Epidemics 52 (2025) 100838

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Incident COVID-19 infections before Omicron in the U.S.

Rachel Lobay [a,*], Ajitesh Srivastava [b], Ryan J. Tibshirani [c], Daniel J. McDonald [a]

[a] Department of Statistics, The University of British Columbia, Earth Sciences Building, 2207 Main Mall, Room 3182, Vancouver, BC, Canada, V6T 1Z4
[b] Department of Computer and Electrical Engineering, University of Southern California, EEB 226, 3740 McClintock Ave, Los Angeles, CA, 90089-2562, United States
[c] Department of Statistics, The University of California, Berkeley, 367 Evans Hall, Berkeley, CA, 94720-3860, United States

ARTICLE INFO

Dataset link: https://github.com/cmu-delphi/latent-infections/

Keywords:
COVID-19
SARS-CoV-2
Infections
Deconvolution
Time series
Seroprevalence
Antibody

ABSTRACT

The timing and magnitude of COVID-19 infections are of interest to the public and to public health, but these are challenging to ascertain due to the volume of undetected asymptomatic cases and reporting delays. Accurate estimates of COVID-19 infections based on finalized data can improve understanding of the pandemic and provide more meaningful quantification of disease patterns and burden. Therefore, we retrospectively estimate daily incident infections for each U.S. state prior to Omicron. To this end, reported COVID-19 cases are deconvolved to their likely date of infection onset using delay distributions estimated from the CDC line list. Then, a novel serology-driven model is used to scale these deconvolved cases to account for the unreported infections. The resulting infection estimates incorporate variant-specific incubation periods, reinfections, and waning antigenic immunity. They clearly demonstrate that reported cases failed to reflect the full extent of disease burden in all states. Most notably, infections were severely underreported during the Delta wave, with an estimated reporting rate as low as 6.3% in New Jersey, 7.3% in Maryland, and 8.4% in Nevada. Moreover, in 44 states, fewer than 1/3 of infections eventually appeared as case reports, and there were sustained periods where surges in infections were virtually undetectable through reported cases. This pattern was clearly illustrated by North and South Dakota during the spring of 2021, as well as by several Northeastern states during the Delta wave of late summer that year. While reported cases offered a convenient proxy of disease burden, they failed to capture the full extent of infections and severely underestimated the true disease burden. Our retrospective analysis also estimates other important quantities for every state, including variant-specific deconvolved cases, time-varying case ascertainment ratios, as well as infection-hospitalization and infection-fatality ratios.

1. Introduction

Reported COVID-19 cases are a staple in tracking the pandemic at varying geographic resolutions (Dong et al., 2020; The New York Times, 2020; The Washington Post, 2020). Yet, for every case that was eventually reported to public health, several infections were likely to have occurred. To see why, it is important to understand *whose* cases were being reported and what differentiates them from unreported cases as well as *when* these case reports happened. Fig. 1 shows an idealized path of a symptomatic infection that is eventually reported to public health. This figure illustrates a number of sources of bias in the reporting pipeline. For instance, diagnostic testing mainly targeted symptomatic individuals; thus, infected individuals exhibiting little to no symptoms were omitted (Centers for Disease Control and Prevention, 2022). In addition, testing practices, availability, and uptake varied temporally and spatially (Pitzer et al., 2021; European Centre for Disease Prevention and Control, 2020; Hitchings et al., 2021). Finally, cases provided a belated view of the pandemic's progression, because they were subject to delays due to the viral incubation period, the speed and severity of symptom onset, laboratory confirmation, test turnaround times, and eventual submission to public health (Pellis et al., 2021; Washington State Department of Health, 2020). For these reasons, reported cases were lagging indicators of the course of the pandemic. Furthermore, they did not represent the actual number of new infections that occurred on any given day based on exposure to the pathogen. Since there was no large-scale surveillance effort in the United States that reliably tracked symptom onset, let alone infection onset, ascertaining the onset of all *infections* is challenging.

* Corresponding author.
E-mail addresses: rachel.lobay@stat.ubc.ca (R. Lobay), ajiteshs@usc.edu (A. Srivastava), ryantibs@berkeley.edu (R.J. Tibshirani), daniel@stat.ubc.ca (D.J. McDonald).

https://doi.org/10.1016/j.epidem.2025.100838
Received 18 October 2024; Received in revised form 28 February 2025; Accepted 26 May 2025
Available online 9 June 2025

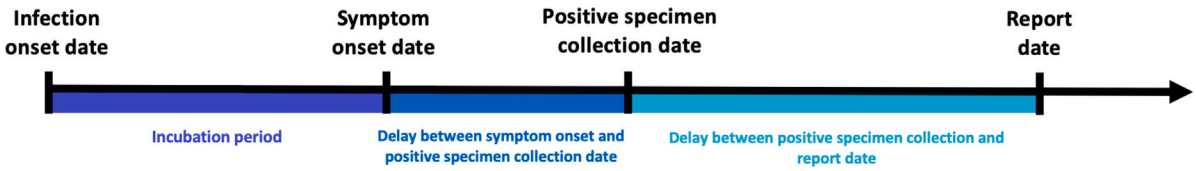

Fig. 1. Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

Infection prevalence studies were conducted in various countries over the course of the pandemic to measure and monitor the prevalence of COVID-19 in real-time. The UK was a leader in such large-scale surveillance with the REACT-1 programme, which used at-home testing of randomly selected individuals across England Imperial College London (2022). In addition, the ONS COVID-19 Infection Survey took the at-home testing approach to households across the UK to ascertain how many people presently and previously had COVID-19 (Office for National Statistics, 2020). Comparable studies were conducted in other countries, including France (Vaux et al., 2023), Germany (Jacob et al., 2021), Hungary (Merkely et al., 2020), and Iceland Gudbjartsson et al. (2020), though efforts were smaller-scale. The U.S. supported some similarly small-scale studies like the Community Prevalence of SARS-CoV-2 Study (COMPASS), which tracked around 22,000 individuals from 15 communities during 2021 (Justman et al., 2024), and the Digital Engagement and Tracking for Early Control and Treatment (DETECT) study, which collected over 35,000 self-reported test results from 2020 to 2022 (Kolb et al., 2023). These studies provided valuable insights, but their voluntary nature meant they were susceptible to participation bias. Large-scale studies like REACT-1 or ONS are costly due to extensive sampling and mass testing (Pavelka et al., 2021), so alternatives, such as seroprevalence from routine blood sampling, may be more sustainable over time (COVID-19 Immunity Task Force, 2023). While infection prevalence studies are important for real-time monitoring, their cost, cross-sectional nature, and logistical complexity limit their feasibility for long-term tracking and retrospective analysis. Consequently, leveraging routine data streams already integrated with existing infrastructure can give a more clear and consistent view of the pandemic's progression, while enhancing transparency and public participation.

Contextualizing the course of the pandemic, understanding the effects of interventions, and drawing insights for future pandemics is challenging because the spatial and temporal behavior of infections is unknown. While reported cases provide a convenient proxy of the disease burden in a population, it is incomplete, delayed, and misrepresents the true size and timing of the pandemic. Regardless of these difficulties, it is important to the public and to public health to perform a pandemic post-mortem. Estimates of daily incident infections are one such way to measure this and can guide understanding of the pandemic burden over space and time.

In this work, we provide a data-driven reconstruction of daily incident infections for each U.S. state before the onset of Omicron. Using state-level line list data, we estimate state-date specific distributions for the delay from symptom onset to positive specimen date and positive specimen to case report date. We combine these with variant-specific incubation period distributions to deconvolve (i.e. push back using the distribution of delays) daily reported COVID-19 cases back to their infection onset, removing the effects of the delays. Finally, we adjust for unreported infections with seroprevalence and reinfection data, accounting for the waning of antigenic immunity over time. This last stage is the key contributor to uncertainty quantification for the estimated incident infections. A graphical depiction of our procedure is shown in Fig. 2. We also provide a brief description of the key data sources used in our procedure in Appendix A.

Our results examine the features of the infection estimates and the implications of using them, rather than reported cases, to assess the impact of the pandemic in U.S. states. We also calculate simple time-varying infection-hospitalization ratios (IHRs) and infection-fatality ratios (IFRs) for each state and compare them with their case-based counterparts: the case-hospitalization ratios (CHRs) and case-fatality ratios (CFRs). While these analyses provide a glimpse into the utility of our infection estimates, we believe that there is much more to be explored, and we hope that our work serves as a benchmark for future retrospective analyses.

2. Methods

In what follows, we describe how we estimate the daily incident infections for each U.S. state from June 1, 2020 to November 29, 2021. Fig. 2 summarizes the major analysis tasks. First, we estimate the delays from positive specimen to report date and use them to push back the reported cases to their sample collection dates. Next, we estimate the delay from symptom onset to sample collection, combine this with variant-specific infection-to-symptom delays, and use these to push back the cases to infection onset. The resulting case estimates are aggregated across variant categories and adjusted by the case ascertainment ratio, estimated with seroprevalence survey data and a model for antigenic immunity.

2.1. From reported cases to positive specimen collection

Deconvolution "pushes back" reported cases to the likely date of positive specimen collection. An important aspect of our methods is that deconvolution is not the same as a simple shift, rather it involves the distribution of delays (specific to each state and date). Simply shifting cases back in time would fail to reflect the fact that some cases take much longer to be reported than others (Appendix B).

We begin by describing how the model for deconvolution infers the likely dates of positive specimen collection from reported cases before describing how the CDC line list (Centers for Disease Control and Prevention, 2020a) is used to estimate the necessary delay distributions. Together, these are the ingredients for Step 1 in Fig. 2. Define $y_{\ell,t}$ to be the number of new cases reported in location ℓ at time t, as reported by the John Hopkins Center for Systems Science and Engineering (JHU CSSE, Dong et al., 2020) and retrieved with the COVIDcast API (Reinhart et al., 2021). Let $\pi_{\ell,t}(k)$ be the probability that cases with positive specimen collection at time $t-k$ are reported at t. Then, we model $y_{\ell,t}$ as a Gaussian with mean

$$\mathbb{E}[y_{\ell,t} \mid x_{\ell,s}, s \le t] = \sum_k \pi_{\ell,t-k}(k) x_{\ell,t-k}, \tag{1}$$

which is a probability-weighted sum of the number of positive specimens collected k days earlier, $x_{\ell,t-k}$. We estimate $\mathbf{x}_\ell = (x_{\ell,1}, \ldots, x_{\ell,T})^\mathsf{T}$ by minimizing the negative log-likelihood with a penalty that encourages smoothness in time. Thus, our estimator is given by

$$\hat{\mathbf{x}}_\ell = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_t \left(y_{\ell,t} - \sum_k \pi_{\ell,t-k}(k) x_{t-k} \right)^2 + \lambda \sum_t \left| x_t - 4x_{t-1} + 6x_{t-2} - 4x_{t-3} + x_{t-4} \right|. \tag{2}$$

The solution to this minimization problem is an adaptive piecewise cubic polynomial (Tibshirani, 2014, 2022) and can be accurately computed easily (Ramdas and Tibshirani, 2016; Jahja et al., 2022). We

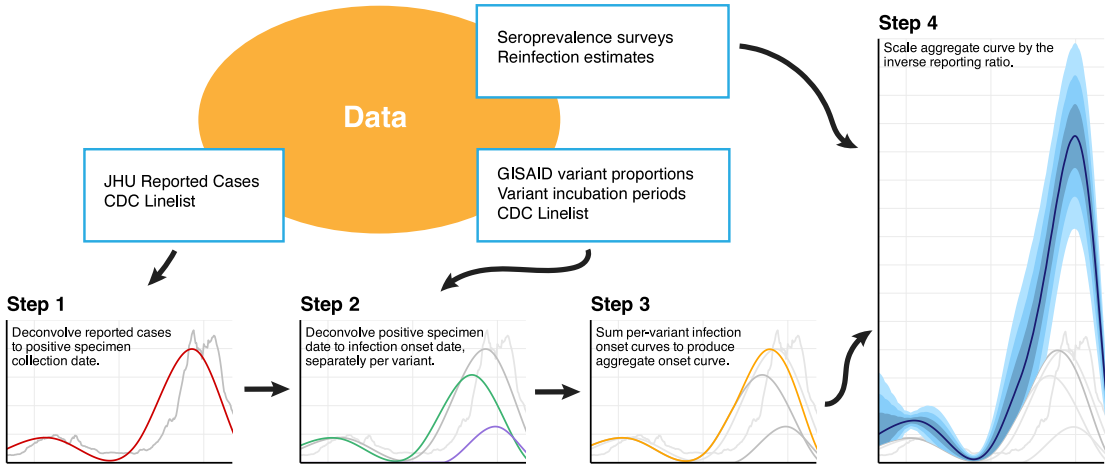

Fig. 2. Flowchart of the data and major analysis steps required to get from reported cases to incident infection estimates. In Step 1, we used the CDC line list data to deconvolve (i.e. push back using the distribution of delays) reported cases (gray) to the date of positive specimen (red). Step 2 separately deconvolved these to the date of infection by variant (Epsilon in Purple, Ancestral in Green), before summing across all variants (orange) in Step 3. Finally, we used seroprevalence survey and time-varying reinfection data to account for the unreported infections. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

select the tuning parameter λ with cross-validation to minimize the out-of-sample reconvolution error. Note that the ℓ_1 penalty is applied to all 4^{th}-order discrete differences. This choice results in piecewise cubic polynomials (discrete splines, Tibshirani, 2022) where the number and length of each piece is chosen adaptively. This property enables the model to better capture localized changes in the estimated infections. In contrast, a squared ℓ_2 penalty would produce smoothing splines resulting in global smoothness.

To estimate the $\pi_{\ell,t}(k)$ for all states ℓ, times t, and delays k, we use the CDC line list (Centers for Disease Control and Prevention, 2020a). The line list contains three key dates of interest for many cases that eventually appear in case reports: symptom onset, positive specimen collection, and report to the CDC. Handling missingness in these dates requires careful attention (Appendix C). Define $z_{\ell,t}$ to be a case report occurring at time t in location ℓ. We assume that positive samples are reported within 60 days and that no test is reported on the same date as it was collected. Under these assumptions, let $N_{\ell,t}$ be the total number of $z_{\ell,r}$ with positive specimen collection date r in a window $r \in [t-75+1, t+60]$ around t. Then, we compute the observed probability mass function (pmf)

$$\tilde{p}_{\ell,t}(k) = \frac{1}{N_{\ell,t}}\big(\#\ z_{\ell,r} \text{ with positive specimen at } r-k\big)\mathbf{1}(0 < k \le 60), \tag{3}$$

where $\mathbf{1}(Z) = 1$ if Z is true and 0 otherwise. We also compute a similar national pmf, $\tilde{p}_t(k)\mathbf{1}(0 < k \le 60)$, without restricting to location ℓ. Next, let $\alpha_{\ell,t}$ be the ratio of $N_{\ell,t}$ to the number of reported cases in the window $[t-60+2, t+75]$. Then, we compute $p_{\ell,t} = \alpha_{\ell,t}\tilde{p}_{\ell,t} + (1-\alpha_{\ell,t})\tilde{p}_t$. This construction allows for more reliance on the state estimate when a larger fraction of case reports appear in the CDC line list. We calculate the mean $m_{\ell,t}$ and variance $v_{\ell,t}$ of the pmf $\{p_{\ell,t}(k)\}$ and estimate the best-fitting gamma distribution by solving the moment equations $m_{\ell,t} = \alpha_{\ell,t}\theta_{\ell,t}$ and $v_{\ell,t} = \alpha_{\ell,t}\theta_{\ell,t}^2$ for the shape $\alpha_{\ell,t}$ and scale $\theta_{\ell,t}$. Finally, we discretize the resulting gamma density to the original support to produce an estimate $\hat{\pi}_{\ell,t}(k)$ of the delay distribution $\pi_{\ell,t}(k)$. This simple procedure of estimating the delay distribution by fitting a probability distribution to observed delays is common but may be susceptible to certain types of bias due to censoring and other considerations (Park et al., 2024). Additional details are deferred to Appendix D.

2.2. From positive specimen collection to infection onset

To continue, pushing positive specimen collection time back to infection onset (Step 2 in Fig. 2), we use a procedure very similar to that described above and specified in Eqs. (1) and (2). However, because the delays involve the time from infection to symptom onset, these are treated as variant-specific. We use our estimates from Section 2.1, $\hat{\mathbf{x}}_\ell$, but we weight them corresponding to the mix of variants in circulation. To estimate the daily proportions of the variants circulating in each state, we use GISAID genomic sequencing data from CoVariants.org (Hodcroft, 2021; Elbe and Buckland-Merrett, 2017), and estimate a multinomial logistic regression model. This procedure is now standard (Obermeyer et al., 2022; Annavajhala et al., 2021; Figgins and Bedford, 2021, see Appendix E for additional details). The resulting estimated probability of variant j is given by $\hat{v}_{j\ell,t}$.

To estimate variant-specific delays from infection to positive specimen collection, we convolve the location-time-specific symptom-to-test distributions (that are estimated from the CDC line list in the same way as in Section 2.1), with variant-specific incubation periods. The convolution of these yields a distribution $\hat{\tau}_{j\ell,t}(k)$. Details on the convolution and its inputs are in Appendices F.1–F.3.

Analogous to Eqs. (1) and (2), for each variant j, we model the variant-specific, deconvolved cases as Gaussian with mean

$$\mathbb{E}\left[\hat{v}_{j\ell,t}\hat{x}_{\ell,t} \mid u_{j\ell,s}, s \le t\right] = \sum_k \hat{\tau}_{j\ell,t-k}(k)u_{j\ell,t-k} \tag{4}$$

and estimate $\mathbf{u}_{j\ell}$ by minimizing the negative log-likelihood with a penalty to promote smoothness:

$$\tilde{\mathbf{u}}_{j\ell} = \underset{\mathbf{u}}{\operatorname{argmin}} \sum_t \left(\hat{v}_{j\ell,t}\hat{x}_{\ell,t} - \sum_k \hat{\tau}_{j\ell,t-k}(k)u_{t-k}\right)^2 + \lambda \sum_t \left|u_t - 4u_{t-1} + 6u_{t-2} - 4u_{t-3} + u_{t-4}\right|. \tag{5}$$

We call the solution $\tilde{\mathbf{u}}_{j\ell}$ the *variant-specific deconvolved cases* and emphasize that these are cases that were eventually reported to public health. Because this deconvolution is performed separately for each location and variant, we sum over the variants at each time t, and denote the total deconvolved cases at location ℓ as $\hat{\mathbf{u}}_\ell = \sum_j \tilde{\mathbf{u}}_{j\ell}$ (Step 3 in Fig. 2). These deconvolved cases are now indexed by the time of infection onset rather than case report.

2.3. Inverse reporting ratio and the antibody prevalence model

To capture the unreported infections, it is necessary to adjust these deconvolved case estimates by the inverse reporting ratio: the ratio of the number of incident infections to incident reported infections (Step 4 in Fig. 2). Seroprevalence of anti-nucleocapsid antibodies represents the percentage of people who have at least one resolving or

past infection (Centers for Disease Control and Prevention, 2020b), so we develop a model that uses the change in subsequent seroprevalence measurements to estimate all new infections. We use two seroprevalence surveys to estimate the proportion of the population with evidence of previous infection in each state over time (Centers for Disease Control and Prevention, 2021a,b, see also Appendix G).

To account for different surveys occurring on different dates with roughly weekly availability and measurement error, we treat actual seroprevalence $s_{\ell,m}$ as a latent variable available on Monday (using m rather than t to denote Mondays). Therefore, the observed seroprevalence survey measurements r_m^1 and r_m^2 are modeled as Gaussian,

$$r_{\ell,m}^1 \mid s_{\ell,m},\ w_{\ell,m}^1 \sim \mathrm{N}(s_{\ell,m},\ w_{\ell,m}^1\sigma_{\ell,r}^2), \tag{6}$$

$$r_{\ell,m}^2 \mid s_{\ell,m},\ w_{\ell,m}^2 \sim \mathrm{N}(s_{\ell,m},\ w_{\ell,m}^2\sigma_{\ell,r}^2), \tag{7}$$

with source-specific measurement errors, $w_{\ell,m}^1$ and $w_{\ell,m}^2$, that scale proportionally to sampling uncertainty.

To complete the model, we assume that latent seroprevalence is modeled as Gaussian with mean given by a fraction of the previous seroprevalence measurement at time m plus the reinfection-adjusted deconvolved cases multiplied by the inverse reporting ratio at time m:

$$\mathbb{E}[s_{\ell,m+1} \mid s_{\ell,m}] = (1-\gamma)s_{\ell,m} + a_{\ell,m}(1-z_m)\sum_{t\in[m,m+1]}\hat{u}_{\ell,t}, \tag{8}$$

where $\hat{u}_{\ell,t}$ are deconvolved cases (from Section 2.2), z_m is the fraction of reinfections, and $a_{\ell,m}$ is the inverse reporting ratio. Note that γ is the fraction of people whose level of infection-induced antibodies falls below the detection threshold between time t and time $t+1$. The daily fraction of new infections z_t are based on surveillance work conducted by the Southern Nevada Health District (Ruff et al., 2022), and these estimates are broadly similar to those in other locations with available data (Ruff et al., 2022; New York State Department of Health, 2023; Hawaii Department of Health, 2022; Washington State Department of Health, 2022). Finally, we specify the time-varying evolution of the inverse reporting ratio as Gaussian with expectation,

$$\mathbb{E}[a_{\ell,m+1} \mid a_{\ell,m},\ a_{\ell,m-1},\ a_{\ell,m-2}] = 3a_{\ell,m} - 3a_{\ell,m-1} + a_{\ell,m-2}. \tag{9}$$

This construction for Eq. (9) results in estimates that vary smoothly in time.

The antibody prevalence model specified by Eqs. (6) to (9) is a state space model with latent variables $\mathbf{s}_\ell$ and $\mathbf{a}_\ell$. In this way, the latent variables and all unknown parameters can be estimated using maximum likelihood, despite missing or irregularly-spaced survey measurements. Additionally, latent quantities can be extrapolated beyond the times of measured seroprevalence. Importantly, the specification of Eqs. (6) to (9) naturally captures uncertainty in the estimates of the $\mathbf{a}_\ell$ curves around their point estimates. While Steps 1–3 in our analysis pipeline (Fig. 2) are also point estimates, and hence uncertain, their relative contribution to the uncertainty of estimated infections is much less than that of the seroprevalence and ascertainment fraction estimates. Additional details of this methodology and the computation of the associated uncertainty measurements are in Appendix H.

2.4. Lagged correlation and time-varying IHRs and IFRs

We use the COVIDcast API (Reinhart et al., 2021) to retrieve the daily number of confirmed COVID-19 hospital admissions for each state that are collected by the U.S. Department of Health and Human Services (HHS). We use our infection estimates $\hat{\mathbf{u}}_\ell$ to compute the lagged correlation with hospitalizations. The goal of this analysis is to find the lag between infection and hospitalization rates that gives the highest average rank-based correlation across U.S. states. Thus, we consider a wide range of possible lag values ranging from 1 to 25 days. Then, to assess the impact of our modeling choices, particularly the contribution of the main steps to the lagged correlation analysis, we conduct an ablation study that is detailed in Appendix I.

For each considered lag, we calculate Spearman's correlation between the daily state infection and hospitalization rates from June 1, 2020, to November 29, 2021, using a center-aligned rolling window of 61 days. We then average these correlations across all states and times for each lag.

The lag $\hat{\ell}$ that leads to the highest average correlation is used to estimate the time-varying IHRs for each state. For each time t, IHR(t) can be computed by dividing the number of individuals who were hospitalized due to COVID-19 on day t by the estimated total number of those who were infected on day $t-\hat{\ell}$. However, to stabilize these IHR estimates, we use the averages of hospitalizations and infections within a window of 61 days centered on the date of interest (t and $t-\hat{\ell}$ respectively).

To evaluate the impact of window size on the resulting stabilized IHR and CHR estimates, we conduct a sensitivity analysis by varying the window size from 15 days to 91 days. The results of this analysis are reported in Appendix J.

The same procedure is used to compute IFRs, with a few data-driven adjustments. New deaths due to COVID-19 for each state and collected by the National Center for Health Statistics (NCHS) are reported weekly rather than daily. We use these death counts rather than those from JHU CSSE because they are aligned by the date of death not the date of report. Since these are weekly while the infection estimates are daily, we convert them to daily by proportionally allocating the weekly counts across the intervening days. First, all weekly totals are divided by 7 and missing weeks are imputed with the previous week's value. Then, since each weekly total corresponds to a Sunday, we use a weighted combination of the preceding Sunday's and the following Sunday's values for the intervening weekdays—for example, a Tuesday is imputed of 5/7 of the preceding Sunday's value and 2/7 of the following Sunday's value.

To identify the optimal lag, we consider a larger range of lag values than for hospitalizations, from 1 to 35 days, as deaths typically follow hospitalization. As with the IHRs, we perform an ablation study to evaluate the effects of our modeling choices (Appendix K). We compute Spearman's rank-based correlation between the state infection and death rates using a 91-day center-aligned rolling window rather than 61 days, and proceed similarly for the construction of time-varying IFRs. A sensitivity analysis varying the window size from 31 to 101 days is provided in Appendix L.

3. Results

3.1. Infection estimates and cases-to-infections ratios across the U.S.

Prior to Omicron, the largest infection outbreaks occurred in the late summer and early fall of 2021 in Louisiana, Georgia, Idaho, and Montana (Figs. 3 to 4). During this time, the state with the highest rate of infections on a single day is Louisiana, with 476 infections per 100K (95% CI: 294, 658) on July 20, 2021. For comparison, the state's 7-day average case rate peaks at 126 cases per 100K on August 13, 2021. Idaho follows with an infections peak of 457 per 100K (95% CI: 319, 595) on September 7, 2021, and a case peak of 76 per 100K occurring shortly thereafter on September 13, 2021. The period of lowest viral transmission is in the summer of 2020, when Vermont has fewer than 10 infections per 100K per week from June to August, the longest such lull for any state.

Nearly all states exhibit two major waves in infections—the Ancestral wave began in the fall of 2020 and extended into the winter season, while the Delta wave started in the late summer of 2021 and continued into mid-fall. In general, greater similarities in the strength and magnitude of outbreaks emerge in small clusters of states that border each other (Idaho and Montana; North and South Carolina), which present waves of infections that mirror each other in amplitude and timing.

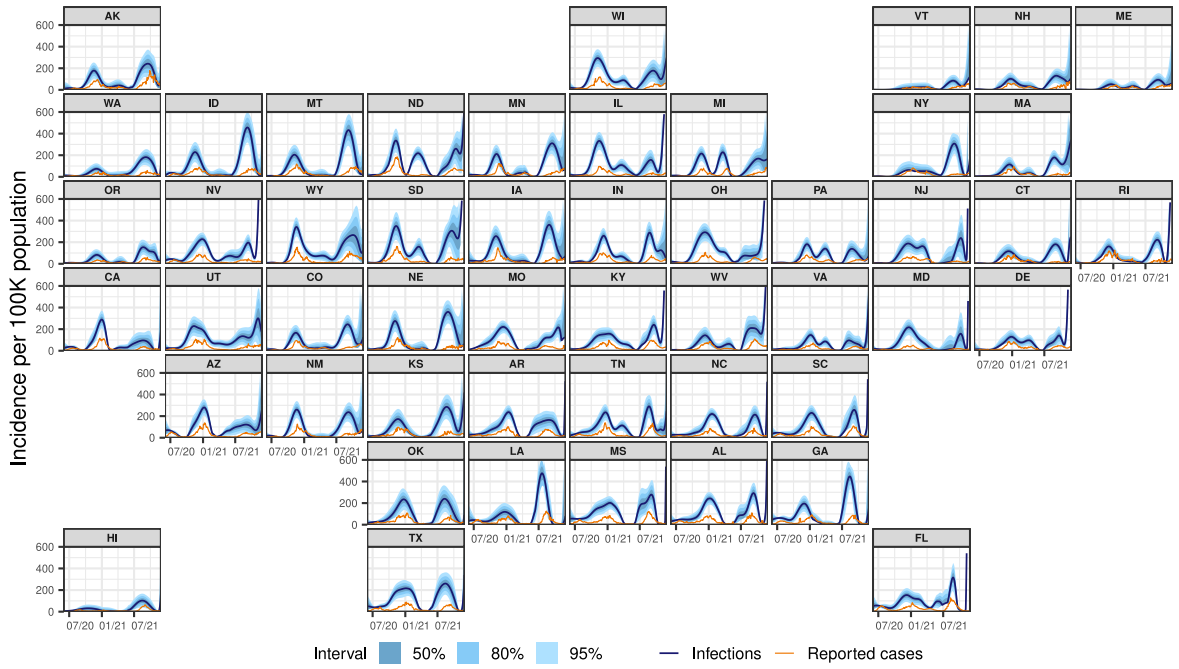

Fig. 3. Estimates of the daily new infections per 100,000 population for each U.S. state from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% intervals for the estimates, while the orange line represents the trailing 7-day average of reported cases per 100,000. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

While the Ancestral, Alpha, and Delta waves are visible for most states, there are clear outbreaks in unreported infections that are not easily detectable from cases alone. For example, a wave of infections is evident in North and South Dakota over the spring of 2021 that is virtually undetectable from reported cases. Similarly, in late-summer 2021, the Delta wave is only faintly detectable from cases in a number of Northeastern states, while infections suggest that it has already begun in earnest.

Moreover, cases severely underestimate infections during Delta for many states, more so than in earlier waves (Fig. 3). The most extreme is New Jersey, where about 6.3% of estimated infections were eventually reported as cases. Similarly low are Maryland (7.3%), Nevada (8.4%), and South Dakota (10.0%). In 44 states, fewer than 1/3 of infections eventually appeared in case reports. The cases-to-infections ratio is larger in earlier waves, and its effects are most apparent in different regions. During Alpha, Louisiana has the lowest ratio of infections to cases (11.9%) followed by California (13.6%). Such patterns are less apparent during the Ancestral wave, where Ohio and Maryland have the lowest ratio of reported cases to infections at 21.4% and 21.7%, respectively.

Fig. 5 shows that using cases as a proxy for infections can lead to misunderstandings in the locations that are affected and the extent to which they are affected. For example, on October 20, 2020, while case rates are elevated in a handful of upper-Midwestern states (namely, North and South Dakota), infection rates are elevated to a similar extent in the surrounding states as well, indicating a wider impact than suggested by cases alone. On July 20, 2021, while the map of case rates shows low and geographically consistent impact, infection rates reveal that Texas, Louisiana, Georgia, and their neighbors are hotspots.

By focusing on states with elevated cases, infection outbreaks may be overlooked. For instance, on August 27, 2021, Montana and Idaho have some of the highest infection rates (Fig. 5). In contrast, their case rates are unremarkable (the highest case rates tend to be in the Southeast). Infection outbreaks tend to precede case outbreaks, though the lead time can vary widely. During the Delta wave, infections in Montana peak about 41 days before cases, while in Idaho, they peak about 6 days before cases (Fig. 3). During the Ancestral wave, infections peak about 12 days earlier than cases in Montana and 24 days earlier in Idaho, demonstrating a notable shift in lead times.

3.2. Cross-correlations, IHRs, IFRs, CHRs, and CFRs

The maximum Spearman's correlation between infections and hospitalizations is 0.48 and occurs at a lag of 13 days (Fig. 6, left panel). In contrast, we find that the largest average Spearman correlation for cases is 0.69 and occurs at a lag of 1 day. That is, case reports are nearly contemporaneous to hospitalizations, while infection estimates clearly precede them. For infections and deaths, the maximum Spearman's correlation is 0.57 and occurs at a lag of 24 days (Fig. 6, right panel), whereas for cases and deaths it is 0.75 at 10 days. Thus, the maximum correlation for infections occurs 14 days earlier than for cases.

We compute the time-varying infection-hospitalization ratios (IHRs) for each state using a 13-day lag and case-hospitalization ratios (CHRs) with a 1-day lag for comparison (Fig. 7). Overall, the relationship between infections and hospitalizations is complex. It is characterized by intermittent spikes that punctuate longer periods where the IHRs are relatively stable, remaining below 0.1 hospitalizations per infection. This pattern of stability and relative comparability across states is supported by the median and 95% confidence intervals for the IHRs described in Appendix M. This trend is also evident in the IFRs, computed using the optimal 24-day lag. These IFRs are generally more stable and smaller than the corresponding CFRs, remaining below 0.02 deaths per infection (Fig. 8 and Appendix M).

While these trends persist for larger window sizes, it is important to acknowledge that the window size plays an important role in smoothing: larger window sizes result in more smoothing, leading to more tapered peaks, whereas smaller window sizes reduce smoothing, resulting in more pronounced peaks and larger intermittent spikes, as detailed in Appendices J and L.

The IHRs and CHRs exhibit similar spatiotemporal trends as those noted above for infections. Namely, states that are proximate (for example, North and South Carolina) show similar temporal patterns

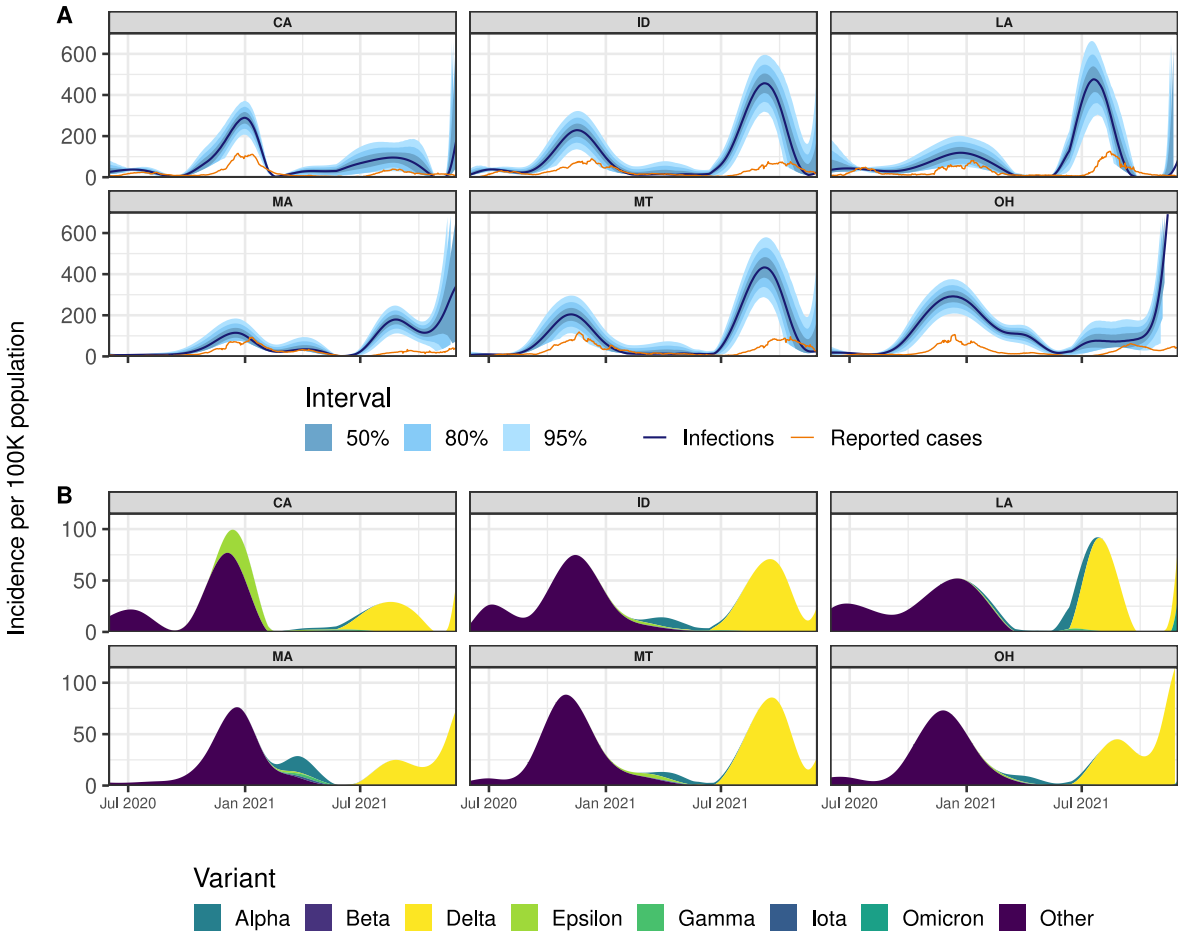

Fig. 4. Panel A: Reported cases (orange) and estimates of daily new infections (dark blue) per 100K inhabitants. The blue shaded regions indicate 50, 80, and 95% confidence bands. Panel B: Deconvolved cases colored by variant per 100K inhabitants. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

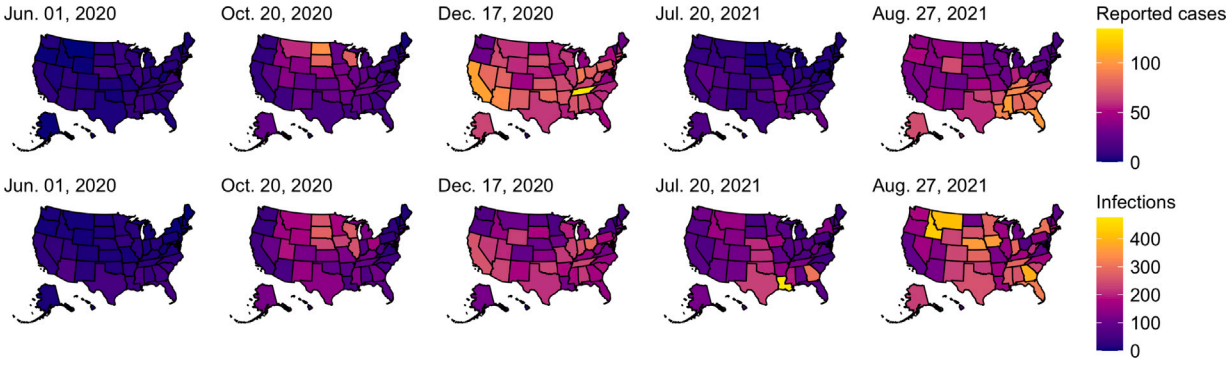

Fig. 5. Choropleth maps of the state-level estimates of daily new cases per 100K (top row) and daily new infections per 100K (bottom row) for five select dates between June 1, 2020 and November 29, 2021. Note that the first date is chosen as a baseline, while the other dates are chosen because they present large counts of infections across all states. In particular, the third and fifth dates show the largest number of total infections across the 50 states within those calendar years. For a more dynamic visualization of the case and infection trends over time, videos depicting the infection and case choropleth maps over the entire period are available in the GitHub repository for this work.

in IHRs and CHRs. In addition, similar spikes are evident across many states during waves of infections that are driven by variants of concern. Many states exhibit a striking increase in hospitalizations in mid-2021, which coincides with the rapid takeover of the Delta variant (Hodcroft, 2021). These trends are less pronounced in the IFRs and CFRs, with notable exceptions such as North and South Dakota, where there is a sharp spike in IFRs and CFRs in mid-2021, also coinciding with the spread of the Delta variant.

4. Discussion

We retrospectively estimated daily incident infections for each U.S. state over the period June 1, 2020 to November 29, 2021. Our estimates support the intuition that the pandemic impacted states earlier and at a larger scale than is indicated by reported cases. They also emphasize that using cases as a proxy for infections can lead to erroneous conclusions about trends in infections. More importantly, we observe

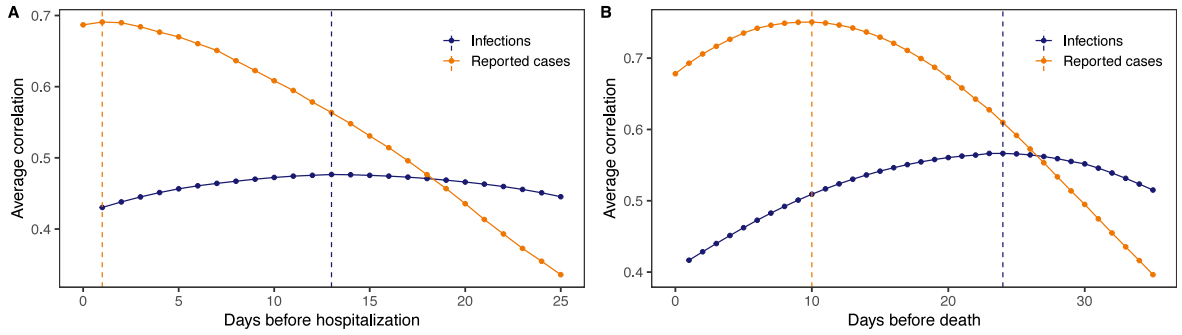

Fig. 6. Left panel: Spearman's rank correlation between each of infections and cases with *hospitalizations* per 100,000. A rolling window of 61 days is applied before averaging across all states and times for each lag. Right panel: Spearman's rank correlation between each of infections and cases with *deaths* per 100,000. A rolling window of 91 days is applied before averaging across all states and times for each lag. Both panels: The vertical dashed lines indicate the lags for which the highest average correlation is attained.

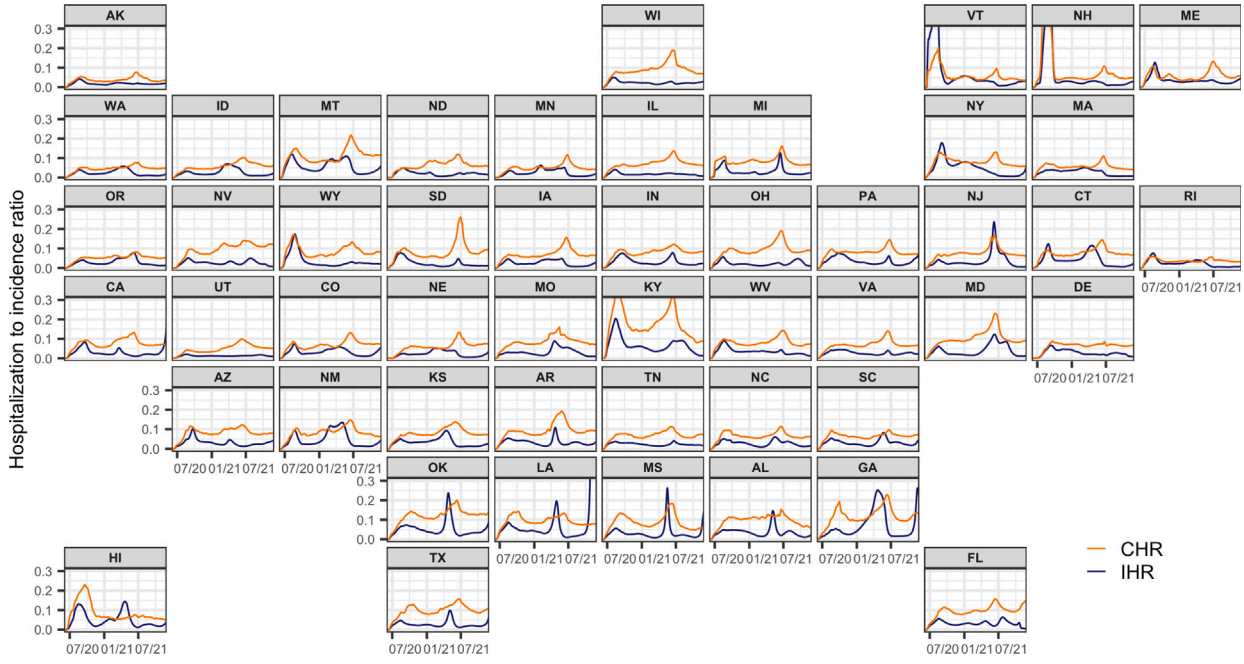

Fig. 7. Time-varying IHR and CHR estimates for each state from June 2020 to November 2021, calculated using the correlation-maximizing lags from Section 3.2. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious day of the week effects. Also note that the different starting points across states are due to the availability of the hospitalization data.

outbreaks in infections that are missed from inspecting cases alone such as the Delta wave in New Jersey, Connecticut, and Maryland. These sorts of omissions serve to emphasize that cases paint an incomplete picture of the pandemic, especially when outbreaks are largely driven by unreported infections. Furthermore, since case reports generally follow symptom and infection onsets, cases have a built-in temporal bias. This is in addition to other biases from differences in reporting across states such as temporary bottlenecks due to influxes of data or more persistent processing issues that increase the average time from case detection to report (Washington State Department of Health, 2020; Dunkel, 2020). Thus, while reported cases provide an indication of the trajectory of the pandemic, it is delayed and incomplete.

Case reporting varied widely over time and across locations globally, not just in the U.S., leading to misrepresentations of disease burden and complicating the development of evidence-based policy decisions. This geographical variation, driven by differing testing capacities and reporting standards, makes it difficult to establish accurate infection trends and investigate appropriate responses. Misleading spatial patterns of reported cases, such as overrepresentation in some areas and underrepresentation in others (a glimpse of which was shown in Fig. 5), further exacerbate the difficulty in evaluating the true scope of the pandemic and likely hindered effective policy decisions.

While countries like England addressed these case-based challenges through large-scale infection-prevalence surveys (e.g. the REACT-1 study, Imperial College London, 2022), these efforts are limited by the biases inherent to the study design, specifically those related to response rates and population coverage. For example, the response rate varied between 11.7% and 30.5% in REACT-1, with demographic variation, particularly in relation to age (Elliott et al., 2023). Additionally, such studies are expensive to design and implement, requiring substantial funding for participant recruitment, data collection, and analysis. Costs are further compounded due to comprehensive national sampling as well as mass PCR-based testing (Pavelka et al., 2021). By leveraging routine data streams, our methodology provides a less expensive, more sustainable way to track infections over time. Thus, our approach has the potential to be adapted into a cost-effective, real-time alternative for pandemic monitoring, offering timely estimates to inform policymakers and overcoming the biases seen in case reporting.

Though this paper focused on retrospective estimation, this approach could be extended to the real-time case. However, such an extension would be challenging due to data-driven issues, particularly

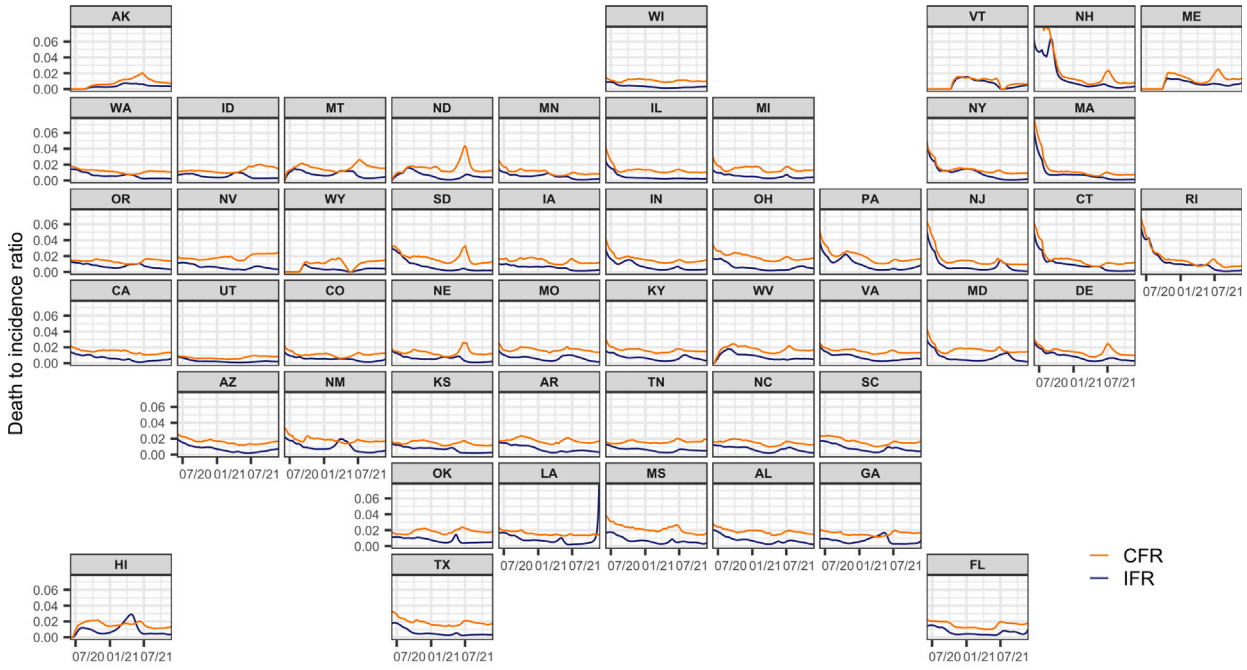

Fig. 8. Time-varying IFR and CFR estimates for each state from June 2020 to November 2021, calculated using the correlation-maximizing lags from Section 3.2. Note that the infection, case, and death counts are calculated with a center-aligned 7-day average to remove spurious day of the week effects.

reporting delays and right censoring. For instance, Jahja et al. (2022) explores estimating infections in real-time using the CDC line list data set in a similar deconvolution-based approach. They highlight complexities with using this data set for estimation, including monthly updates and right censoring. To handle the right censoring problem, they created a Kaplan–Meier-like method and they applied specialized regularization techniques to manage fluctuations in estimates around the date of interest. Similar techniques would likely need to be applied to manage the CDC line list dataset if we were to extend our approach to a real-time analysis. For a brief overview of the other main data sources we used, their data collection period, update frequency, and major sources of delay, which are factors that would likely impact right censoring and increase volatility in real-time estimation, see Appendix A.

Our approach offers a number of additional advantages. By incorporating state-level case, line list, and variant circulation data, we are able to construct incubation and delay distributions that are spatiotemporally specific. Time-varying and state-specific seroprevalence data allows the reporting ratio estimates to similarly vary over space and time, a departure from existing work (Unwin et al., 2020; Center for the Ecology of Infection Diseases, 2020). Unlike previous approaches that use a single delay distribution to generate estimates for all states (Chitwood et al., 2022; Jahja et al., 2022; Miller et al., 2022), our work avoids this assumption of geographic invariance, an assumption that is far from realistic due to differences in the reporting pipelines, pandemic response, and variants in circulation, among other things. Similarly, prior methodology relies on only one incubation period distribution (Miller et al., 2022), whereas our method incorporates variant-specific incubation periods. This enhances our infection onset estimation by accounting for the differences across variants–specifically, that newer variants tend to have shorter incubation periods (Tanaka et al., 2022; Ogata et al., 2022; Wu et al., 2022).

Another limitation of previous approaches to estimate infections is that they often fail to account for reinfections. While reinfections constitute a small portion of the total infections until the arrival of high immune-escape variants (Omicron BA.1), disregarding them means that the infection-reporting ratio will tend to be underestimated with seroprevalence data alone. By accounting for reinfections as well as the waning of seropositivity, we more accurately estimate this ratio. Future work could further refine this analysis. Because the waning of immunity is likely to be variant-dependent (Pooley et al., 2023), our model's single waning parameter would be more accurately estimated as a mixture of variant-specific parameters with weights determined by the proportion of the variants circulating.

While we did consider the waning of seropositivity in our modeling procedure, we did not explicitly consider the relationship between the epidemic growth rates and the probability of positive tests. Specifically, during upswings in an epidemic, a greater proportion of infections come from individuals early in their infection, who have higher viral loads and are more likely to test positive (Frediani et al., 2024). During a decline, the opposite occurs. This dynamic relationship between infection stage, viral load, and test positivity has been well-described by Hay et al. (2021). This effect may be substantial and could influence the accuracy of the deconvolved case and infection estimates, especially during rapid changes in the epidemic trajectory.

Another area for potential improvement is the use of the constant-lag relationship between infections and hospitalizations. We used a single lag in our analysis, emphasizing the contrast between case-based calculations and those based on estimated infections. However, using a single lag is problematic as it is likely to vary over space and time (Imperial College London, 2022) due to changing variants, local public health availability, and policy decisions, among other reasons. Additionally, this fixed lag of estimator is likely to create bias, even if correctly specified, and should generally be avoided (Goldwasser et al., 2024). Future work could explore multiple lags or heterogeneous convolutional estimators (Overton et al., 2022) using distributions of delays rather than single lags.

Another aspect of our analysis that warrants consideration is the time period under study. We chose to end our analysis on November 29, 2021, for two main reasons. The first is that Omicron and subsequent variants come with substantial increases in the risk of reinfection in comparison to previous variants, likely due to increased immune escape (Wei et al., 2024; Pulliam et al., 2022; Eythorsson et al., 2022). Access to reinfection data that is representative of each location under study is paramount for extending the analysis. While it would be ideal

to use the reinfection rates over time for each U.S. state, many states do not publicly report reinfection data over the entire time period under examination, if at all.

The second reason is that the case-ascertainment ratio after December 2021 can no longer be estimated with seroprevalence data alone. Specifically, while most state-level data suggests that reinfections still account for less than 20% of reported cases during Omicron (Ruff et al., 2022; New York State Department of Health, 2023; Hawaii Department of Health, 2022; Washington State Department of Health, 2022), seropositivity rapidly reaches nearly 100% of the population. Therefore, alternative data sources for estimating the case-ascertainment ratio must be considered. For example, wastewater surveillance data may be complementary to seroprevalence data, especially when testing is low, or serve as a substitute when it is unavailable (McManus et al., 2023). An alternative approach could integrate surveillance streams from surveys, helplines, or medical records if they offer a sufficiently strong signal of the disease intensity over time (Reinhart et al., 2021; European Centre for Disease Prevention and Control, 2020).

Our work develops a deconvolution-based approach to inferring infection onset, combining available line list data with variant circulation estimates and literature derived incubation periods. This approach is complemented with the development of a model that incorporates waning detectable antibody levels and major seroprevalence surveys. The resulting infection estimates as well as their geospatial and temporal trends are strongly grounded in both data and statistical models.

These well-informed, localized estimates of COVID-19 infections provide a clear and comprehensive understanding of the pandemic's progression over time. They contribute important information on the timing and magnitude of the disease burden for each location, and highlight trends that may not be visible from reported case data alone. Therefore, these infection estimates provide key information for the ongoing investigation on the true size and impact of the pandemic.

CRediT authorship contribution statement

Rachel Lobay: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Ajitesh Srivastava:** Writing – review & editing, Methodology. **Ryan J. Tibshirani:** Writing – review & editing, Methodology, Conceptualization. **Daniel J. McDonald:** Writing – review & editing, Visualization, Software, Methodology, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank members of the Delphi research group for valuable feedback, and Change Healthcare and Optum/United Health Group for their invaluable data partnership and collaboration.

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative (Elbe and Buckland-Merrett, 2017), on which this research is based.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and the Centers for Disease Control and Prevention.

DJM and RJT were supported by Centers for Disease Control and Prevention (CDC), United States Grant No. 75D30123C15907. DJM and RL received support from the National Sciences and Engineering Research Council of Canada and the University of British Columbia, Canada. AS was supported by the Centers for Disease Control and Prevention and the National Science Foundation, United States under Award No. 2223933 and 2333494.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.epidem.2025.100838.

Data availability

The required materials and code for reproducing all figures and the numerical results are available at https://github.com/cmu-delphi/latent-infections/.

References

Annavajhala, M.K., Mohri, H., Wang, P., Nair, M., Zucker, J.E., Sheng, Z., Gomez-Simmonds, A., Kelley, A.L., Tagliavia, M., Huang, Y., et al., 2021. Emergence and expansion of SARS-CoV-2 B. 1.526 after identification in New York. Nature 597 (7878), 703–708.

Center for the Ecology of Infection Diseases, 2020. COVID-19 portal. https://www.covid19.uga.edu/nowcast.html.

Centers for Disease Control and Prevention, 2020a. COVID-19 case surveillance restricted access detailed data. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t.

Centers for Disease Control and Prevention, 2020b. COVID data tracker. https://covid.cdc.gov/covid-data-tracker/#national-lab.

Centers for Disease Control and Prevention, 2021a. 2020–2021 nationwide blood donor seroprevalence survey infection-induced seroprevalence estimates. https://data.cdc.gov/Laboratory-Surveillance/2020-2021-Nationwide-Blood-Donor-Seroprevalence-Su/mtc3-kq6r.

Centers for Disease Control and Prevention, 2021b. Nationwide commercial laboratory seroprevalence survey. https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv.

Centers for Disease Control and Prevention, 2022. Estimated COVID-19 burden. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html.

Chitwood, M.H., Russi, M., Gunasekera, K., Havumaki, J., Klaassen, F., Pitzer, V.E., Salomon, J.A., Swartwood, N.A., Warren, J.L., Weinberger, D.M., et al., 2022. Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: Results of a Bayesian evidence synthesis model. PLoS Comput. Biol. 18 (8), e1010465.

COVID-19 Immunity Task Force, 2023. Seroprevalence provides an accurate measure of SARS-CoV-2 infection compared to PCR testing. https://www.covid19immunitytaskforce.ca/seroprevalence-provides-an-accurate-measure-of-sars-cov-2-infection-compared-to-pcr-testing/.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 (5), 533–534.

Dunkel, S., 2020. COVID-19 case numbers: Why the delay in reporting? https://www.tpchd.org/Home/Components/Blog/Blog/21448.

Elbe, S., Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob. Chall. 1 (1), 33–46.

Elliott, P., Whitaker, M., Tang, D., Eales, O., Steyn, N., Bodinier, B., Wang, H., Elliott, J., Atchison, C., Ashby, D., et al., 2023. Design and implementation of a national SARS-CoV-2 monitoring program in England: REACT-1 study. Am. J. Public Health 113 (5), 545–554.

European Centre for Disease Prevention and Control, 2020. Strategies for the Surveillance of COVID-19. Technical Report, ECDC, Stockholm, Sweden.

Eythorsson, E., Runolfsdottir, H.L., Ingvarsson, R.F., Sigurdsson, M.I., Palsson, R., 2022. Rate of SARS-CoV-2 reinfection during an omicron wave in iceland. JAMA Netw. Open 5 (8), e2225320–e2225320.

Figgins, M.D., Bedford, T., 2021. SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. http://dx.doi.org/10.1101/2021.12.09.21267544, medRxiv.

Frediani, J.K., Parsons, R., McLendon, K.B., Westbrook, A.L., Lam, W., Martin, G., Pollock, N.R., 2024. The new normal: delayed peak SARS-CoV-2 viral loads relative to symptom onset and implications for COVID-19 testing programs. Clin. Infect. Dis. 78 (2), 301–307.

Goldwasser, J., Hu, A.J., Bilinski, A., McDonald, D.J., Tibshirani, R.J., 2024. Challenges in estimating time-varying epidemic severity rates from aggregate data. http://dx.doi.org/10.1101/2024.12.27.24319518, medRxiv.

Gudbjartsson, D.F., Helgason, A., Jonsson, H., Magnusson, O.T., Melsted, P., Norddahl, G.L., Saemundsdottir, J., Sigurdsson, A., Sulem, P., Agustsdottir, A.B., et al., 2020. Spread of SARS-CoV-2 in the Icelandic population. N. Engl. J. Med. 382 (24), 2302–2315.

Hawaii Department of Health, 2022. COVID-19 reinfection data. https://health.hawaii.gov/coronavirusdisease2019/files/2022/09/reinfection_report_2022-09-28.pdf.

Hay, J.A., Kennedy-Shaffer, L., Kanjilal, S., Lennon, N.J., Gabriel, S.B., Lipsitch, M., Mina, M.J., 2021. Estimating epidemiologic dynamics from cross-sectional viral load distributions. Science 373 (6552), eabh0635.

Hitchings, M.D., Dean, N.E., García-Carreras, B., Hladish, T.J., Huang, A.T., Yang, B., Cummings, D.A., 2021. The usefulness of the test-positive proportion of severe acute respiratory syndrome coronavirus 2 as a surveillance tool. Am. J. Epidemiol. 190 (7), 1396–1405.

Hodcroft, E., 2021. CoVariants: SARS-CoV-2 mutations and variants of interest. https://covariants.org.

Imperial College London, 2022. The REACT-1 programme. https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/studies/the-react-1-programme/).

Jacob, L., Koyanagi, A., Smith, L., Haro, J.M., Rohe, A.M., Kostev, K., 2021. Prevalence of and factors associated with COVID-19 diagnosis in symptomatic patients followed in general practices in Germany between March 2020 and March 2021. International Journal of Infectious Diseases 111, 37–42.

Jahja, M., Chin, A., Tibshirani, R.J., 2022. Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. Statist. Sci. 37 (2), 207–228.

Justman, J., Skalland, T., Moore, A., Amos, C.I., Marzinke, M.A., Zangeneh, S.Z., Kelley, C.F., Singer, R., Mayer, S., Hirsch-Moverman, Y., et al., 2024. Prevalence of SARS-CoV-2 infection among children and adults in 15 US communities, 2021. Emerg. Infect. Dis. 30 (2), 245.

Kolb, J.J., Radin, J.M., Quer, G., Rose, A.H., Pandit, J.A., Wiedermann, M., 2023. Prevalence of positive COVID-19 test results collected by digital self-report in the US and Germany. JAMA Netw. Open 6 (1), e2253800–e2253800.

McManus, O., Christiansen, L.E., Nauta, M., Krogsgaard, L.W., Bahrenscheer, N.S., von Kappelgaard, L., Christiansen, T., Hansen, M., Hansen, N.C., Kähler, J., et al., 2023. Predicting COVID-19 incidence using wastewater surveillance data, Denmark, October 2021–June 2022. Emerg. Infect. Dis. 29 (8), 1589.

Merkely, B., Szabó, A.J., Kosztin, A., Berényi, E., Sebestyén, A., Lengyel, C., Merkely, G., Karády, J., Várkonyi, I., Papp, C., et al., 2020. Novel coronavirus epidemic in the Hungarian population, a cross-sectional nationwide survey to support the exit policy in Hungary. Geroscience 42, 1063–1074.

Miller, A.C., Hannah, L.A., Futoma, J., Foti, N.J., Fox, E.B., D'Amour, A., Sandler, M., Saurous, R.A., Lewnard, J.A., 2022. Statistical deconvolution for inference of infection time series. Epidemiology 33 (4), 470–479.

New York State Department of Health, 2023. COVID-19 reinfection data. https://coronavirus.health.ny.gov/covid-19-reinfection-data.

Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S.F., Pyle, J.D., Yurkovetskiy, L., Bosso, M., Park, D.J., Babadi, M., MacInnis, B.L., et al., 2022. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science 376 (6599), 1327–1332.

Office for National Statistics, 2020. Coronavirus (COVID-19) infection survey (CIS): About the study. https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/covid19infectionsurveycis.

Ogata, T., Tanaka, H., Irie, F., Hirayama, A., Takahashi, Y., 2022. Shorter incubation period among unvaccinated Delta variant coronavirus disease 2019 patients in Japan. Int. J. Environ. Res. Public Heal. 19 (3), 1127.

Overton, C.E., Webb, L., Datta, U., Fursman, M., Hardstaff, J., Hiironen, I., Paranthaman, K., Riley, H., Sedgwick, J., Verne, J., Willner, S., Pellis, L., Hall, I., 2022. Novel methods for estimating the instantaneous and overall COVID-19 case fatality risk among care home residents in England. PLoS Comput. Biol. 18 (10), e1010554.

Park, S.W., Akhmetzhanov, A.R., Charniga, K., Cori, A., Davies, N.G., Dushoff, J., Funk, S., et al., 2024. Estimating epidemiological delay distributions for infectious diseases. http://dx.doi.org/10.1101/2024.01.12.24301247, medRxiv.

Pavelka, M., Van-Zandvoort, K., Abbott, S., Sherratt, K., Majdan, M., working group, C., Analỳz, I.Z., Jarčuška, P., Krajčí, M., Flasche, S., et al., 2021. The impact of population-wide rapid antigen testing on SARS-CoV-2 prevalence in Slovakia. Science 372 (6542), 635–641.

Pellis, L., Scarabel, F., Stage, H.B., Overton, C.E., Chappell, L.H., Fearon, E., Bennett, E., Lythgoe, K.A., House, T.A., Hall, I., et al., 2021. Challenges in control of COVID-19: Short doubling time and long delay to effect of interventions. Philos. Trans. R. Soc. B 376 (1829), 20200264.

Pitzer, V.E., Chitwood, M., Havumaki, J., Menzies, N.A., Perniciaro, S., Warren, J.L., Weinberger, D.M., Cohen, T., 2021. The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. Am. J. Epidemiol. 190 (9), 1908–1917.

Pooley, N., Abdool Karim, S.S., Combadière, B., Ooi, E.E., Harris, R.C., El Guerche Seblain, C., Kisomi, M., Shaikh, N., 2023. Durability of vaccine-induced and natural immunity against COVID-19: A narrative review. Infect. Dis. Ther. 12 (2), 367–387.

Pulliam, J.R., van Schalkwyk, C., Govender, N., von Gottberg, A., Cohen, C., Groome, M.J., Dushoff, J., Mlisana, K., Moultrie, H., 2022. Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. Science 376 (6593), eabn4947.

Ramdas, A., Tibshirani, R.J., 2016. Fast and flexible ADMM algorithms for trend filtering. J. Comput. Graph. Statist. 25 (3), 839–858.

Reinhart, A., Brooks, L., Jahja, M., Rumack, A., Tang, J., Agrawal, S., Al Saeed, W., Arnold, T., Basu, A., Bien, J., et al., 2021. An open repository of real-time COVID-19 indicators. Proc. Natl. Acad. Sci. 118 (51), e2111452118.

Ruff, J., Zhang, Y., Kappel, M., Rathi, S., Watkins, K., Zhang, L., Lockett, C., 2022. Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. Emerg. Infect. Dis. 28 (10), 1977.

Tanaka, H., Ogata, T., Shibata, T., Nagai, H., Takahashi, Y., Kinoshita, M., Matsubayashi, K., Hattori, S., Taniguchi, C., 2022. Shorter incubation period among COVID-19 cases with the BA. 1 Omicron variant. Int. J. Environ. Res. Public Heal. 19 (10), 6330.

The New York Times, 2020. Coronavirus in the U.S.: Latest map and case count. https://www.nytimes.com/interactive/2021/us/covid-cases.html.

The Washington Post, 2020. Tracking U.S. COVID-19 cases, deaths and other metrics by state. https://www.washingtonpost.com/graphics/2020/national/coronavirus-us-cases-deaths/?state=US.

Tibshirani, R.J., 2014. Adaptive piecewise polynomial estimation via trend filtering. Ann. Statist. 42 (1), 285–323.

Tibshirani, R.J., 2022. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. Found. Trends Mach. Learn. 15 (6), 694–846.

Unwin, H.J.T., Mishra, S., Bradley, V.C., Gandy, A., Mellan, T.A., Coupland, H., Ish-Horowicz, J., Vollmer, M.A., Whittaker, C., Filippi, S.L., et al., 2020. State-level tracking of COVID-19 in the United States. Nat. Commun. 11 (1), 6189.

Vaux, S., Gautier, A., Soullier, N., Levy-Bruhl, D., 2023. SARS-CoV-2 testing, infection and places of contamination in France, a national cross-sectional study, December 2021. BMC Infect. Dis. 23 (1), 279.

Washington State Department of Health, 2020. COVID-19 data dashboard. https://doh.wa.gov/emergencies/covid-19/data-dashboard.

Washington State Department of Health, 2022. Reported COVID-19 reinfections in Washington State. https://doh.wa.gov/sites/default/files/2022-02/421-024-ReportedReinfections.pdf.

Wei, J., Stoesser, N., Matthews, P.C., Khera, T., Gethings, O., Diamond, I., Studley, R., Taylor, N., Peto, T.E., Walker, A.S., et al., 2024. Risk of SARS-CoV-2 reinfection during multiple Omicron variant waves in the UK general population. Nat. Commun. 15 (1), 1008.

Wu, Y., Kang, L., Guo, Z., Liu, J., Liu, M., Liang, W., 2022. Incubation period of COVID-19 caused by unique SARS-CoV-2 strains: a systematic review and meta-analysis. JAMA Netw. Open 5 (8), e2228008–e2228008.