Retrospective estimation of latent COVID-19 infections before Omicron in the U.S.

Rachel Lobay[a,1], Ajitesh Srivastava[b], Ryan J. Tibshirani[c], and Daniel J. McDonald[a]

[a]Department of Statistics, The University of British Columbia
[b]Department of Computer and Electrical Engineering, University of Southern California
[c]Department of Statistics, The University of California, Berkeley

Version: August 29, 2024

Abstract

Background: The timing and magnitude of COVID-19 infections are of interest to the public and to public health, but these are challenging to discern due to the volume of undetected asymptomatic cases and reporting delays. Retrospective estimation of COVID-19 infections improves our understanding of the size and scope of the pandemic and provides more meaningful quantification of disease patterns and burden.
Methodology: We retrospectively estimate daily incident infections for each U.S. state prior to Omicron. Reported COVID-19 cases are deconvolved to their date of infection onset using delay distributions estimated from the CDC line list. Then, we develop a serology-driven model to scale the deconvolved cases and account for unreported infections. The resulting daily infection estimates incorporate variant-specific incubation periods, time-varying reinfection estimates, and waning antigenic immunity.
Results: The disease burden from infections emerges earlier and more extensively than is reflected by cases. 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 30% of infections were eventually reported as cases during this period.
Conclusions: While reported cases offer a convenient proxy of disease burden, they fail to capture the full extent of infections. Our estimates of infections enhance the understanding of the progression of the pandemic in the U.S. prior to the onset of Omicron and its descendants.

Keywords: Infections; Case ascertainment ratio; COVID-19; SARS-CoV-2; Deconvolution; Time series; Seroprevalence; Antibody

1 Introduction

Reported COVID-19 cases are a staple in tracking the pandemic at varying geographic resolutions [1–3]. Yet, for every case that is eventually reported to public health, several infections are likely to have occurred, and likely much earlier. To see why, it is important to understand *whose* cases are being reported and what differentiates them from unreported cases as well as *when* these case reports happen. Figure 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 targets symptomatic individuals; thus, infected individuals exhibiting little to no symptoms are omitted [4]. In addition, testing practices, availability, and uptake vary temporally and spatially [5–7]. Finally, cases provide a belated view of the pandemic's progression, because they are 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 [8,9]. For these reasons, reported cases are lagging indicators of the course of the pandemic. Furthermore, they do not represent the actual number of new infections that occur 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.

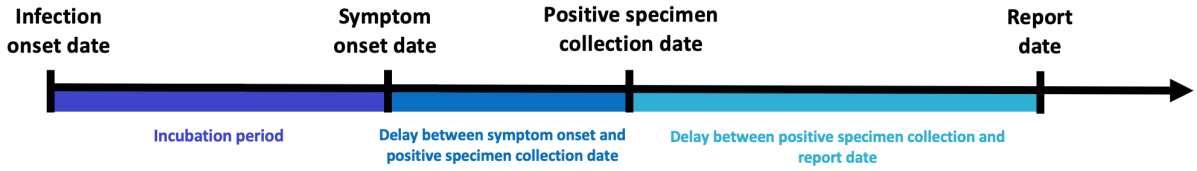

Figure 1: Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

Contextualizing the course of the pandemic, understanding the effects of interventions, and drawing insights for future pandemics is challenging because the spatial and temporal behaviour 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

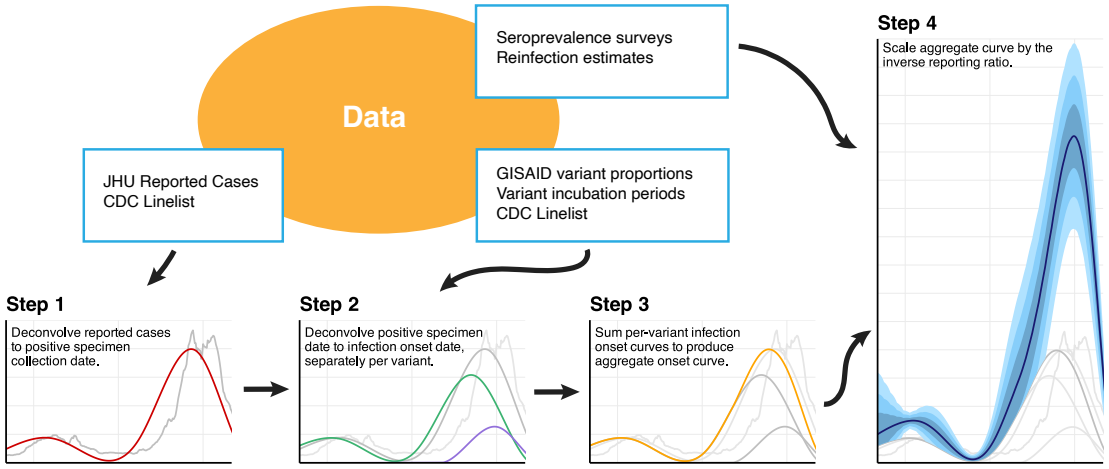

Figure 2: Flowchart of the data and major analysis steps required to get from reported cases to incident infection estimates. In Step 1, we use the CDC line list data to deconvolve reported cases (grey) backward to the date of positive specimen (red). Step 2 separately deconvolves these to the date of infection by variant (Epsilon in Purple, Ancestral in Green), before summing across all variants (orange) in Step 3. Finally, we use seroprevalence survey and time-varying reinfection data to account for the unreported infections.

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 from June 1, 2020 to November 29, 2021. 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 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. A graphical depiction of our procedure is shown in Figure 2. Our results examine features of our infection estimates and the implications of using them, rather than reported cases, to assess the impact of the pandemic. We also produce simple time-varying infection-hospitalization ratios (IHRs) for each state and compare these with case-hospitalization ratios (CHRs). 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 and will prove an important benchmark for others to undertake retrospective analyses. Our estimates, as well as the R and Python code used to produce them, are available on GitHub.

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. Figure 2 provides a visual summary of the major analysis tasks. First, we estimate spatiotemporal-specific delays from positive specimen to report date and use them to deconvolve reported cases to their sample collection date. Next, we estimate the delay from symptom onset to positive specimen, combine this with variant specific infection-to-symptom delays, and push back reported cases to the corresponding date of infection. These reported infection estimates are aggregated across the variant categories and adjusted by the case ascertainment ratio, estimated with seroprevalence survey data and a model for antigenic immunity. Additional methodological details are provided in the Supplement.

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 (Supplement A).

We will start 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 [10] was used to estimate the necessary delay distributions. Together, these are the ingredients for Step 1 in Figure 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) [1] and retrieved with the COVIDcast API [11]. 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

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

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

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

The solution to this minimization problem is an adaptive piecewise cubic polynomial [12,13] and can be accurately computed with ease [14,15]. We select the tuning parameter λ with cross-validation to minimize the out-of-sample reconvolution error.

To estimate the $\pi_{\ell,t}(k)$ for all states ℓ, times t, and delays k, we use the CDC line list [10]. 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 (Supplement B). Define $z_{\ell,t}$ to be a case report occurring at time t in location ℓ. We assume that positive specimens will be reported within 60 days and that no test will be 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)

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

where $\mathbf{1}(Z) = 1$ if Z is true and 0 otherwise. We also compute a similar national pmf, $\tilde{p}_t(k)\mathbf{1}(0 < k \leq 60)$, without restricting to location ℓ. Next, let $\alpha_{\ell,t}$ be the ratio of $N_{\ell,t}$ to the number of cases reported by JHU CSSE [1] in the window $[t-60+2, t+75]$. Then, compute $p_{\ell,t} = \alpha_{\ell,t}\tilde{p}_{\ell,t} + (1-\alpha_{\ell,t})\tilde{p}_t$. This construction was adopted to allow for more reliance on the state estimate when a larger fraction the JHU cases 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 $\widehat{\pi}_{\ell,t}(k)$ of the delay distribution $\pi_{\ell,t}(k)$. Additional details are deferred to Supplement C.

2.2 From positive specimen collection to infection onset

To continue, pushing positive specimen collection time back to infection onset (Step 2 in Figure 2), we use a procedure very similar to that described above and specified in Equations (1) and (2). However, because the delays involve the time from infection to symptom onset, these must be variant-specific. We use our estimates from Section 2.1, $\widehat{\mathbf{x}}_\ell$, 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 [16,17], and estimate a multinomial logistic regression model. This procedure is now standard [18–20] (see also Supplement D). 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 $\widehat{\tau}_{j\ell,t}(k)$. Details on the convolution and its inputs are in Supplements F to H.

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

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

and estimate $\mathbf{u}_{j\ell}$ by minimizing the negative loglikelihood with a penalty to encourage smoothness:

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

We call the solution $\widetilde{\mathbf{u}}_{j\ell}$ the *variant-specific deconvolved cases* and emphasize that these are cases that will eventually be 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 $\widehat{\mathbf{u}}_\ell = \sum_j \widetilde{\mathbf{u}}_{j\ell}$ (Step 3 in Figure 2). Note that 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 Figure 2). Seroprevalence of anti-nucleocapsid antibodies represents the percentage of people who have at least one resolving or past infection [21], 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 [22,23] (Supplement I).

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^1_m and r^2_m are modelled as Gaussian,

$$r^1_{\ell,m} \mid s_{\ell,m},\ w^1_{\ell,m} \sim \mathrm{N}(s_{\ell,m},\ w^1_{\ell,m}\sigma^2_{\ell,r}), \tag{6}$$

$$r^2_{\ell,m} \mid s_{\ell,m},\ w^2_{\ell,m} \sim \mathrm{N}(s_{\ell,m},\ w^2_{\ell,m}\sigma^2_{\ell,r}), \tag{7}$$

with source-specific measurement errors, $w^1_{\ell,m}$ and $w^2_{\ell,m}$, that scale proportional to reported uncertainty.

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

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

where $\widehat{u}_{\ell,t}$ 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 [24], and these estimates are broadly similar to those in other locations with available data [24–27]. 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 Equation (9) results in estimates that vary smoothly in time.

The antibody prevalence model specified by Equations (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. Additional details of this methodology and the computation of the associated uncertainty measurements are in Supplement J.

2.4 Lagged correlation to hospitalizations and time-varying IHRs

From the COVIDcast API[11], we 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 modelling choices, particularly the contribution of the main steps to the lagged correlation analysis, we conduct an ablation study that is detailed in Supplementary Methods Supplement K.

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

The lag that leads to the highest average correlation is used to estimate the time-varying IHRs for each state. The IHR is computed by dividing the number of individuals who are hospitalized due to COVID-19 by the estimated total number who were infected on the lagged number of days before. To stabilize these lagged IHR estimates, we average these hospitalizations and infections within a window of 31 days centered on the date of interest, rather than just using one pair of dates for each computation.

3 Results

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

Prior to Omicron, the largest infection (as opposed to case) outbreaks were observed in the late summer and early fall of 2021 in Louisiana, Georgia, Idaho, and Montana (Figures 3 to 4). During this time, the states that have the highest rate of infections on single day are Louisiana (476 infections per 100K, on July 20, 2021) and Idaho (also 457 infections per 100K, on September 7, 2021). The period of lowest viral transmission was in the summer of 2020, where Vermont had fewer than 10 infections per 100K per week from June to August, the longest such lull observed for any state. ATTN: Can you put one sentence that compares these to the case peaks around the same time, just for context?

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) present waves of infections that mirror each other in amplitude and timing.

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 tend to severely underestimate infections during Delta for many states, more so than in earlier waves (Figure 3). The most extreme was 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 appear in case reports. The cases-to-infections ratio was larger in earlier waves, and its effects were most apparent in different regions. During Alpha, Louisiana had 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 had the lowest ratio of reported cases to infections at 21.4% and 21.7%, respectively.

Figure 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.

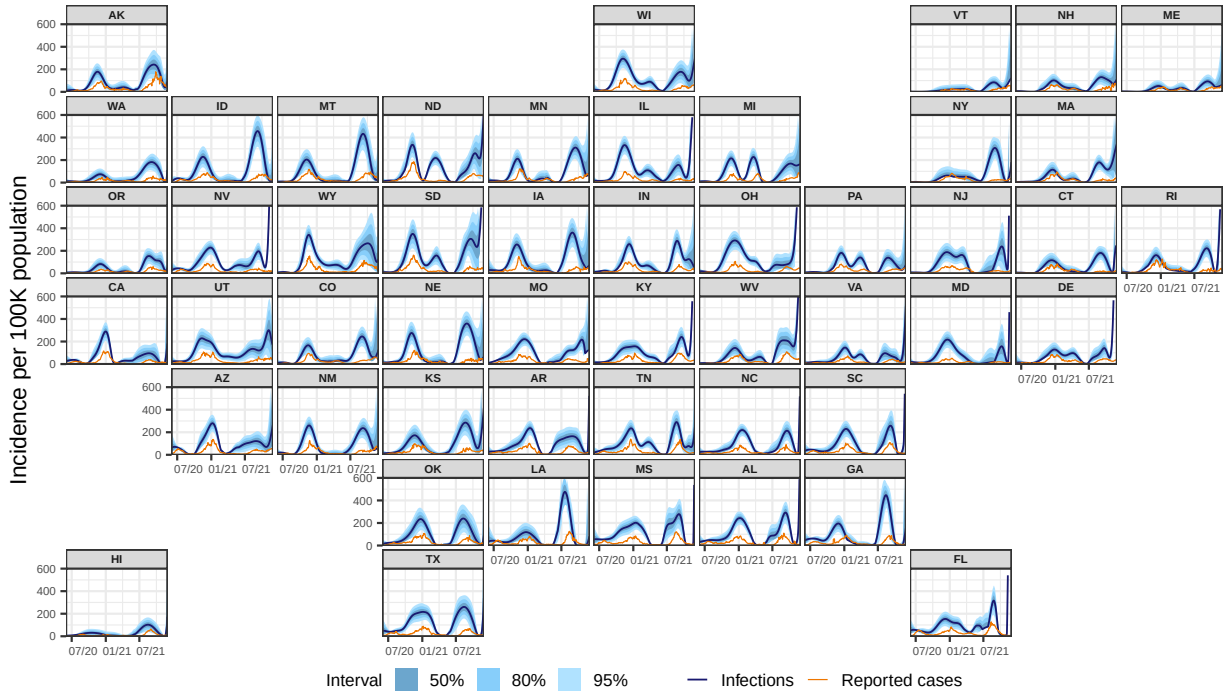

Figure 3: Estimates of the daily new infections per 100,000 population for each U.S. state from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% intervals for the estimates, while the orange line represents the trailing 7-day average of reported cases per 100,000.

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 (Figure 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 and Idaho lead cases by about 41 and 6 days at the peaks (Figure 3). Such trends are also observed during the Ancestral wave, where infections peak about 12 and 24 days earlier than cases for these same states. ATTN: These last two sentences read strangely: "41 and 6 days"? what does that mean? These temporal discrepancies underscore the importance of clearly distinguishing between infections and cases when assessing the disease burden and spread of COVID-19.

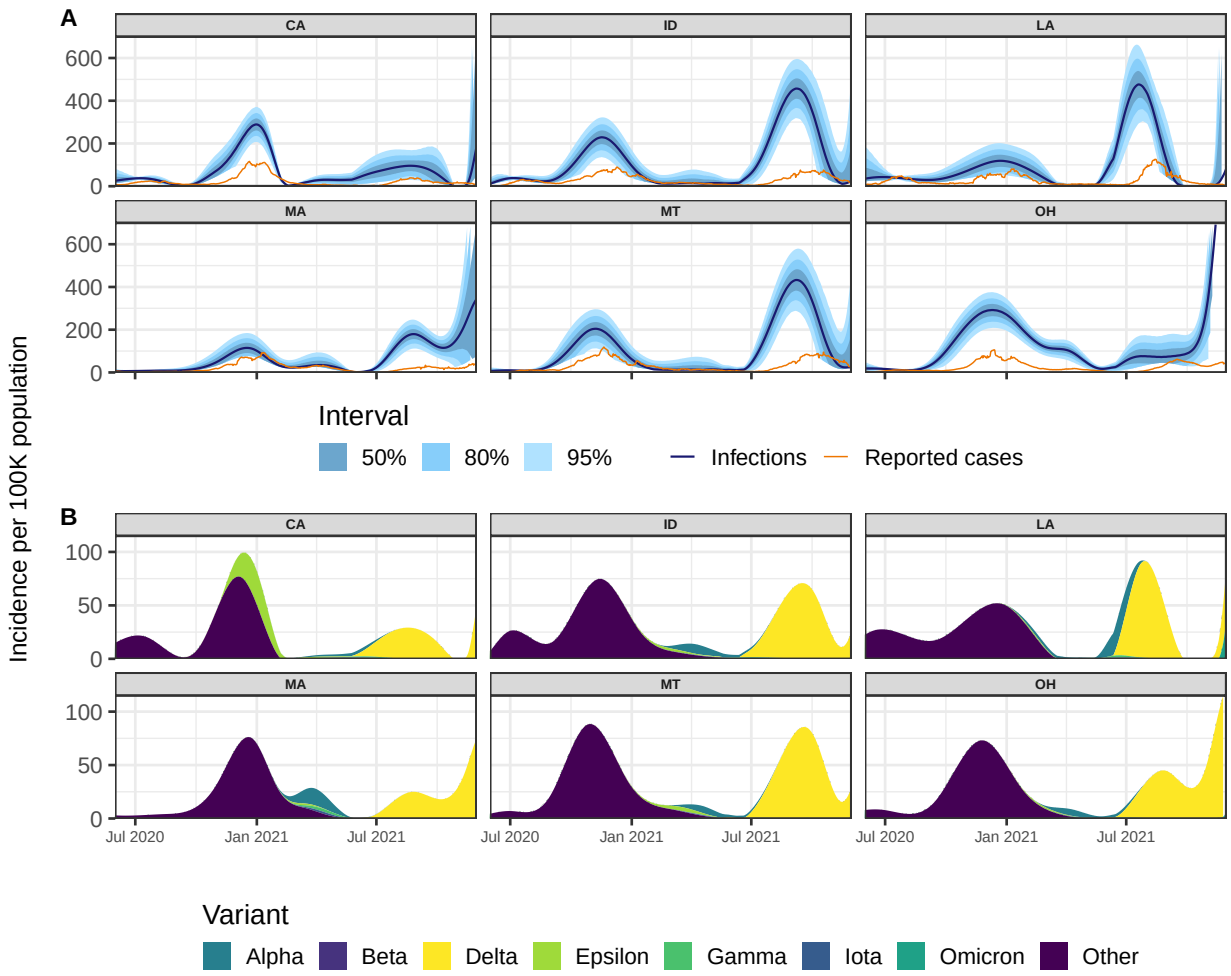

Figure 4: Panel A: Reported cases (orange) and estimates of daily new infections (dark blue) per 100K inhabitants. The blue shaded regions indicate 50, 80, and 95% confidence bands. Panel B: Deconvolved cases colored by variant per 100K inhabitants.

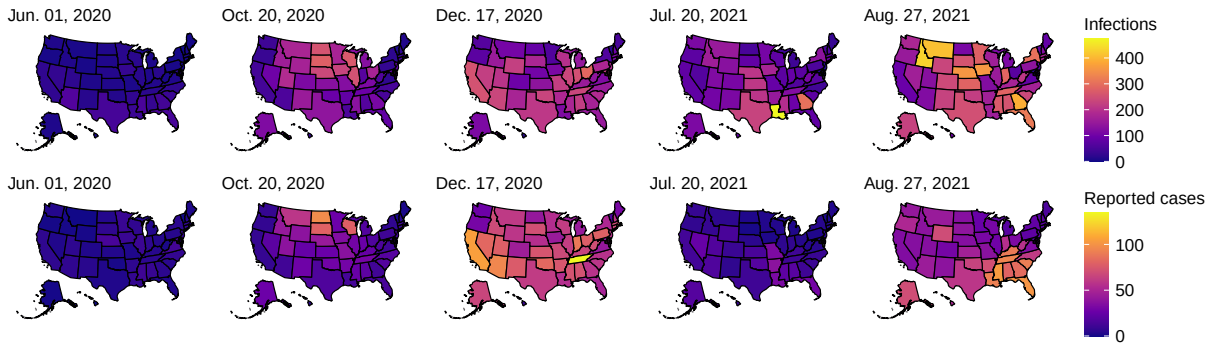

Figure 5: Choropleth maps of the state-level estimates of the daily new infections per 100K (top row) and the daily new cases per 100K (bottom row) for five select dates between June 1, 2020 and November 29, 2021. Note that the first date was chosen as a baseline, while the other dates were chosen because they present large counts of infections across all states. In particular, the third and fifth dates present the largest number of total infections across the 50 states within those calendar years.

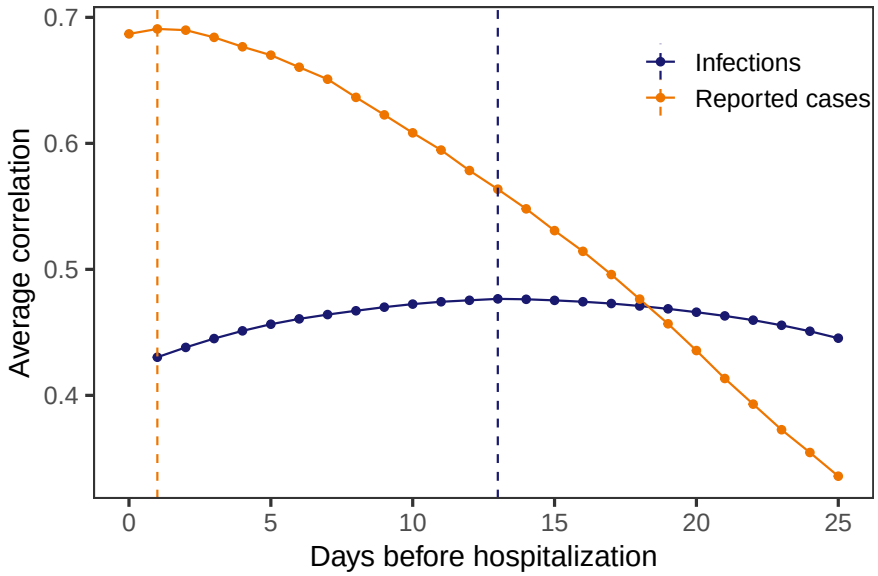

Figure 6: Spearman's rank correlation between each of cases and infections with hospitalizations per 100,000. These are calculated for each lag, state, and rolling window of 61 days before averaging. The vertical dashed lines indicate the lags for which the highest average correlation is attained.

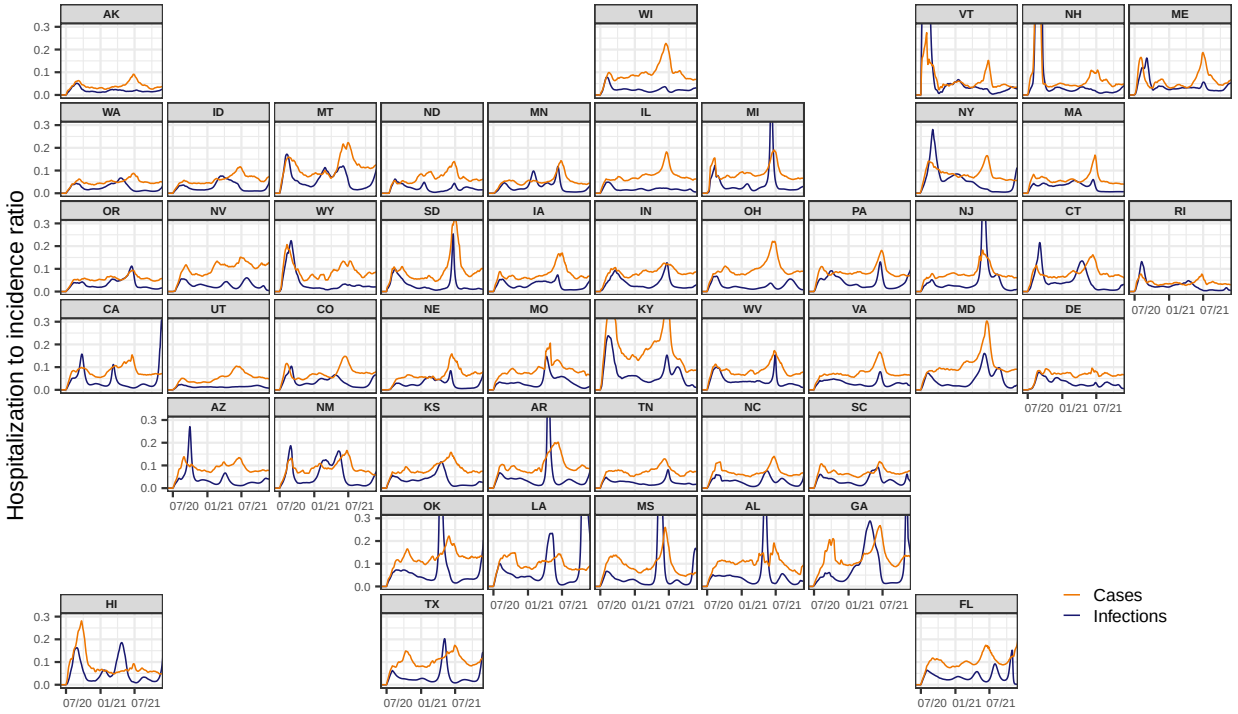

Figure 7: Time-varying IHR and CHR estimates for each state from June 1, 2020 to November 29, 2021, obtained using the respective correlation maximizing lag from Section 3.2. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious day of the week effects. Also note that the different starting points across states are due to the availability of the hospitalization data.

3.2 Insights from cross-correlations, IHRs and CHRs

ATTN: Need a connecting sentence here. Is this next sentence about infections? The maximum Spearman's correlation, averaged across states, is 0.48, and occurs at a lag of 13 days (Figure 6). 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.

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 Figure 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.

Both IHRs and CHRs exhibit similar spatiotemporal trends as those noted for infections. Namely, states that are proximate (for example, North and South Carolina) show similar temporal patterns in IHRs and CHRs. In addition, similar spikes are evident across many states during waves of infections that are driven by variants of concern. For example, many states exhibit a striking increase in hospitalizations in mid-2021, which coincides with the rapid takeover of the Delta variant [16].

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 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 influxes of data or more persistent processing issues that increase the average time from case detection to report [9,28]. Thus, while reported cases provide an indication of the trajectory of the pandemic, it is delayed and incomplete.

Our approach offers a number of 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 [29,30]. Unlike previous approaches that use a single delay distribution to generate estimates for all states [15,31,32], 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 [32], 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 [33–35].

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 (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. However, future work could refine this analysis. Because the waning of immunity is likely to be variant-dependent [36], 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.

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 [37–39]. 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 [24–27], 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 [40]. 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 [6,11].

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. The result removes the effects of the delay from reported cases, pushing them back to infection onset. This approach is complemented with the development of a model that leverages the measurements of waning detectable antibody levels and 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 over time provide a clearer and more comprehensive understanding of the course of the pandemic. Such estimates contribute important information on the timing and magnitude of the disease burden for each location, and they highlight trends that may not be visible from reported case data alone. Therefore, our infection estimates provide key information for the ongoing investigation on the true size and impact of the pandemic.

Code availability

Code for reproducing all figures and the numerical results is available at https://github.com/cmu-delphi/latent-infections/.

References

[1] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, 533–534 (2020).

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

[3] The Washington Post. 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 (2020).

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

[5] Pitzer, V. E. *et al.* The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. *American Journal of Epidemiology* **190**, 1908–1917 (2021).

[6] European Centre for Disease Prevention and Control. Strategies for the surveillance of COVID-19. Technical report, ECDC, Stockholm, Sweden (2020).

[7] Hitchings, M. D. *et al.* The usefulness of the test-positive proportion of severe acute respiratory syndrome coronavirus 2 as a surveillance tool. *American Journal of Epidemiology* **190**, 1396–1405 (2021).

[8] Pellis, L. *et al.* Challenges in control of COVID-19: Short doubling time and long delay to effect of interventions. *Philosophical Transactions of the Royal Society B* **376**, 20200264 (2021).

[9] Washington State Department of Health. COVID-19 data dashboard. https://doh.wa.gov/emergencies/covid-19/data-dashboard (2020).

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

[11] Reinhart, A. *et al.* An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences* **118**, e2111452118 (2021).

[12] Tibshirani, R. J. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42**, 285–323 (2014).

[13] Tibshirani, R. J. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. *Foundations and Trends in Machine Learning* **15**, 694–846 (2022).

[14] Ramdas, A. & Tibshirani, R. J. Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics* **25**, 839–858 (2016).

[15] Jahja, M., Chin, A. & Tibshirani, R. J. Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. *Statistical Science* **37**, 207–228 (2022).

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

[17] Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).

[18] Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).

[19] Annavajhala, M. K. *et al.* Emergence and expansion of SARS-CoV-2 B. 1.526 after identification in New York. *Nature* **597**, 703–708 (2021).

[20] Figgins, M. D. & Bedford, T. SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. *MedRxiv* 2021–12 (2021).

[21] Centers for Disease Control and Prevention. COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/#national-lab (2020).

[22] Centers for Disease Control and Prevention. 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 (2021).

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

[24] Ruff, J. *et al.* Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. *Emerging Infectious Diseases* **28**, 1977 (2022).

[25] New York State COVID-19 reinfection data. https://coronavirus.health.ny.gov/covid-19-reinfection-data (2021).

[26] Hawaii Department of Health COVID-19 reinfection data. https://health.hawaii.gov/coronavirusdisease2019/files/2022/09/reinfection_report_2022-09-28.pdf (2022).

[27] Reported COVID-19 reinfections in Washington State. https://doh.wa.gov/sites/default/files/2022-02/421-024-ReportedReinfections.pdf (2022).

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

[29] Unwin, H. J. T. *et al.* State-level tracking of COVID-19 in the United States. *Nature Communications* **11**, 6189 (2020).

[30] Center for the Ecology of Infection Diseases. COVID-19 portal. https://www.covid19.uga.edu/nowcast.html (2020).

[31] Chitwood, M. H. *et al.* Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: Results of a Bayesian evidence synthesis model. *PLOS Computational Biology* **18**, e1010465 (2022).

[32] Miller, A. C. *et al.* Statistical deconvolution for inference of infection time series. *Epidemiology* **33**, 470–479 (2022).

[33] Tanaka, H. *et al.* Shorter incubation period among COVID-19 cases with the BA. 1 Omicron variant. *International Journal of Environmental Research and Public Health* **19**, 6330 (2022).

[34] Ogata, T., Tanaka, H., Irie, F., Hirayama, A. & Takahashi, Y. Shorter incubation period among unvaccinated delta variant coronavirus disease 2019 patients in Japan. *International Journal of Environmental Research and Public Health* **19**, 1127 (2022).

[35] Wu, Y. *et al.* Incubation period of COVID-19 caused by unique SARS-CoV-2 strains: a systematic review and meta-analysis. *JAMA network open* **5**, e2228008–e2228008 (2022).

[36] Pooley, N. *et al.* Durability of vaccine-induced and natural immunity against COVID-19: A narrative review. *Infectious Diseases and Therapy* **12**, 367–387 (2023).

[37] Wei, J. *et al.* Risk of SARS-CoV-2 reinfection during multiple Omicron variant waves in the UK general population. *Nature Communications* **15**, 1008 (2024).

[38] Pulliam, J. R. *et al.* Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science* **376**, eabn4947 (2022).

[39] Eythorsson, E., Runolfsdottir, H. L., Ingvarsson, R. F., Sigurdsson, M. I. & Palsson, R. Rate of SARS-CoV-2 reinfection during an Omicron wave in Iceland. *JAMA Network Open* **5**, e2225320–e2225320 (2022).

[40] McManus, O. *et al.* Predicting COVID-19 incidence using wastewater surveillance data, Denmark, October 2021–June 2022. *Emerging Infectious Diseases* **29**, 1589 (2023).

[41] U.S. Census Bureau, Population Division. Annual estimates of the resident population for the United States, regions, states, District of Columbia, and Puerto Rico: April 1, 2020 to july 1, 2022. https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html (2022).

[42] World Health Organization. Tracking SARS-CoV-2 variants. https://www.who.int/activities/tracking-SARS-CoV-2-variants (2021).

[43] Yang, S. *et al.* Investigation of SARS-CoV-2 Epsilon variant and hospitalization status by genomic surveillance in a single large health system during the 2020-2021 winter surge in Southern California. *American Journal of Clinical Pathology* **157**, 649–652 (2022).

[44] Duerr, R. *et al.* Dominance of Alpha and Iota variants in SARS-CoV-2 vaccine breakthrough infections in New York City. *The Journal of Clinical Investigation* **131**, e152702 (2021).

[45] Tindale, L. C. *et al.* Evidence for transmission of COVID-19 prior to symptom onset. *eLife* **9**, e57149 (2020).

[46] Grant, R. *et al.* Impact of SARS-CoV-2 Delta variant on incubation, transmission settings and vaccine effectiveness: Results from a nationwide case-control study in France. *The Lancet Regional Health–Europe* **13**, 100278 (2022).

[47] Public Health Agency of Canada. COVID-19 for health professionals: Transmission. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/health-professionals/transmission.html (2021).

[48] Zaki, N. & Mohamed, E. A. The estimations of the COVID-19 incubation period: A scoping reviews of the literature. *Journal of Infection and Public Health* **14**, 638–646 (2021).

[49] Cortés Martínez, J. *et al.* SARS-CoV-2 incubation period according to vaccination status during the fifth COVID-19 wave in a tertiary-care center in Spain: A cohort study. *BMC Infectious Diseases* **22**, 1–7 (2022).

[50] Jones, J. M. *et al.* Estimated US infection-and vaccine-induced SARS-CoV-2 seroprevalence based on blood donations, July 2020-May 2021. *JAMA* **326**, 1400–1409 (2021).

[51] Bajema, K. L. *et al.* Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. *JAMA Internal Medicine* **181**, 450–460 (2021).

[52] Durbin, J. & Koopman, S. J. *Time Series Analysis by State Space Methods*, vol. 38 (OUP Oxford, 2012).

Acknowledgements

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

Author contributions

RT and DJM conceptualized the project. RL acquired the data, performed the statistical analysis, and wrote the initial draft of the paper. DJM, RT, and AS, reviewed and revised the paper. All authors were involved in significant discussion and development of the methodology.

Competing interests

The authors declare no competing interests.

Supplement for:

Retrospective estimation of latent COVID-19 infections before Omicron in the U.S.

This Supplement contains additional information about the data used as well as details about statistical methodology.

A A general description and depiction of convolution

In general, the goal of convolution is to propagate the input signal forward in time using a probability distribution. In the 1D and discrete context, it is simply a rolling, weighted average of the past. So for an input sequence $\{x_t\}_{t=1}^n$ and time-constant weights $\{z(k)\}_{k=-\infty}^0$, the output sequence $\{y_t\}_{t=1}^n$ is given by

$$y_t = \sum_{s=0}^{t} z(s) x_{t-s}. \tag{10}$$

Figure 8 presents a depiction of the convolution procedure for an example signal x_t (smoothed cases, orange line). Essentially, to push the cases forward in time, we take the appropriately aligned (forward-in-time) delay distribution $z(s)$ (blue shaded region) and convolve it with the smoothed cases signal counts by it to get the convolved estimates (blue line). This process is repeated as we march forward in time, as shown through the stop-motion panels, such that it eventually covers the entire line of cases. An important takeaway from this is that convolution is not the same as a simple shift of the data. A shift is a special case: when $z(d) = 1$ and $z(s) = 0$, $s \neq d$, we shift x forward by d. Rather, convolution generally weights the entire past by non-zero probabilities. Deconvolution proceeds in the same fashion, but in the opposite direction, going backward in time and undoing the effect of a convolution.

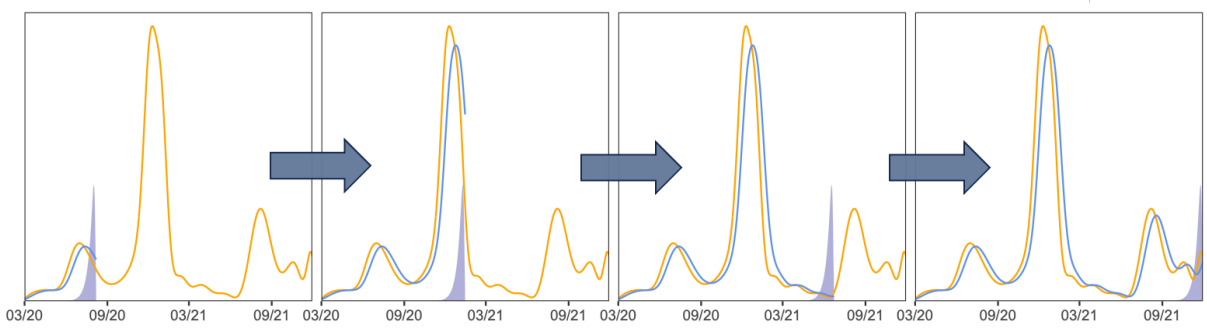

Figure 8: A general depiction of convolving smoothed cases (orange line) with the corresponding delay probabilities (shaded blue area) to get the convolved estimates (blue line) over four different times.

B Additional details on the date fields in the CDC line list

Because the restricted CDC line list is updated monthly and cases may undergo revision, we use a single version of it that was released on June 6, 2022. We consider this version to be finalized in that it is well-beyond our study end date such that the dataset is unlikely to be subject to further significant revisions.

Table 1 presents the percent of pairwise occurrences for the different possible permutations of events in the line list. Essentially, most cases follow the idealized ordering shown by Figure 1 and so we adhere to this construction as much as possible. Unfortunately, the line list has significant missing data, notably with respect to our variables of interest. Approximately 62.3% of cases are missing the symptom onset date, 55.4%

Order of events	Percent pairwise occurrence	Handling
IO → SO → PS → RE	PS ≥ SO: 97.1 PS = SO: 33.6 PS > RE: 1.74 PS = RE: 14.6	This is the idealized order of events and so the support sets for SO → PS and PS → RE delay distribution constructions around this such that IO comes first by construction, SO typically precedes PS, but may be the same or come before, and RE comes after PS and SO
IO → PS → SO → RE	PS < SO: 2.91 SO ≤ RE: 99.3 SO < RE: 86.1	Allowed for negative delays up to the largest non-outlier value for the 0.05 quantile of delay from PS to SO by state
IO → PS → RE → SO	RE < SO: 0.7 RE < PS: 1.7	Current handling by the CDC of the line list ensures that the most concerning cases are handled where SO = PO = RE, SO = RE and PO = RE

Table 1: Percent pairwise occurrence for the different permutations of events considered in the restricted CDC line list. The abbreviation IO stands for infection onset, SO is symptom onset, PS is positive specimen, and RE is report date. We consider a restricted set of permutations because we assume that IO must come first and that PS must precede report date for a case to be legitimate. Finally, the underlying assumption for the percent pairwise occurrence calculations is that the cases must have both elements present (not missing).

are missing positive specimen date, and 8.96% of cases are missing the report date. Furthermore, cases with missing report or positive specimen dates may be filled with their symptom onset date[15]. So it is possible that all three variables may have the same date for a case. However, we only actually deal with pairs of these events; we do not use all three at once in our construction of the delay distributions. Therefore, we restrict our investigation of coincident missingness to the possible pairs.

Due to the contamination in the zero delay cases (those whose symptom onset was used to fill missing positive specimen or report date, the true extent of which is unknown to us), we omit all cases where the positive specimen and report dates have zero delay from our analysis. We choose to allow for zero and negative delay for symptom onset to report because correspondence with the CDC confirms the possibility that a person could test positive before symptom onset and it is a reasonable ordering to expect if, for example, the individual is aware that they have been exposed to an infected individual.

The restricted CDC line list contains 74,849,225 cases (rows) in total compared to 84,714,805 cases reported by the JHU CSSE; that is, the line list is missing about 10 million cases. The extent that this issue impacts each state is shown in Figure 9, from which it is clear the fraction of missing cases is substantial for many states, often surpassing 50%[15]. In addition, the probability of being missing does not appear to be the same for states, so there is likely bias introduced from using the complete case line list data. We consider such bias to be unavoidable in our analysis due to a lack of alternative line list sources.

C Additional details on delay distribution calculations

In the line list, we observe unusual spikes in reporting in 2020 in comparison to majority of 2021. When stratified by report date, a few states contribute unusually large case counts on isolated dates very late in the reporting process (often more than 100 days following specimen collection). These large accumulations of cases over time are likely due breakdowns of the reporting pipeline. Such anomalies are not likely to be reliable indicators of the delay from positive specimen to case report. Therefore, we prune these reporting backlogs systematically. The heuristic is to find large batches of cases that were all simultaneously reported on the same date with a lengthy delay (as would happen if a state "found" a tranche of previously unreported positive tests). For each of three time intervals (delimited by July 16, 2020, October 16, 2020, and January 15, 2021), we apply the following pruning procedure: Let $x_{\ell t}(d)$ be the number of cases in state ℓ with positive specimen collection date t and delay d. For each $j = 1, 2, \ldots$, define $z_{\ell bj} = \sum_{d=50j}^{50(j+1)} \sum_t x_{\ell t}(d)$. Then, locate

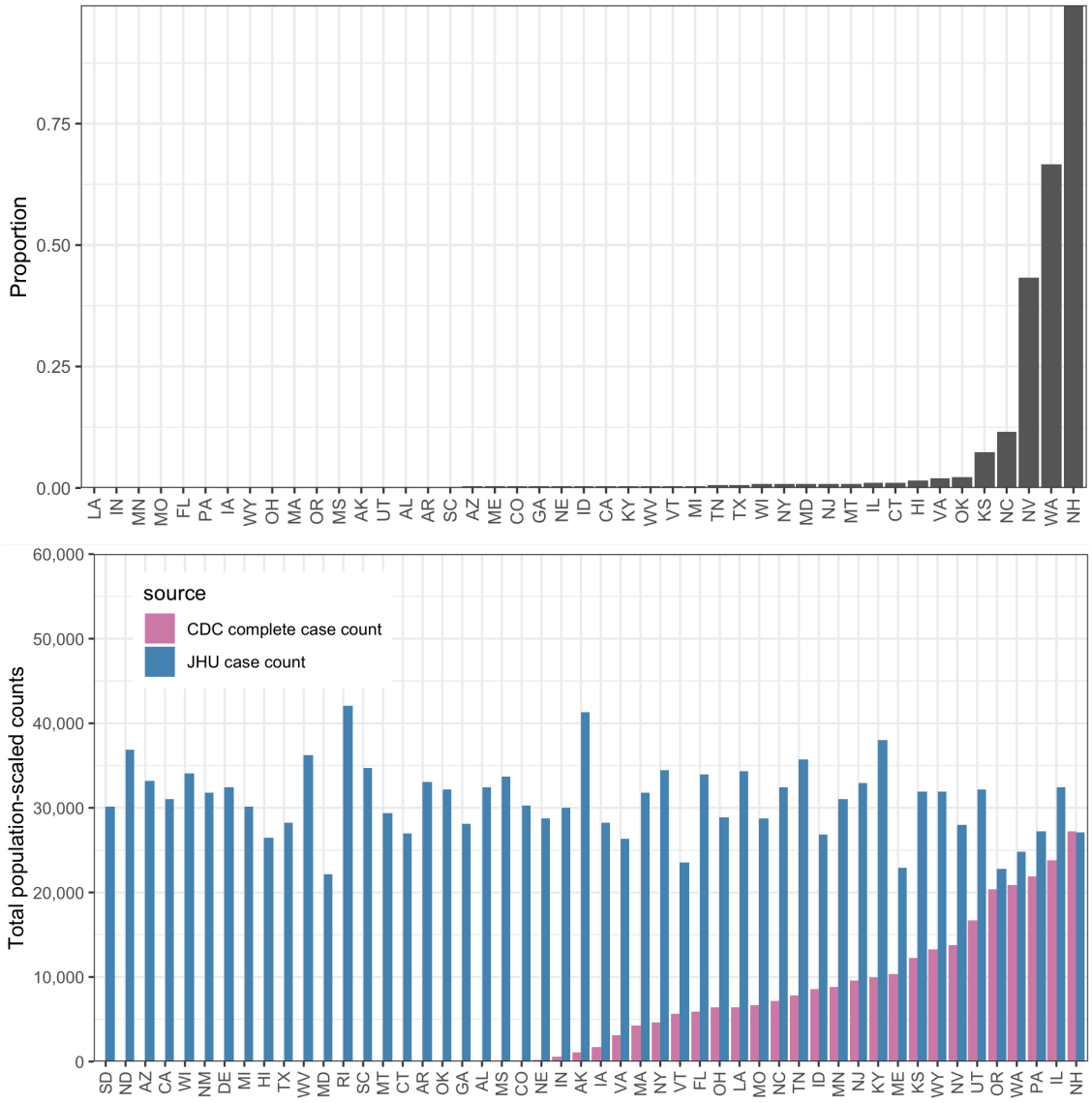

Figure 9: Top panel: Proportion of complete cases with zero delay between positive specimen and report date in the restricted CDC line list dataset. Bottom panel: Complete case counts by state in the CDC line list versus the cumulative complete case counts from JHU CSSE as of June 6, 2022. All counts have been scaled by the 2022 state populations as of July 1, 2022 from [41].

the collection of potentially problematic combinations

$$\mathcal{H} = \{(\ell, b, j) : \log(z_{\ell bj}) > \text{median}_\ell(\log(z_{\ell bj})) + 1.5 \times \text{IQR}_\ell(\log(z_{\ell bj}))\}.$$

Remove any case reports from the line list in $\mathcal{H}$ where the total number of cases with the same report date exceeds 2000 if $j = 1$ or 500 otherwise.

Finally, in New Hampshire and for a small handful of report dates, all cases reported by JHU appear in the CDC line list and all are recorded as having positive specimen collection date equal to the report date. The resulting estimate of the delay distribution (see Section 2.1) $\widetilde{p}_{\text{NH},t}(k)$ would be a point mass at $k = 0$ and the weight $\alpha_{\text{NH},t} = 1$ resulting $\widehat{\pi}_{\text{NH},t}(k)$ also being a point mass at $k = 0$. In this specific case, we force $\alpha_{\text{NH},t} = \min_{\ell \neq \text{NH}} \alpha_{\ell,t}$.

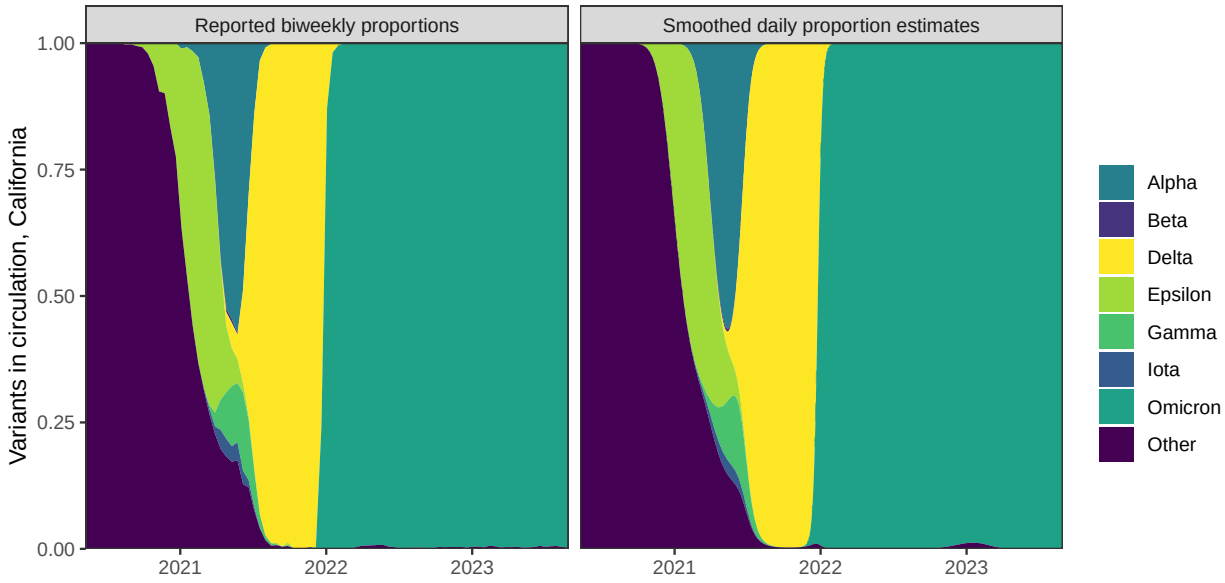

Figure 10: Left: Original biweekly proportions of the variants in circulation for California. Right: Daily proportions of the variants in circulation for California.

D Variant circulation proportions

To estimate the daily proportions of the variants circulating in each state, we obtain the GISAID genomic sequencing data from CoVariants.org[16,17]. These counts represent the total number of cases belonging to a particular variant using a sample of positive tests over a biweekly period. To estimate the population proportion of each variant, we apply multinomial logistic regression for the eight variant categories separately for each state. Multinomial logistic regression is a standard technique to model the frequency of SARS-CoV-2 variants[18–20].

We let $V_{j\ell,t}$ to be the probability of a new cases at time t in location ℓ corresponding to variant j. Let $v_{j\ell,t}$ be the analogous observed proportion. Then nonparametric multinomial logistic regression models the log odds as the system

$$\log\left(\frac{V_{j\ell,t}}{1-V_{j\ell,t}}\right) = \beta_{j\ell,0} + \beta_{j\ell,1}t + \beta_{j\ell,2}t^2 + \beta_{j\ell,3}t^3, \quad j = 1, \dots J. \tag{11}$$

This is estimated along with a constraint to ensure that the estimated proportions will sum to 1 across all J variants. The specification of the log odds as a third-order polynomial in time produces smoothness of the estimated proportions. Figure 10 shows the proportions by variant for California before (left) and after (right) the smoothing procedure.

E Constructing the delay from infection to test

The result of Step 1 (Section 2.1) is $\widehat{x}_{\ell,t}$, case estimates by positive specimen date for each state. To continue, pushing this back to infection estimates, we need the variant-specific delays from infection to positive specimen collection. As shown in Figure 1, this delay can be broken into two separate pieces: (1) the delay from infection to symptom onset, and (2) the delay from symptom onset to positive specimen collection. The first requires different methods and is specific to the variant causing the infection, while the second is estimable from the CDC line list.

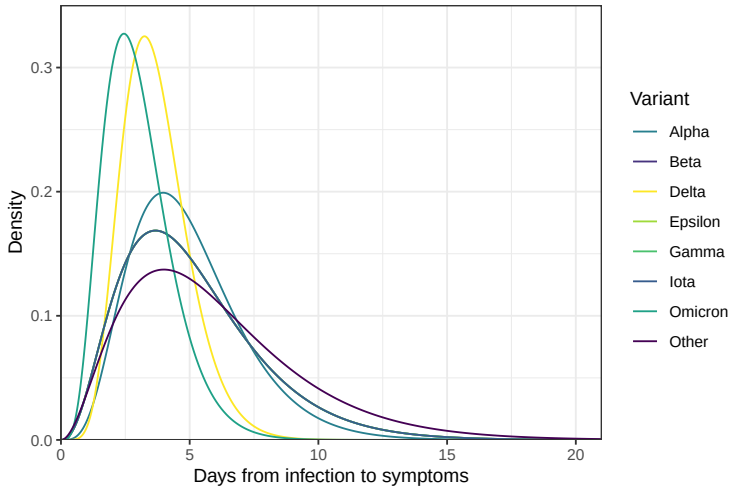

Figure 11: Gamma density for the incubation period of each of the eight variant categories. Note that the Ancestral variant uses reported shape and scale parameters[45], while the remaining variants convert reported estimates for the mean and variance[33,34,46] using the method of moments to produce the gamma parameters.

F Estimating the incubation period distributions

To account for the incubation period, the time between infection and symptom onset, we use estimates from the existing literature, modified slightly for coherence with each other: we model each incubation as a gamma distribution with different parameters. We focus on the following eight variants (shown in Figure 10), which saw significant circulation in one of the U.S. states during our study period: Ancestral/Other, Alpha, Beta, Epsilon, Iota, Gamma, Delta, and Omicron. Alpha, Beta, Delta, Gamma, and Omicron are all variants of concern[42], while we include the Epsilon (California) and Iota (New York) variants because of large impact on those and neighbouring states[43,44].

The incubation period of the Ancestral variant has been modelled as a gamma distribution[45], so we simply use the reported shape and scale parameters. For the Alpha, Beta, Gamma, Delta and Omicron variants, the mean and standard deviation are reported[33,34,46]. Therefore, we use method of moments to match the mean and variance to estimate the gamma parameters, using the moment equations given in Section 2.1. Then, we discretize each resulting density shown in Figure 11 to the support set, which is taken to be from 1 and 21 days. This range assumes that symptoms require at least 1 day to develop[47] and that an asymptomatic infection will resolve within 21 days[48,49].

We were unable to locate incubation period estimates for the geo-specific Epsilon and Iota variants, so we use the incubation period for Beta because Epsilon, Iota, and Beta are all children from the same parent in the phylogenetic tree of the Nextstrain Clades[16]. All other circulating variants are grouped together with the Ancestral variant. There was little available sequencing data prior to Alpha-emergence, but unfortunately, later in the pandemic, it is impossible to separate Ancestral from other rare variants, though these also saw minimal circulation after the middle of 2021.

G Estimating the delay distributions for symptom onset to positive specimen

Estimating the delay from symptom onset to positive specimen date follows a similar procedure as described in Section 2.1 with a minor adjustment. Here, we allow k to range from -3 to 21 (rather than 1 to 60). These upper and lower bounds are based on the largest delay values for the state-wide 0.05 and 0.95 quantiles. The median delay is very short at approximately 2 days, and an asymptomatic individual may test positive following a known exposure, before the onset of symptoms. We show both types of delays for a sample of states over several dates in Figure 12. Unlike the delay from positive specimen collection to report, the delay

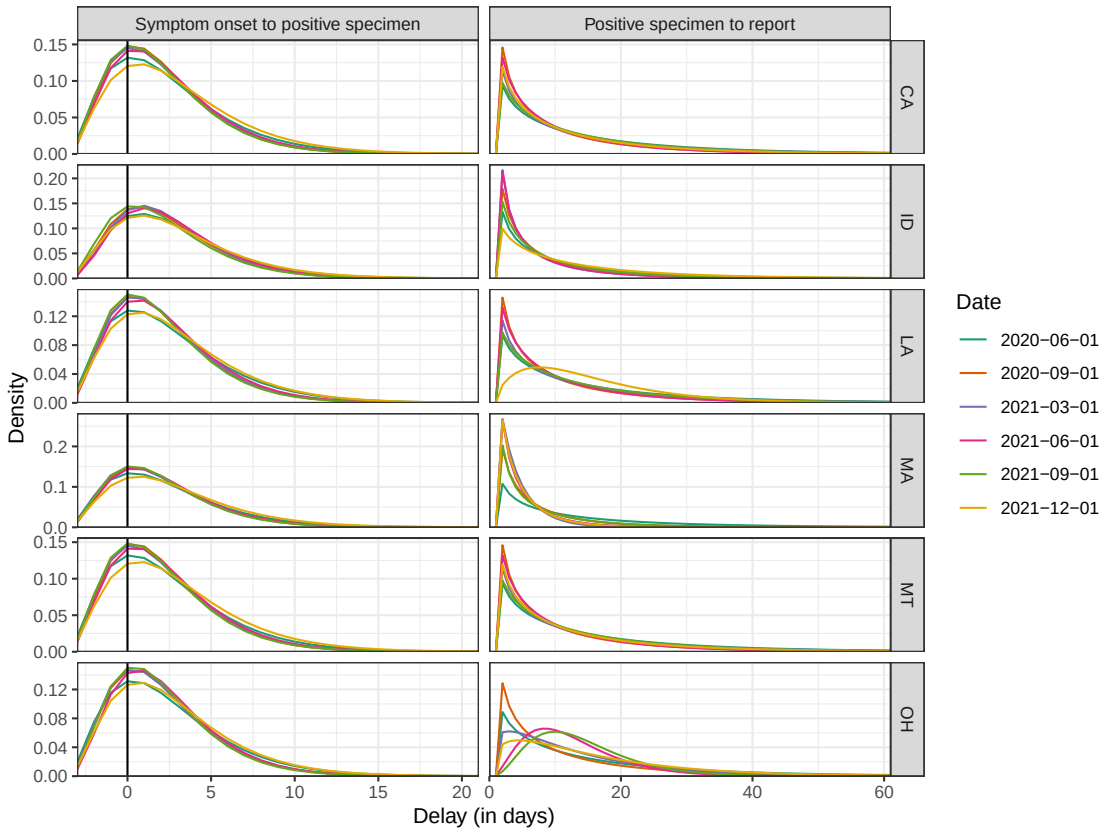

Figure 12: Depictions of the estimated delay from symptom onset to positive specimen date (left) and from positive specimen date to report date (right) for a sample of six states over several dates.

from symptoms to positive specimen can conceivably be negative. The most obvious reason for this would be if a person knew they had been exposed to an infectious individual and so got tested prior to the development of symptoms. Required regular testing for jobs in health care settings, construction, or the film industry could also produce negative delays.

H Details on constructing the infection-to-test distributions

Finally, to produce the delay from infection to positive specimen collection we convolve the variant-specific incubation periods from Supplement F, denoted by $\{i_j(k) : 0 < k \leq 21\}$ with the location-time-specific symptom-to-positive-test distributions from Supplement G, denoted by $\{q_{\ell,t}(k) : -3 \leq k \leq 21\}$. The convolution of these yields a distribution $\{\tau_{j\ell,t}(k) : -3 \leq k \leq 42\}$. Figure 12 shows the delays used for a sample of 6 states: symptom to positive specimen (left column) and positive specimen to report (right column). The convolution distribution $\tau_{j\ell,t}$ requires convolving the distribution in the left column with the variant-specific incubation periods shown in Figure 11.

I Details about seroprevalence data

We use two major contemporaneous surveys to estimate the proportion of the population with evidence of previous infection in each state over time: the 2020–2021 Blood Donor Seroprevalence Survey and the Nationwide Commercial Lab Seroprevalence Survey [22,23]. In the former, the CDC collaborated with 17 blood collection organizations in the largest nationwide COVID-19 seroprevalence survey to date [22]. The blood donation samples were used to construct monthly seroprevalence estimates for nearly all states from July 2020

to December 2021[50]. In the latter survey, the CDC collaborated with two private commercial laboratories to test blood samples from people that were in for routine or clinical management (presumably unrelated to COVID-19[51]) for the antibodies to the virus. The resulting dataset contains seroprevalence estimates for a number of multi-week collection periods starting in July 2020 to February 2022.

Both datasets are based on repeated, cross-sectional studies that estimate the percentage of people who were previously infected with COVID-19 using the percentage of people from a convenience sample who had antibodies against the virus[21,50,51]. Adjustments were made in both for age and sex to account for the demographic differences between the sampled and the target populations. However, both datasets are incomplete and they differ in the number and the timing of the data points for each state (Figure 13). For example, in the commercial dataset, the last estimate for North Dakota is in September 2020. In the blood donor dataset, Arkansas does not have estimates available until October 2020.

A major difference in the structure of the two datasets is that the commercial dataset always has the seroprevalence estimates at the level of the state, while the blood donor dataset can either have estimates for the state or for multiple separate regions within the state. For the commercial dataset, we use the midpoint of the provided specimen collection date variable. For the blood donor dataset, we use the median donation date if the seroprevalence estimates are designated to be for entire state. If they are instead for regions in the state, since there is reliably one measurement per region per month, we aggregate the measurements into one per month per state by using a weighted average (to account for the given sample sizes of the regions). The median of the median dates is taken to be the date for the weighted average. If there are multiple measurements in a week from a seroprevalence source, then the average is used.

J State space representation of the antibody prevalence model

The antibody prevalence model described in Section 2.3 can be expressed as a linear Gaussian state space model[52]. For $m = 1, \ldots, M$, let α_m be a vector of latent state processes at time m and y_m be a vector of observations at time m. The form of the (general) linear Gaussian state space model is

$$y_m = Z\alpha_m + \sigma_r^2 \epsilon_m, \qquad \epsilon_m \sim N(0, H_m) \tag{12}$$

$$\alpha_{m+1} = T_m \alpha_m + R\eta_m, \quad \eta_m \sim N(0, Q) \tag{13}$$

where $\alpha_1 \sim N(a_1, P_1)$ and ϵ_m and η_m are mutually and serially independent.

To express the antibody prevalence model in state space form, we relate the model in Equations (6) to (9) to the components in Equations (12) and (13) as follows (omitting the location subscript for simplicity):

$$Z = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad H_m = \begin{bmatrix} w_m^1 & 0 \\ 0 & w_m^2 \end{bmatrix} \tag{14}$$

$$\alpha_m = \begin{bmatrix} s_m \\ a_m \\ a_{m-1} \\ a_{m-2} \end{bmatrix} \qquad T_m = \begin{bmatrix} (1-\gamma) & \widehat{u}_m^{\Sigma}(1-z_m) & 0 & 0 \\ 0 & 3 & -3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{15}$$

$$Q = \begin{bmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{bmatrix} \qquad a_1 = \begin{bmatrix} \tilde{s}_1 \\ \tilde{a}_1 \\ \tilde{a}_1 \\ \tilde{a}_1 \end{bmatrix} \qquad P_1 = \begin{bmatrix} \sigma_{\tilde{s}_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\tilde{a}_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\tilde{a}_1}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\tilde{a}_1}^2 \end{bmatrix} \tag{16}$$

where σ_r^2 is the variance of observations, σ_ϵ^2 is the variance of the seroprevalence estimates, σ_η^2 is the trend variance, and $\widehat{u}_m^{\Sigma}$ denotes the new deconvolved cases between m and $m+1$. Because the inverse reporting ratios should be more variable than the seroprevalence estimates, we enforce that the estimate of σ_η^2 is a multiple of σ_ϵ^2.

Finally, w_m^1 and w_m^2 are the time-varying inverse variance weights computed from the commercial and blood donor datasets, respectively. For each source, we compute the weights for the observed seroprevalence estimates using the formula for the standard error of a proportion. These weights are then re-scaled to sum to the number of observed seroprevalence measurements for the source. Finally, the ratio of the average observed

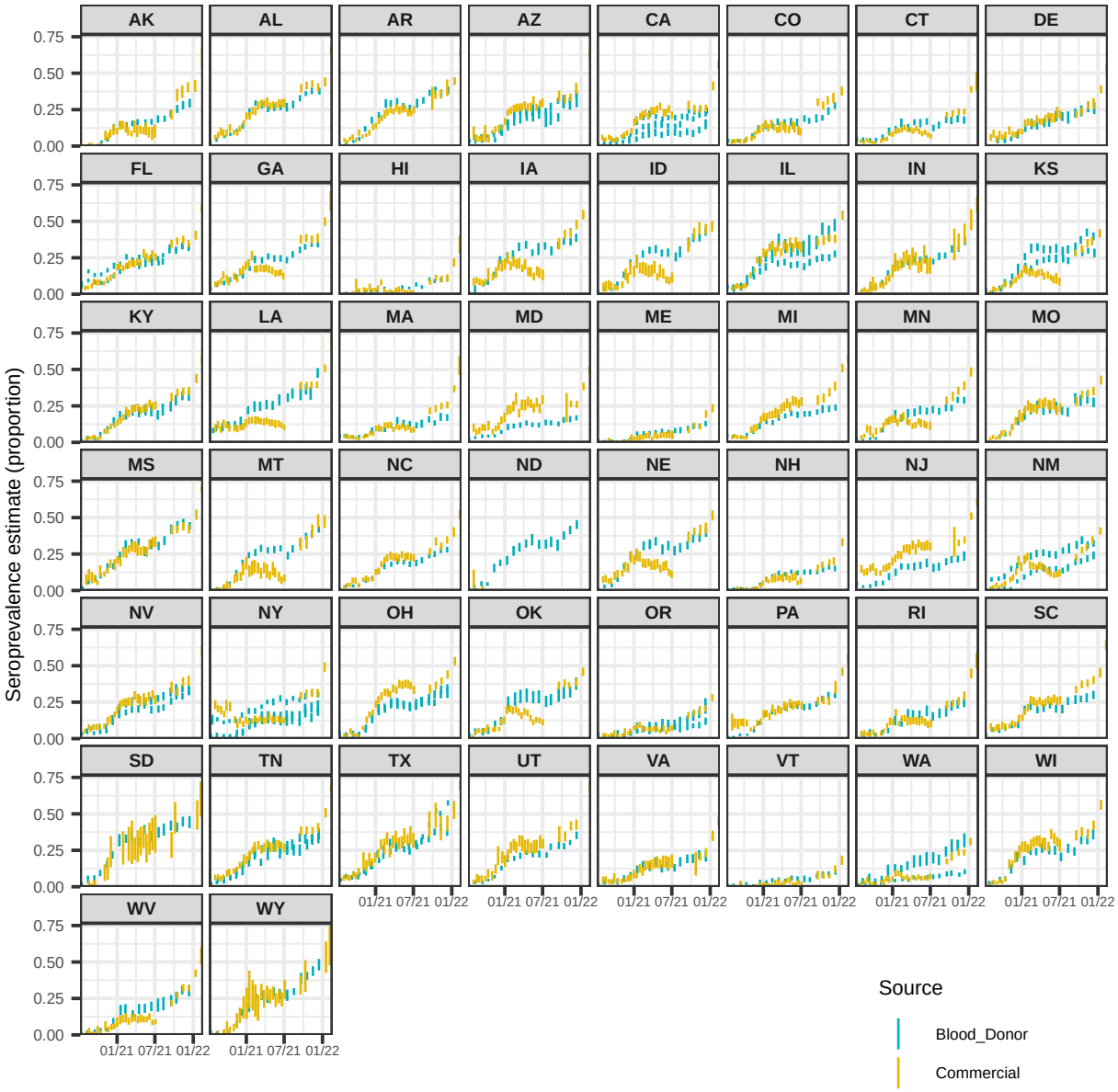

Figure 13: A comparison of the seroprevalence estimates from the Commercial Lab Seroprevalence Survey dataset (yellow) and the 2020–2021 Blood Donor Seroprevalence Survey dataset (blue). Note that the maximum and the minimum of the line ranges are the provided 95% confidence interval bounds to give a rough indication of uncertainty.

weights from the two sources is used to scale all weights relative the commercial source. This transformation is done purely for computational purposes to make estimation of the unknown parameters easier.

The prior distribution for α_1 is estimated using both data-driven constraints and externally sourced information. The initial value of the seroprevalence component, $\tilde{s}_1$ is the average of the initial seroprevalence measurements from each source rounded down to two decimal places. The corresponding initial variance estimate, $\sigma^2_{\tilde{s}_1}$, is taken to be the mean of the standard errors of the two seroprevalence estimates. For the initial value of the trend component, we use the inverse of the ascertainment ratio estimate for each state as of June 1, 2020 from Table 1 in [29].

The initial σ_r^2 is taken to be the average of the estimated variances from the observed seroprevalence measurements regressed linearly on time. The initial value of the multiplier is set to be 100 for all states. The σ_ϵ^2 and γ values are estimated separately for each state, then fixed to their averages on the log-scale.

Following maximum likelihood estimation of remaining parameters we use the Kalman filter to obtain the smoothed point estimates and variances of the weekly inverse reporting ratios. We use forward and backward extrapolation to extend these estimates outside of the observed seroprevalence range[52], followed by linear interpolation to produce daily values. We then multiply these by the corresponding deconvolved case estimates before converting to per-capita values. Annual estimates of the state populations as of July 1 of 2020 are taken from the Dec. 2022 press release from the U.S. Census Bureau[41].

K Ablation analysis of infection-hospitalization correlations

In this ablation study, the lagged correlation is re-computed by using the following infection estimates: 1. those from the deconvolution procedure under the assumption that the infection onset is the same as the positive specimen date (i.e., excluding the positive specimen to infection onset data and deconvolution); 2. those from the deconvolution procedure under the assumption that the infection onset is the same as the symptom onset date (excluding the incubation period data); 3. those from the deconvolution procedure when utilizing all incubation period and delay data (the deconvolved case estimates); 4. those from applying the antibody prevalence model to produce estimates for both the reported and the unreported cases (the infection estimates).

The results of this study are shown in Figure 14. From this, we can see that the deconvolved case and infection estimates from the intermediate steps are all leading indicators of hospitalizations. However, the degree that each such set of estimates lead hospitalizations depend on its location in the sequence of deconvolution steps and how close the estimates are to infection onset. For example, the deconvolved cases by positive specimen date tend to precede hospitalizations by about 11 days, while those for the subsequent step indicate that the deconvolved cases by symptom onset tend to precede hospitalizations by a longer time of 13 days. Finally, after adding the variant-specific incubation period data into the deconvolution and obtaining the deconvolved case estimates, we can observe that the reported infections precede hospitalizations by about 19 days.

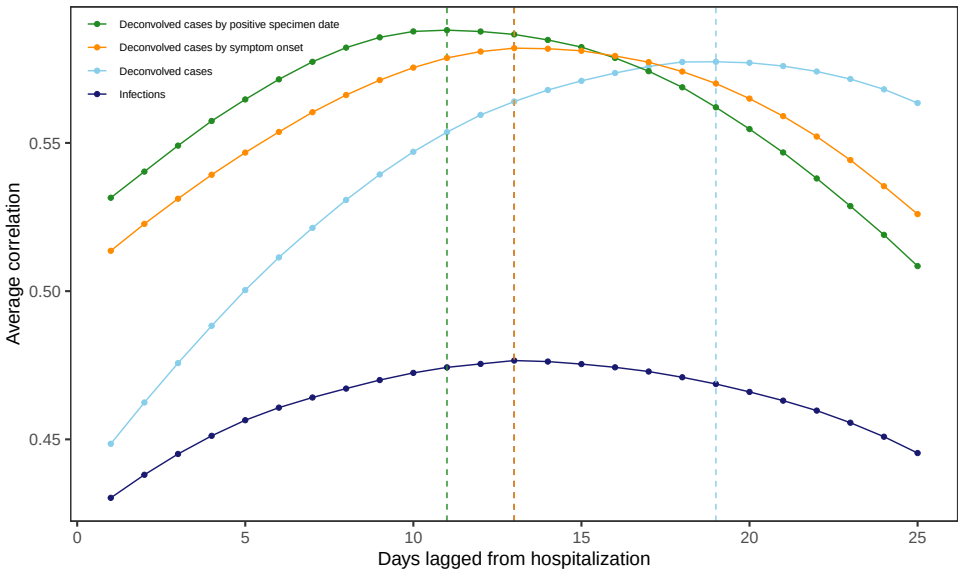

Figure 14: Lagged Spearman's correlation between the infection and hospitalization rates per 100,000 averaged for each lag across U.S. states and days over June 1, 2020 to November 29, 2021, and taken over a rolling window of 61 days. The infection rates are based on the counts for the deconvolved case and infection estimates as well as the reported infections by symptom onset and when the report is symptom onset. Note that each such set of infection counts is subject to a center-aligned 7-day averaging to remove spurious day of the week effects. The dashed lines indicate the lags for which the highest average correlation is attained.