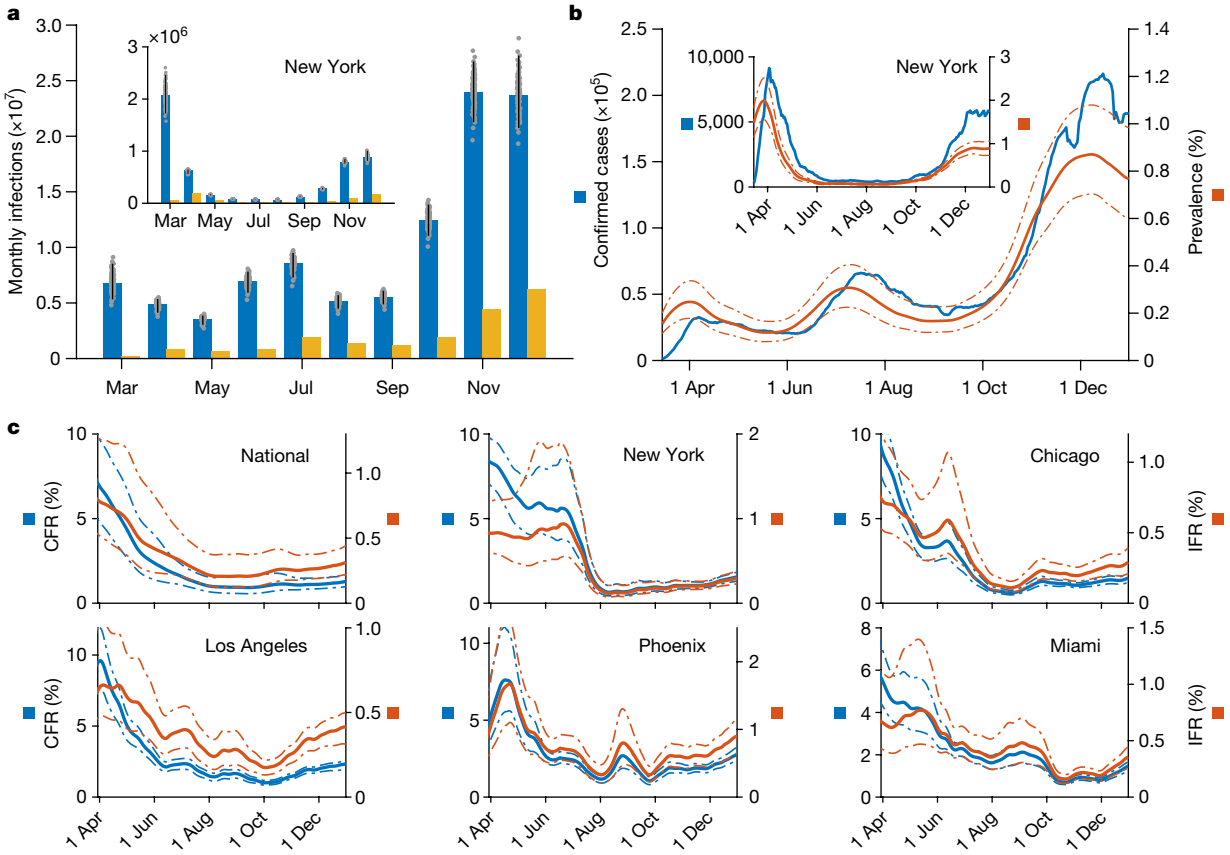


Fig. 3 | Estimated transmission and characteristics of COVID-19 in the United States. a, Estimated monthly total infections (blue bars) and confirmed cases (orange bars) in the United States and the New York metropolitan area (inset). Distributions are obtained from $n = 100$ ensemble members. The blue bars represent the medians and whiskers show 95% CIs. **b**, Daily confirmed cases (blue line, 7-day moving average) and estimated prevalence of contagious infections (red line, median; red dashed lines, 95% CIs) in the United States. Inset, result for the New York metropolitan area. **c**, Estimated CFR (blue lines) and IFR (red lines) in the United States and five metropolitan areas. Solid and dashed lines show median estimate and 95% Cis, respectively.

1 August 2020 (Fig. 2b), expanded considerably by 31 December 2020 (Fig. 2c). In particular, areas of the upper Midwest and Mississippi valley, including the Dakotas, Minnesota, Wisconsin and Iowa, are estimated to have population susceptibility below 40% as of 31 December 2020.

The structure of the outbreak is evident in both incidence and prevalence estimates (Fig. 3, Extended Data Fig. 7). Incidence indicates the daily number of newly infectious individuals—both confirmed cases of COVID-19 and those whose infections remain undocumented. The majority of infections each month are undocumented (Fig. 3a), as indicated by the low ascertainment rates (Fig. 1c). For all of 2020, an estimated 78.2% of infections in the United States were undocumented. Estimates of daily prevalence provide a measure of the community infectious rate (CIR), the fraction of the population currently harbouring a contagious infection. The national SARS-CoV-2 CIR was 0.77% (0.60–0.98%) on 31 December 2020, indicating that roughly 1 in 130 people was contagious (a similar percentage, 0.83% (0.52–1.26%), was estimated to be latently infected—that is, infected but not yet contagious) (Fig. 3b). Among the 5 focal metropolitan areas, the CIR varied considerably: in mid-November, Chicago reached a CIR of 1.51% (1.27–1.82%); whereas in Miami CIR increased to 1.25% (1.03–1.53%) during July. Los Angeles was even more burdened at the end of 2020, with a CIR of 2.42% (2.05–2.86%) as of 31 December 2020 (Extended Data Fig. 7).

The model fitting enables estimation of the case fatality rate (CFR) and the infection fatality rate (IFR). Using public line-list data from the CDC[14], we estimated the distribution of time lag from case confirmation to death for each county and, using these estimates, deconvolved observed deaths to their date of case reporting[15] (Extended Data Figs. 8, 9, Supplementary Information). CFR and IFR were then generated using these deconvolved death data. Both rates were highest nationally at the beginning of the spring wave: the CFR was 7.1% (4.8–9.8%) and the IFR was 0.77% (0.51–1.25%) during April (Fig. 3c). The national cumulative IFR up to 1 June was 0.69% (0.47–1.04%), in line with previous studies[5–7] (Extended Data Fig. 2, Supplementary Information). Over the course of the year, with earlier diagnosis and treatment, improved patient care[16–18] and—in the case of CFR—increased reporting of mild infections, the CFR and IFR dropped to 1.29% (0.98–1.68%) and 0.31% (0.22–0.44%) by December 2020, respectively. Both rates varied by location and over time; for instance, intermediate drops of CFR and IFR began for Chicago, Phoenix and Miami during the summer wave, in association with a decrease of the average age of hospitalized patients (Extended Data Fig. 8). During the winter of 2020, the CFR and IFR in most metropolitan areas increased slightly, possibly driven by greater hospitalization rates among older individuals (Extended Data Fig. 8) and strained healthcare resources[19]. Overall, these findings delineate the mortality risk associated with infection broadly. The national IFR during the latter half of 2020 hovers around 0.30%, well above estimates for both seasonal influenza[20] (<0.08%) and the 2009 influenza pandemic[21] (0.0076%). As COVID-19 deaths are likely to be under-reported, our estimate of IFR could be biased low.

We further examined the change of the reproduction number R_t, in response to changing local, reported COVID-19 case numbers in five US regions (Northeast, Southeast, Midwest, Southwest and West) during the spring, summer and autumn–winter (Supplementary Information). Results indicate that communities with increasing cases showed greater reductions of R_t (Extended Data Fig. 10). However, the rate of reduction

in R_t decreased over successive waves. These findings are potentially driven by a number of factors modulating the reproduction number, including changing compliance with non-pharmaceutical interventions[22] and seasonal modulation of virus transmissibility[23]. A more thorough analysis of this preliminary finding is needed.

The United States experienced the highest numbers of confirmed COVID-19 cases and deaths in the world during 2020[1]. Our findings provide quantification of the time-evolving epidemiological characteristics associated with successive pandemic waves in the United States, as well as conditions at the end of the year and prospects for 2021. Critically, despite more than 19.6 million reported cases by the end of 2020, an estimated 69% of the population remained susceptible to viral infection. Several factors will considerably alter population susceptibility in the coming months. First, ongoing transmission will infect naive hosts and continue to deplete the susceptible pool. Second, as more vaccine is distributed and administered, more individuals will be protected against symptomatic infection and the IFR will decrease. Finally, our model does not represent reinfection, either through waning immunity or immune escape; however, reinfection has been documented[24,25], evidence of waning antibody levels exists[26,27], and new variants of concern have emerged[28,29] and will probably continue to do so. All these processes will affect population susceptibility over time and help to determine when society enters a post-pandemic phase, the pattern of endemicity the virus ultimately assumes and its long-term public health burden[30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03914-4.

1. WHO Coronavirus (COVID-19) Dashboard. *World Health Organization* https://covid19.who.int (2021).
2. Holshue, M. L. et al. First case of 2019 novel coronavirus in the United States. *N. Engl. J. Med.* **382**, 929–936 (2020).
3. COVID Data Tracker. *US Centers for Disease Control and Prevention* https://covid.cdc.gov/covid-data-tracker (2021).
4. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
5. Brazeau, N. et al. *Report 34: COVID-19 Infection Fatality Ratio: Estimates From Seroprevalence*, http://spiral.imperial.ac.uk/handle/10044/1/83545 (2020).
6. O'Driscoll, M. et al. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140–145 (2021).
7. Meyerowitz-Katz, G. & Merone, L. A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. *Int. J. Infect. Dis.* **101**, 138–148 (2020).
8. Kalish, H. et al. Undiagnosed SARS-CoV-2 seropositivity during the first six months of the COVID-19 pandemic in the United States. *Sci. Transl. Med.* **13**, abh3826 (2021).
9. Pei, S., Kandula, S. & Shaman, J. Differential effects of intervention timing on COVID-19 spread in the United States. *Sci. Adv.* **6**, eabd6370 (2020).
10. Yamana, T., Pei, S., Kandula, S. & Shaman, J. Projection of COVID-19 cases and deaths in the US as individual states re-open May 4, 2020. Preprint at https://doi.org/10.1101/2020.05.04.20090670 (2020).
11. Buss, L. F. et al. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science* **371**, 288–292 (2021).
12. Shioda, K. et al. Estimating the cumulative incidence of SARS-CoV-2 infection and the infection fatality ratio in light of waning antibodies. *Epidemiology* **32**, 518–524 (2021).
13. Estimated Disease Burden of COVID-19. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html (2021).
14. COVID-19 Case Surveillance Public Use Data with Geography. *Centers for Disease Control and Prevention* https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4 (2021).
15. Goldstein, E. et al. Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proc. Natl Acad. Sci. USA* **106**, 21825–21829 (2009).
16. RECOVERY Collaborative Group Dexamethasone in hospitalized patients with Covid-19. *N. Engl. J. Med.* **384**, 693–704 (2021).
17. Horwitz, L. I. et al. Trends in COVID-19 risk-adjusted mortality rates. *J. Hosp. Med.* **16**, 90–92 (2021).
18. Beigel, J. H. et al. Remdesivir for the treatment of Covid-19—final report. *N. Engl. J. Med.* **383**, 1813–1826 (2020).
19. Lefrancq, N. et al. Evolution of outcomes for patients hospitalised during the first 9 months of the SARS-CoV-2 pandemic in France: a retrospective national surveillance data analysis. *Lancet Reg. Health Eur.* **5**, 100087 (2021).
20. Burden of influenza. *Centers for Disease Control and Prevention* https://www.cdc.gov/flu/about/burden/index.html (2020).
21. Riley, S. et al. Epidemiological characteristics of 2009 (H1N1) pandemic influenza based on paired sera from a longitudinal community cohort study. *PLOS Med.* **8**, e1000442 (2011).
22. Du, Z. et al. Pandemic fatigue impedes mitigation of COVID-19 in Hong Kong. *Res. Sq.* https://doi.org/10.21203/rs.3.rs-591241/v1 (2021).
23. Ma, Y., Pei, S., Shaman, J., Dubrow, R. & Chen, K. Role of meteorological factors in the transmission of SARS-CoV-2 in the United States. *Nat. Commun.* **12**, 3602 (2021).
24. To, K. K.-W. et al. Coronavirus disease 2019 (COVID-19) re-infection by a phylogenetically distinct severe acute respiratory syndrome coronavirus 2 strain confirmed by whole genome sequencing. *Clin. Infect. Dis.* ciaa1275 (2020).
25. Tillett, R. L. et al. Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect. Dis.* **21**, 52–58 (2021).
26. Self, W. H. Decline in SARS-CoV-2 antibodies after mild infection among frontline health care personnel in a multistate hospital network—12 states, April–August 2020. *Morb. Mortal. Wkly. Rep.* **69**, 1762–1766 (2020).
27. Choe, P. G. et al. Waning antibody responses in asymptomatic and symptomatic SARS-CoV-2 infection. **27**, 327–329 (2020).
28. Fiorentini, S. et al. First detection of SARS-CoV-2 spike protein N501 mutation in Italy in August, 2020. *Lancet Infect. Dis.* **21**, s1473-3099 (2021).
29. Rambaut, A. et al. Preliminary genomic characterization of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological* https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2020).
30. Shaman, J. & Galanti, M. Will SARS-CoV-2 become endemic? *Science* **370**, 527–529 (2020).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The human mobility and COVID-19 surveillance data that support the findings of this study are available at GitHub (https://github.com/SenPei-CU/COVID_US_2020). The county-level COVID-19 surveillance data for the United States are available at Johns Hopkins University coronavirus resource center (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series). County-to-county commuting data were downloaded from the US Census Bureau (https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html). Human mobility data in 2020 were provided by SafeGraph (https://safegraph.com/), which aggregates anonymized location data from numerous applications to provide insights about physical places, via the SafeGraph Community. To enhance privacy, SafeGraph excludes census block group information if fewer than five devices visited an establishment in a month from a given census block group. We aggregated the mobility data to county level to estimate change of inter-county mobility in 2020. Aggregated and derived data are allowed to be shared publicly by SafeGraph. Seroprevalence data were published by the CDC (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html). The line-list datasets are available at the CDC website (https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf and https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4). Source data are provided with this paper.

Code availability

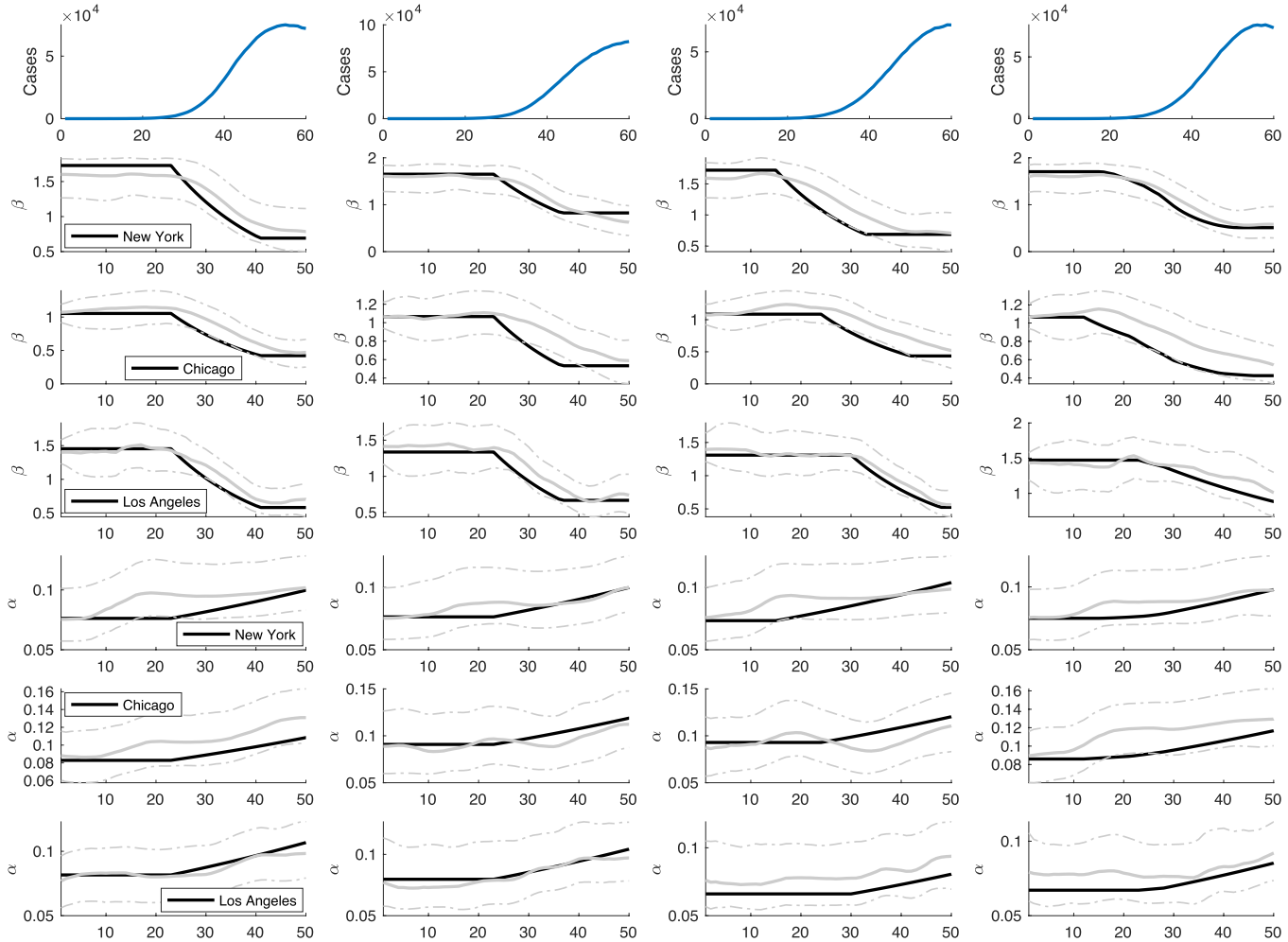
Custom code supporting this study is available at GitHub (https://github.com/SenPei-CU/COVID_US_2020).

Acknowledgements This study was supported by funding from the National Science Foundation (DMS-2027369) and a gift from the Morris-Singer Foundation. We thank SafeGraph for providing human mobility data and Columbia University Mailman School of Public Health for high-performance computing resources

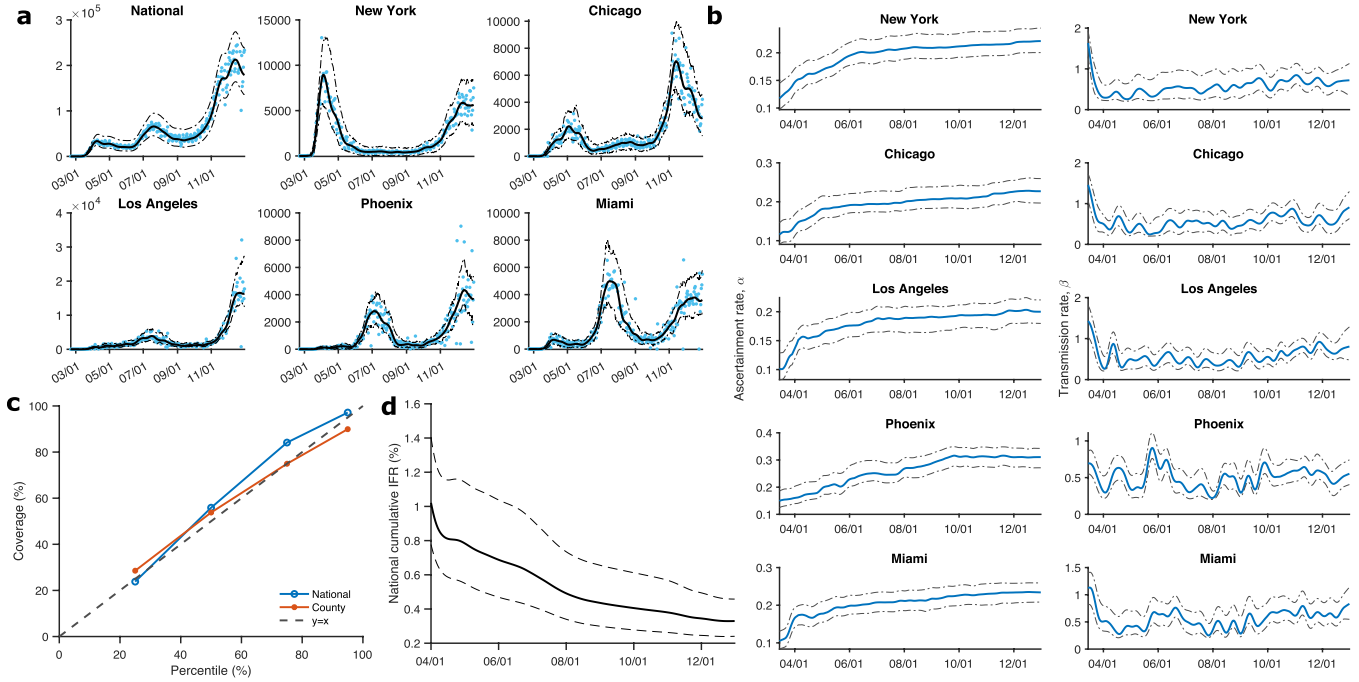
Author contributions S.P. and J.S. conceived the study; S.P., T.K.Y., S.K. and M.G. performed the analysis; and S.P. and J.S. drafted the manuscript. All authors revised and reviewed the manuscript.

Competing interests J.S. and Columbia University disclose partial ownership of SK Analytics. J.S. discloses consulting for BNI. All other authors declare no competing interests.

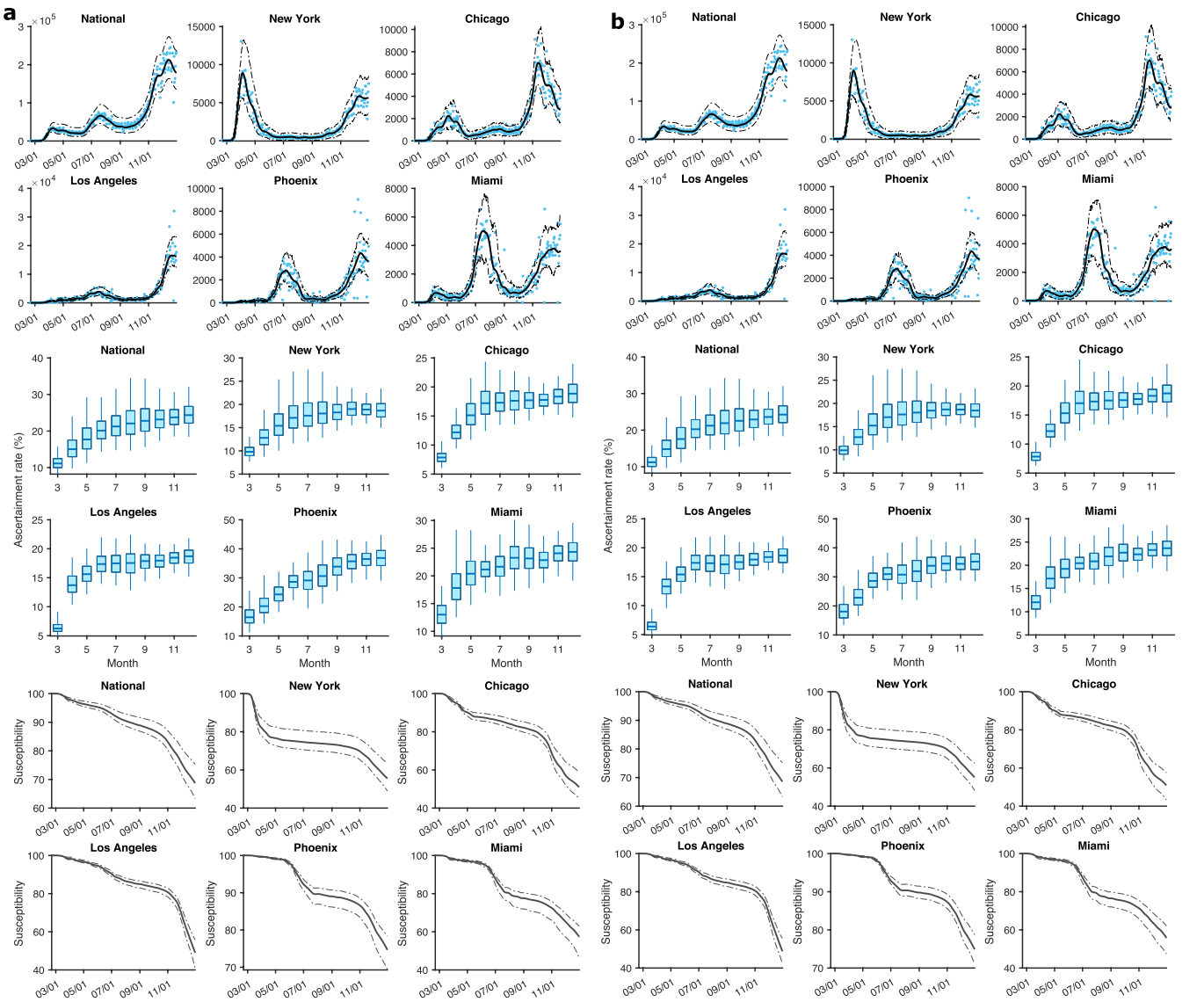
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03914-4.
Correspondence and requests for materials should be addressed to P.S. or J.S.
Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.



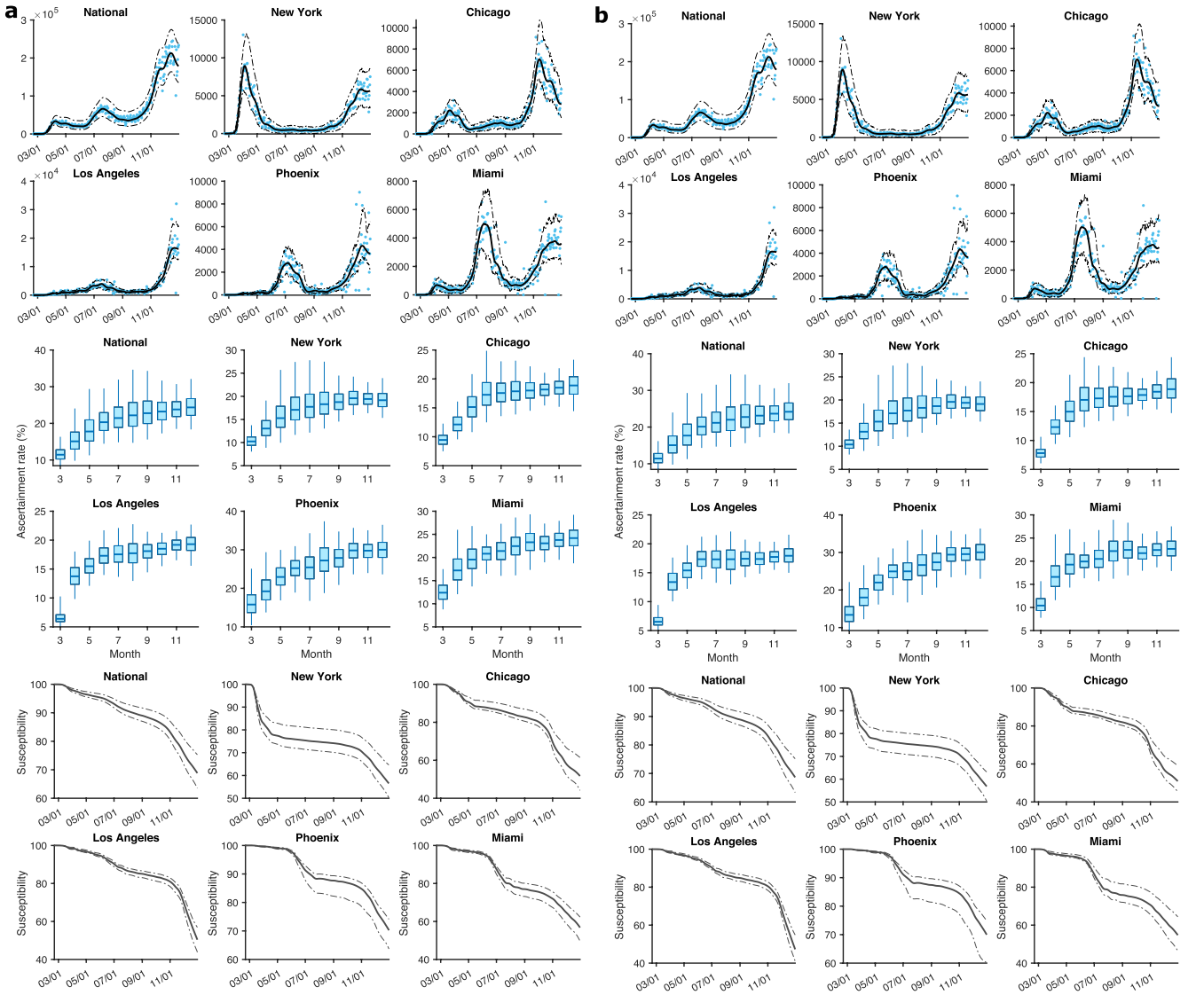
Extended Data Fig. 1 | Parameter inference for simulated outbreaks. Results are shown for three major metropolitan areas – New York, Chicago, and Los Angeles. Outbreaks were generated for 60 days using four prescribed scenarios. National daily cases are shown in the top row. Parameter estimates for the last 10 days are not displayed as there is not enough data at the end of the time series to constrain the model. Solid and dashed lines show the median estimate and 95% CIs respectively.



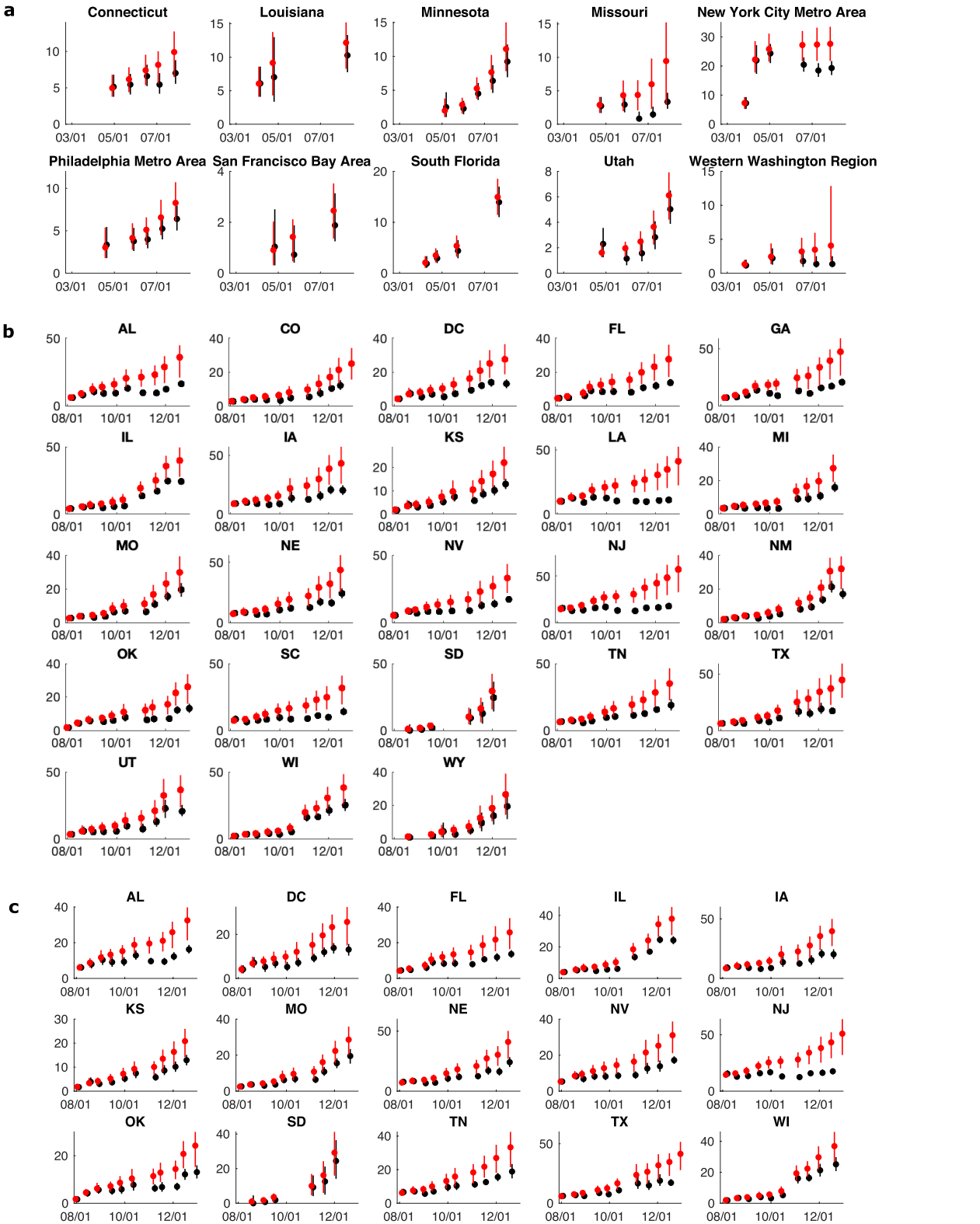
Extended Data Fig. 2 | Model fitting and inference results. (**a**) Model fitting to daily case numbers (blue dots) in the US and five metropolitan areas. Solid and dashed lines show the median estimate and 95% CIs respectively. (**b**) Estimated daily ascertainment rates (left column) and transmission rates (right column) for five metropolitan areas. Solid and dashed lines show the median estimate and 95% CIs respectively. (**c**) Reliability plot for model calibration. Data points show the coverage of the 25%, 50%, 75% and 95% CIs of the posterior fitting at county and national levels. (**d**) The estimated national cumulative IFR in 2020. The cumulative IFR is computed using the estimated cumulative numbers of death (deconvolved) and infections prior to a given date.



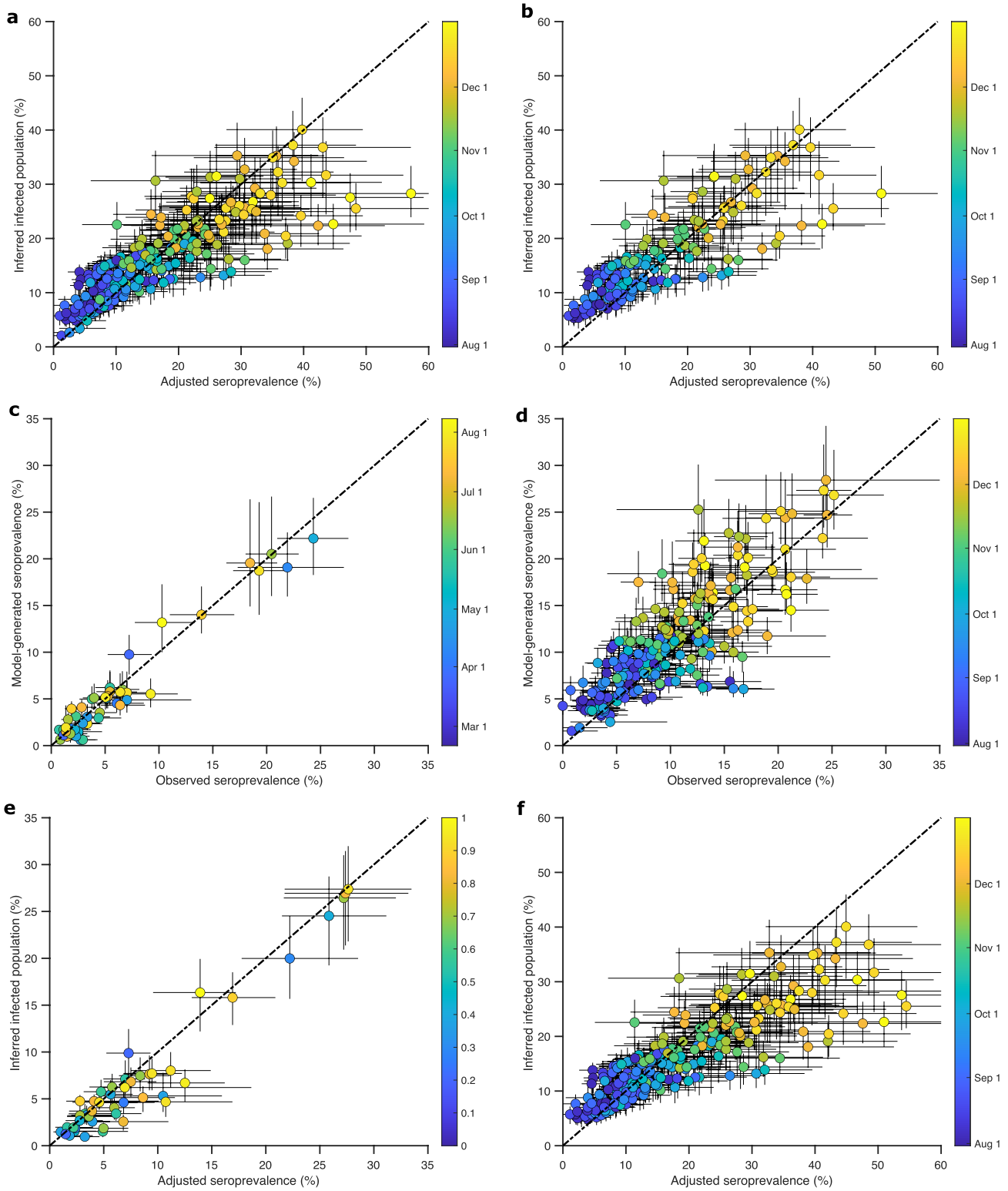
Extended Data Fig. 3 | Sensitivity analyses on inference results. (**a**) Inference results using fixed parameters (Z, D, μ, θ) estimated from case data prior to April 2 2020. (**b**) Inference results from a modified version of the transmission model in which the relative infectiousness of undocumented infections, μ, is allowed to vary over time. Fitting to case data (top two rows), estimated monthly ascertainment rate (middle two rows) and population susceptibility (bottom two rows) are shown. Distributions are obtained from $n = 100$ ensemble members. In the top two and bottom two rows, the solid line represents the median, and the dash lines show 95% CIs. In the middle two rows, centre and box bounds represent the median, 25th, and 75th percentiles, and whiskers show 2.5th and 97.5th percentiles.



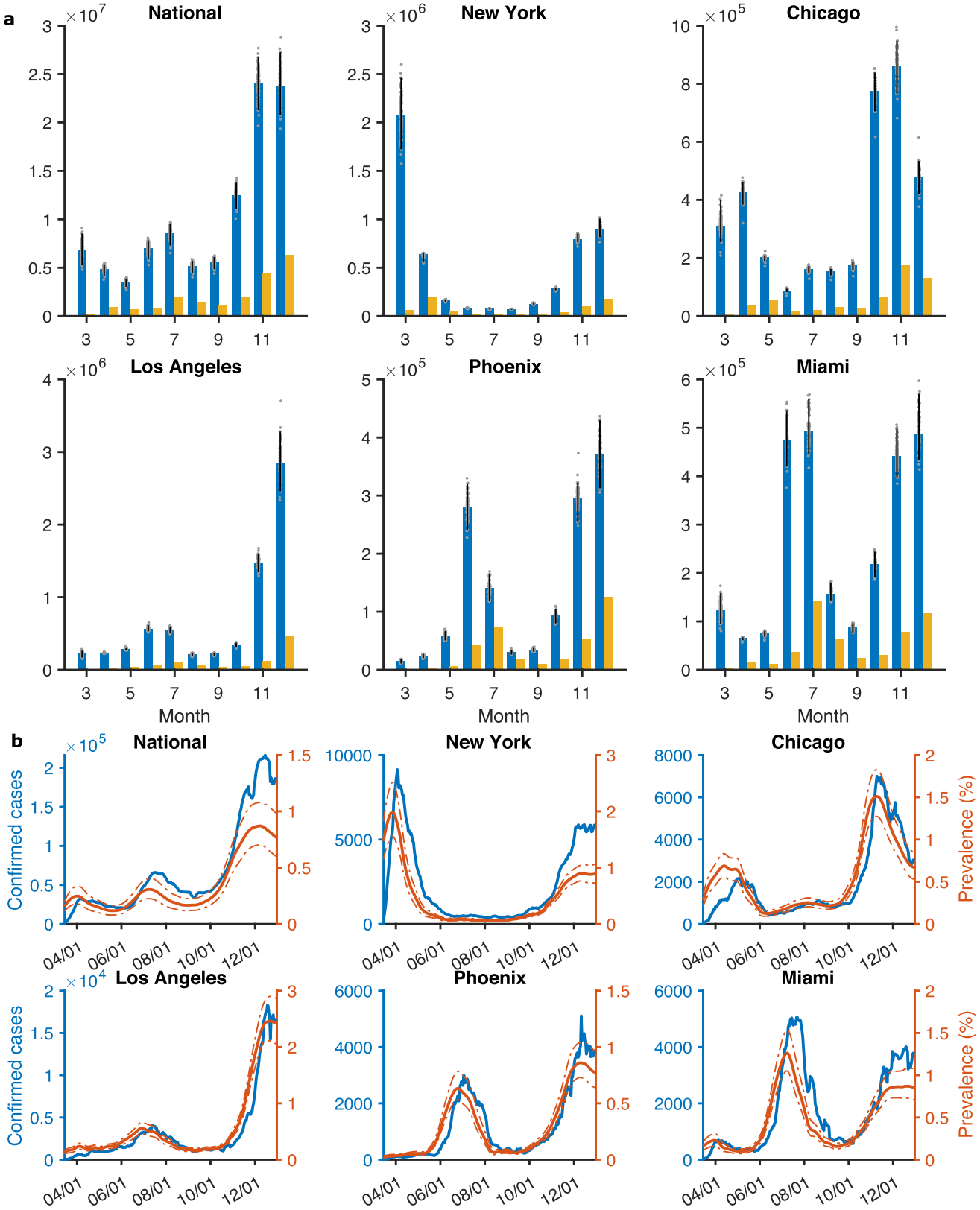
Extended Data Fig. 4 | Inference results from a modified version of the transmission model permitting movement of documented infections among counties. (**a**) 25% of documented infections are allowed to move among counties. (**b**) 50% of documented infections are allowed to move among counties. Fitting to case data (top two rows), estimated monthly ascertainment rate (middle two rows) and population susceptibility (bottom two rows) are shown. Distributions are obtained from $n = 100$ ensemble members. In the top two and bottom two rows, the solid line represents the median, and the dash lines show 95% CIs. In the middle two rows, the centre and box bounds represent the median, 25th, and 75th percentiles, and the whiskers show 2.5th and 97.5th percentiles.



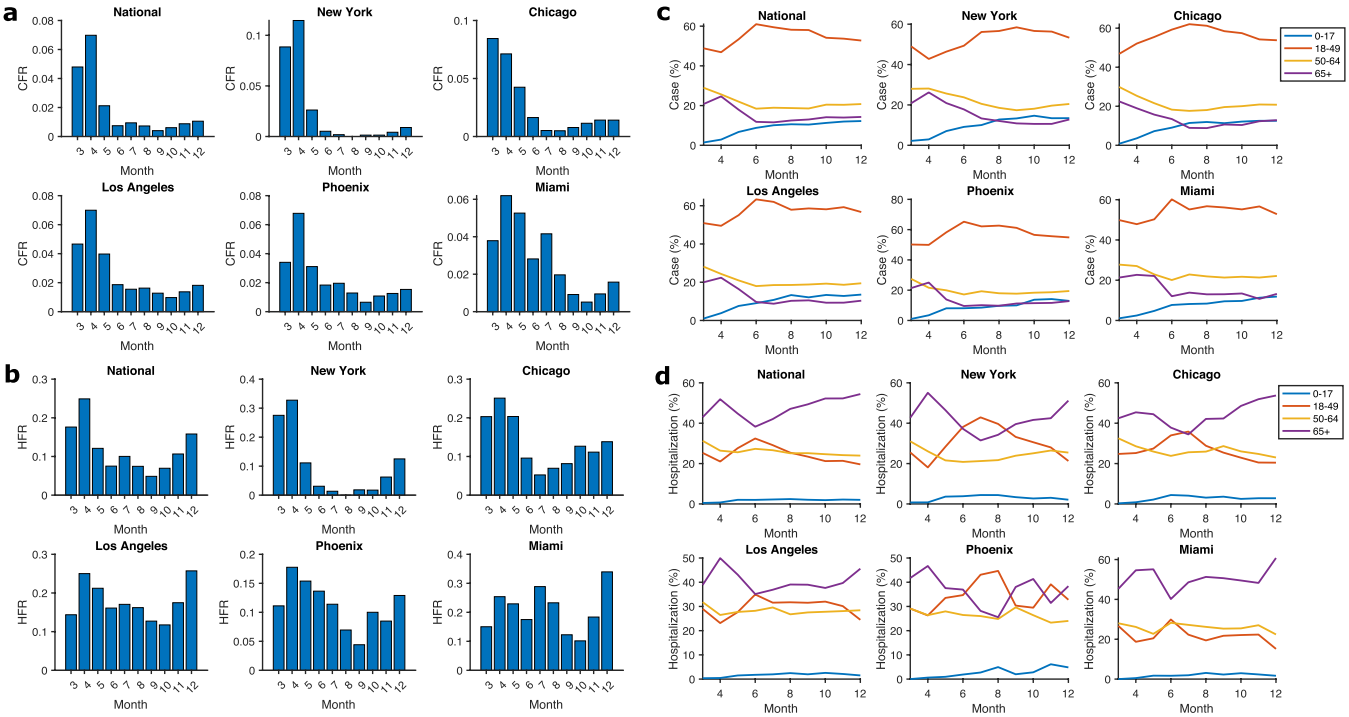
Extended Data Fig. 5 | The reported (black) and adjusted (red) seroprevalence. (**a**) Results for the 10-site study. (**b**) and (**c**) show the results for state-level serological surveys obtained using a maximum monthly attenuation rate of 17.5% and 15%, respectively. Dots and whiskers show the median and 95% CIs respectively. Distributions are obtained from $n = 1{,}000$ simulated seroprevalence samples.



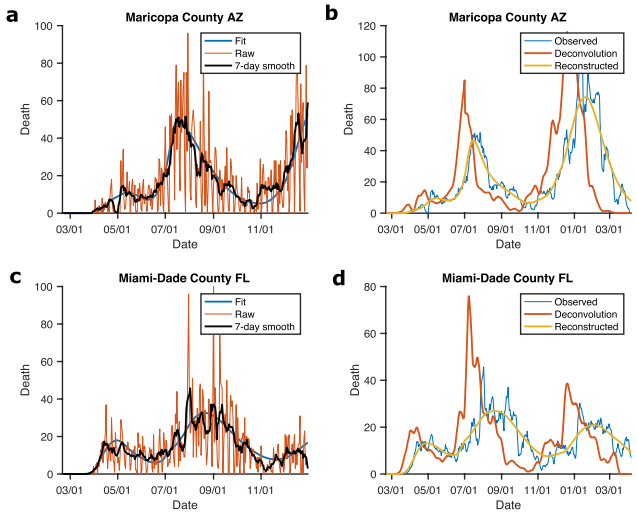
Extended Data Fig. 6 | Validation of inference using seroprevalence data. (a) – (b) Comparison between the inferred percentage of cumulative infections and seroprevalence at the state level adjusted for antibody waning. Seroprevalence data adjusted using a maximum monthly attenuation rate of 17.5% (**a**) and 15% (**b**) are included in the analysis. (c) – (d) Comparison between the model-generated seroprevalence and observed seroprevalence in 10 locations (**c**) and at the state level (**d**). (e) – (f) Comparison between the inferred percentage of cumulative infections and seroprevalence in 10 locations (**e**) and at the state level (**f**) adjusted for antibody waning using lower sensitivity and specificity. Distributions are obtained from $n = 100$ ensemble members. Centre and whiskers show median and 95% CIs. Color indicates the sample collection date for each location.



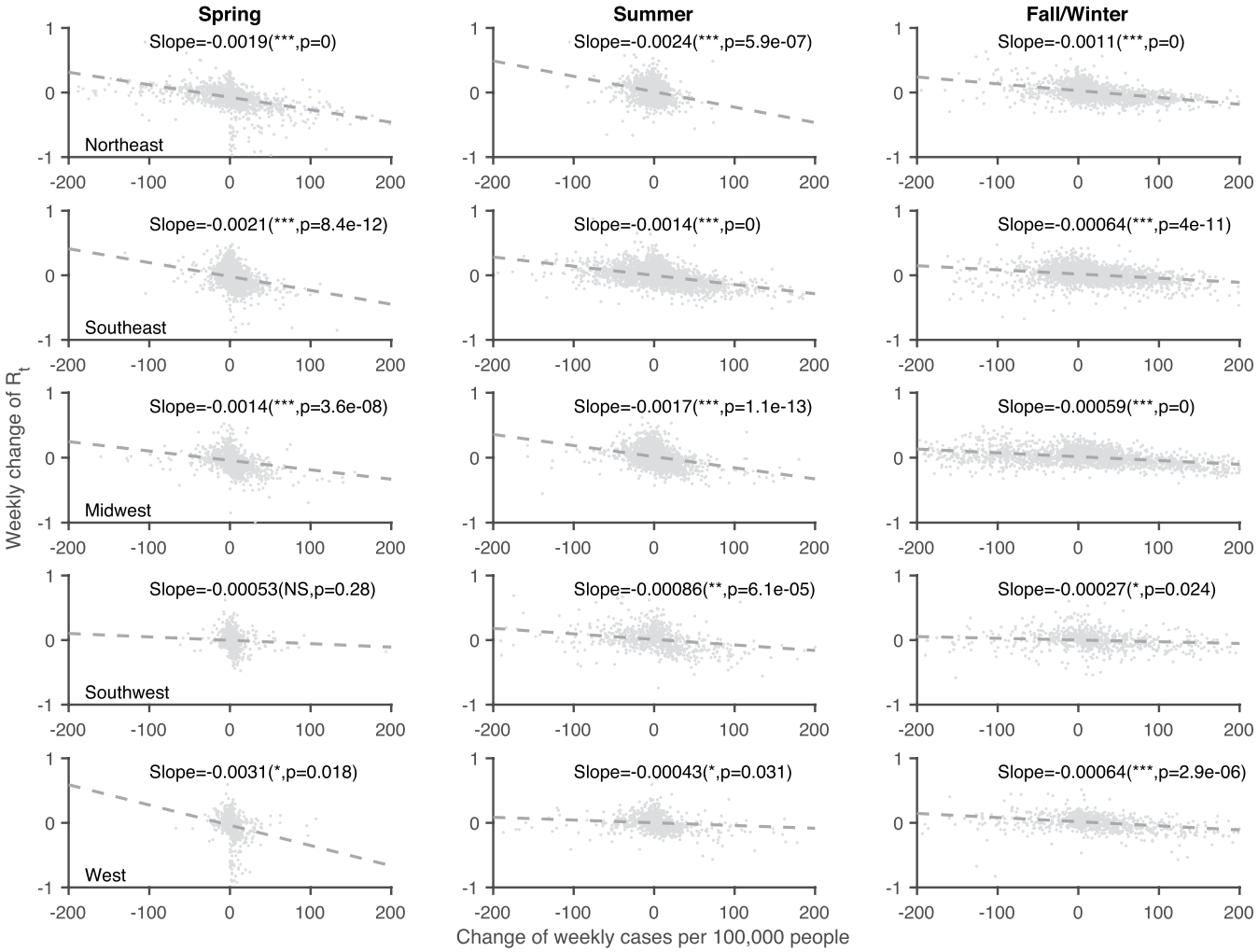
Extended Data Fig. 7 | Inference results in the US and five metropolitan areas. (**a**) Estimated monthly total infections (blue bars) and confirmed cases (orange bars) in the US and five metropolitan areas. Distributions are obtained from $n = 100$ ensemble members. The blue bars show medians and whiskers show 95% CIs. (**b**) Daily confirmed cases (blue line, 7-day moving average) and estimated prevalence of contagious infections (red line, median and 95% CIs) for the US and five metropolitan areas.



Extended Data Fig. 8 | Key statistics obtained from line-list data for the US and five metropolitan areas. (a) – (b) The crude monthly CFR (**a**) and HFR (**b**) obtained from line-list data for the US and five metropolitan areas. Note that due to incomplete reporting of deaths in the line-list data, these estimates are likely low. (c) – (d) The proportion of confirmed cases (**c**) and hospitalizations (**d**) in four age groups (0-17, 18-49, 50-64, 65+) in the line-list data. Data are shown monthly for the US and five metropolitan areas.



Extended Data Fig. 9 | Estimation of the time-to-event distribution from case confirmation to death for Maricopa County AZ (a) and Miami-Dade County FL (c). Deconvolution of daily deaths using the estimated delay distributions for Maricopa County AZ (**b**) and Miami-Dade County FL (**d**).



Extended Data Fig. 10 | Weekly change of R_t in response to the change of weekly cases per 100,000 people at county level. The analysis was performed for five US regions (Northeast, Southeast, Midwest, Southwest, West) during the spring (Feb 21 – May 31), summer (Jun 1 – Sep 15), and fall/winter (Sep 16 – Dec 31) waves. In the five US regions, 116, 162, 126, 45 and 54 counties that reported cumulative cases over 100 per 100,000 people during all three waves and had a population over 100,000 were included in the analysis. A positive/negative change of weekly cases in the x-axis indicates increasing/decreasing community prevalence of COVID-19. The dash lines are the linear fits. The statistical significance of the slope is indicated by asterisks (two-sided t-test. ***: $p < 10^{-5}$, **: $p < 0.001$, *: $p < 0.05$; NS: not significant. P-values are reported in the legends).

nature portfolio

Corresponding author(s): Jeffrey Shaman, Sen Pei

Last updated by author(s): Aug 6, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We implemented the EAKF algorithm in MATLAB R2021a. Data analysis was performed using MATLAB R2021a. The custom codes are publicly posted at GitHub: https://github.com/SenPei-CU/COVID_US_2020.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The human mobility and COVID-19 surveillance data that support the findings of this study are available at GitHub (https://github.com/SenPei-CU/COVID_US_2020). The county-level COVID-19 surveillance data in the US are available at Johns Hopkins University coronavirus resource center (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series). County-to-county commuting data were downloaded from the US Census Bureau (https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html). Human mobility data in 2020 were provided by SafeGraph (https://safegraph.com/), a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places, via the Placekey Community (https://placekey.io/). To enhance privacy, SafeGraph excludes census block group information if fewer than five devices visited an

establishment in a month from a given census block group. We aggregated the mobility data to county level to estimate change of inter-county mobility in 2020. Aggregated and derived data are allowed to be shared publicly by SafeGraph. Seroprevalence data were published by the US CDC (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html). The line-list datasets are available at the US CDC website (https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf and https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not collect data from samples. In the EAKF algorithm, we used n=100 ensemble members to represent the distributions of model states, parameters, and inference outcomes.
Data exclusions	No data were excluded.
Replication	This is a modeling study without experiments. We repeated the inference multiple times and the inference results are reproducible and robust.
Randomization	This is a modeling study without experiments. Randomization is not relevant.
Blinding	This is a modeling study without experiments. Blinding is not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging