

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

Rachel Lobay<sup>a,1</sup>, Ajitesh Srivastava<sup>b</sup>, Ryan J. Tibshirani<sup>c</sup>, and Daniel J. McDonald<sup>a</sup>

<sup>a</sup>Department of Statistics, The University of British Columbia

<sup>b</sup>Department of Computer and Electrical Engineering, University of Southern California

<sup>c</sup>Department of Statistics, The University of California, Berkeley

Version: June 18, 2024

## Abstract

The true timing and magnitude of COVID-19 infections are of interest to both the public and to public health, but these are challenging to pin down for a variety of reasons such as the underdetection of asymptomatic cases and delays in reporting. Accurate estimates of latent COVID-19 infections can improve our understanding of the size and scope of the pandemic and provide more meaningful and timely quantification of disease patterns and burden. In this work, we describe methodology to estimate daily incident *infections* for each U.S. state. To this end, we first deconvolve reported COVID-19 cases to their date of infection onset using delay distributions estimated from the CDC line list. We then use a serology-driven model to scale these deconvolved cases to account for the unreported infections. The resulting daily infection estimates incorporate variant-specific incubation periods, reinfections, and waning antigenic immunity. This analysis also produces estimates for other important quantities such as the number of deconvolved cases specific to each variant and the infection-case-report ratio (also referred to as the case ascertainment rate). Finally, we discuss some implications of our results: a disease burden that appears earlier and more extensively than previously quantified and differential infection-hospitalization ratio estimates. Our findings help to better understand the progression of the pandemic in the U.S. prior to the onset of Omicron and its descendants.

## 1 Introduction

Reported COVID-19 cases are a staple in tracking the pandemic at varying geographic resolutions<sup>1–3</sup>. 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 illustration of the path of a symptomatic infection that is eventually reported to public health. Using this figure, we can discern 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<sup>4</sup>. In addition, testing practices, availability, and uptake vary temporally and spatially<sup>5–7</sup>. 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<sup>8,9</sup>. For these reasons, reported cases are a lagging indicator of the course of the pandemic. Furthermore, they do not represent the actual number of new infections that occur on a 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.

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

---

<sup>1</sup>To whom correspondence should be addressed. E-mail: [rachel.lobay@stat.ubc.ca](mailto:rachel.lobay@stat.ubc.ca)

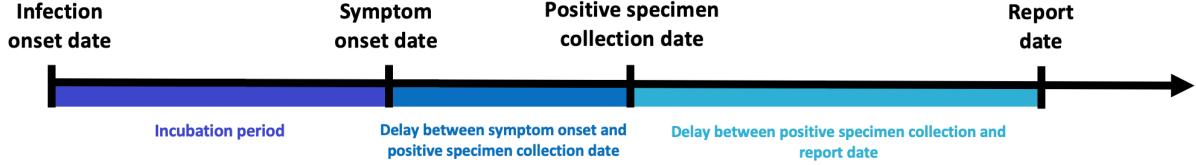


Figure 1: Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

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 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 from June 1, 2020 to November 29, 2021. Using state-level line list data, we construct time-varying delay distributions for the time 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 effect of the delay. Finally, we adjust for the unreported infections by using seroprevalence and reinfection data to account for the waning of antigenic immunity over time. A graphical depiction of our procedure is shown in Figure 2 with methodological details deferred to Section 4. Our results examine some features of our infection estimates and the implications of using them rather than reported cases in assessing the impact of the pandemic. To this end, we produce simple time-varying infection-hospitalization ratios (IHRs) for each state and compare those to similarly derived 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 publicly-available estimates 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](#).

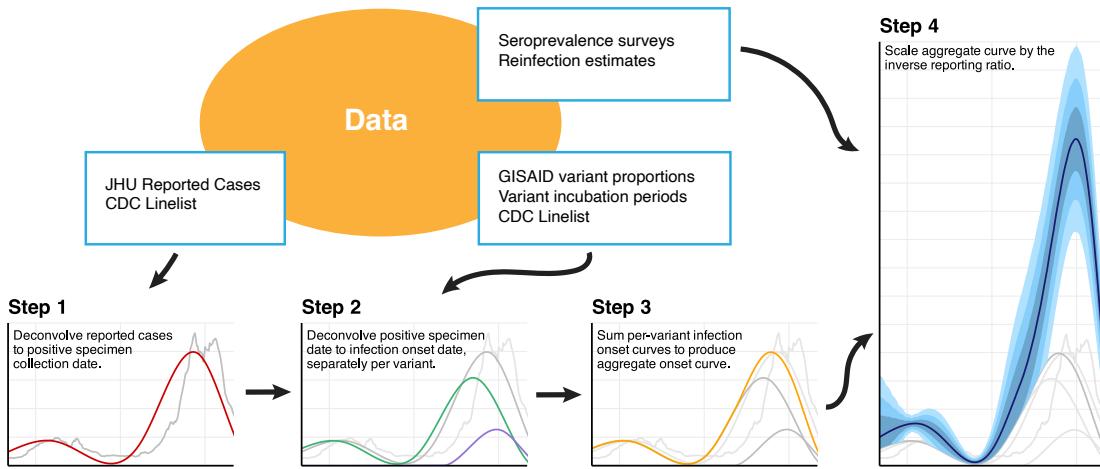


Figure 2: Flowchart of the data and major analysis steps required to get from reported cases to incident infection estimates. In Step 1, we use the CDC line list data to deconvolve reported cases (grey) backward to the date of positive specimen (red). Step 2 separately deconvolves these to the date of 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.

## 2 Results

By estimating the time series of COVID-19 infections per 100,000 inhabitants for each U.S. state from June 1, 2020 to November 29, 2021, we observe rates of infections that vary in intensity and disease burden across space and time (Figures 3 to 5). Outbreaks in infections precede those in cases and are consistently larger in magnitude (we will use “cases” to mean “reported cases”).

The largest per-capita outbreaks prior to Omicron were observed in the late summer or early fall of 2021 in Louisiana, Georgia, Idaho, and Montana matching the intuition that clusters of geographically proximate states are likely to exhibit similar viral spread. During this time, the two 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 is observed in the summer of 2020. From June 2020 to the end of August, Vermont saw less than 10 infections per 100K per week, the longest such lull for any state.

### 2.1 Infection estimates reveal waves missed by reported cases

Relative to reported cases, examining estimated infections reveals a rather different pattern. Figure 3 shows estimates of the number of daily new infections per 100,000 inhabitants for each U.S. state from June 1, 2020 to November 29, 2021 compared with reported cases.

Nearly all states exhibit at least two major waves of infections—the first begins in the fall of 2020 and extends into the winter season, while the second starts in the late summer of 2021 and proceeds into mid-fall. These represent major waves driven by the Ancestral and Delta variants, respectively. In general, greater similarities in the strength and magnitude of outbreaks emerge in small clusters of states that border each other. For instance, in the Western states of Idaho and Montana or in the Southern states of North and South Carolina the crests and troughs in the waves of infections appear in sync with each other. Compared to the other states, consistently low rates of infections are attained in the Northeastern states of Vermont, New Hampshire, and Maine, even during the aforementioned Ancestral and Delta waves.

While the major Ancestral, Alpha, and Delta waves tend to be visible for most states, there are clear outbreaks in unreported infections that are not easily detectable from cases alone in the falls of 2020 and 2021. For example, a wave of infections is evident in North Dakota and South Dakota over the spring of 2021 that is virtually undetectable from the reported cases. In the late summer of 2021, the Delta wave is only faintly detectable from cases in a number of Northeastern states (New York, Massachusetts, Connecticut, and New Hampshire), and yet the infections suggest that it has already begun in earnest.

### 2.2 The cases-to-infections ratio varies by state and variant

While it is clear from Figure 3 that cases underestimate the true burden of infections for every state, the degree to which this problem persists varies widely across states and variants. For the Ancestral wave, the largest discrepancies are more frequently in the Midwest: states such as Illinois, Indiana, and Ohio. For the Delta wave, some of the largest discrepancies between cases and infections are visible in the Western states of Idaho and Montana, the Southeastern states of Louisiana and Georgia, and the Midwestern states of Iowa and Nebraska. Early in the pandemic, such discrepancies between cases and infections may be attributable to state-specific issues with the reporting pipeline, while later, they more likely due to the rise in asymptomatic infections across variants<sup>10,11</sup>.

While the Delta wave is somewhat apparent from the case counts for all states (Figure 3), infection estimates suggest that case counts severely underestimate infections during this period for many states, more so than in earlier waves. The most extreme was New Jersey, where about 6.3% of estimated infections were eventually reported as cases. Similarly low are Maryland (7.3%), and Nevada (8.4%), and South Dakota (10.0%). This issue extends to most states: in 44 states fewer than 30% 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.

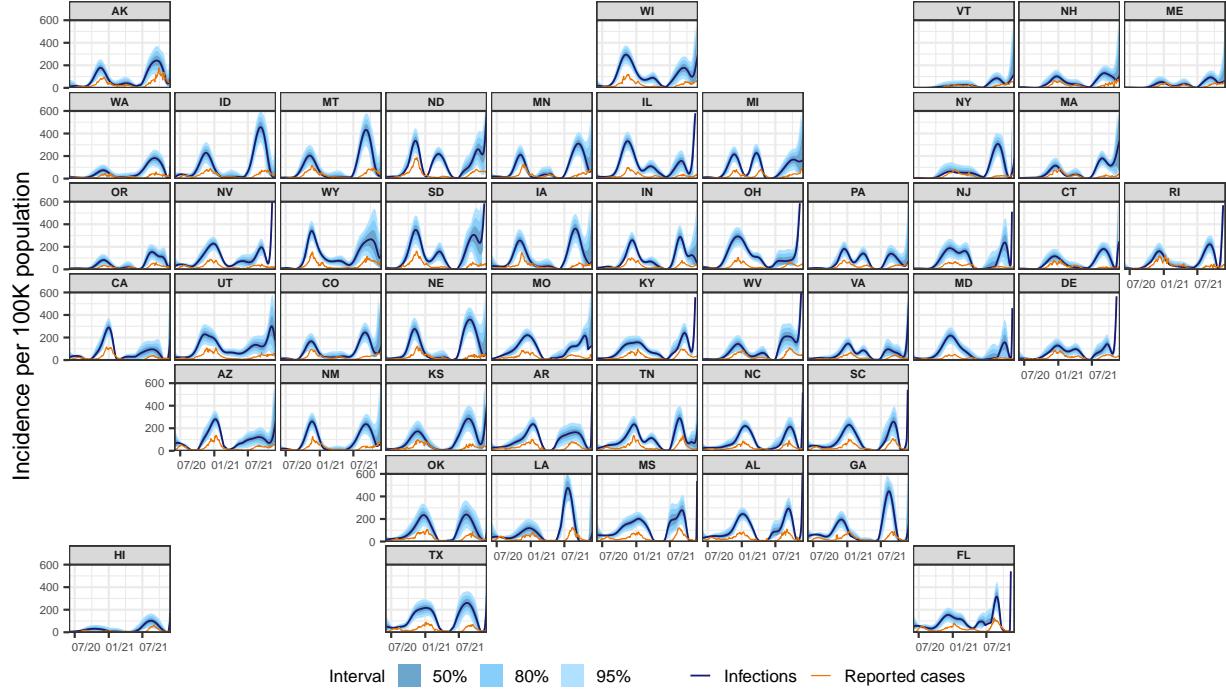


Figure 3: Estimates of the number of daily new infections per 100,000 population for each U.S. state from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% intervals for the estimates, while the orange line represents the trailing 7-day average of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

Figure 4 displays the state-level daily new infections and cases per 100,000 for five dates over June 2020 to November 2021, allowing a closer examination of their spatial patterns. For instance, it shows that on June 1, 2020, there is little difference between case and infection rates across the states, while later on, the differences become more pronounced. Furthermore, using cases as a proxy for infections can lead to misunderstandings in the states that are affected and the extent to which they are affected. For some days, the spatial extent of infections is understated by cases. For example, on October 20, 2020, while case rates are elevated in a handful of upper-Midwestern states (namely, North and South Dakota), infection estimates are elevated to a similar extent in the surrounding states, emphasizing a wider impact than is indicated by cases. 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 at that time.

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. In contrast, the case rates are unremarkable for these two states, whereas the highest case rates tend to be localized in the Southeast. However, the opposite occurs as well: on December 17, 2020, Tennessee and California have the highest case rates but infections are largely similar to other states.

### 2.3 Infections, overall and by variant, emphasize earlier outbreaks

Figure 5 examines the infection estimates for a selection of states more closely. The top panel shows infection estimates for these states, while the bottom panel disaggregates their deconvolved cases based on the locally circulating variant proportions at the time. These figures show times when the total infections emphasize earlier outbreaks than are indicated by cases alone. During the Ancestral wave, infections in Massachusetts, Idaho, Montana, Louisiana, and Ohio peak earlier than cases. For these states, the infections peak about 17 days earlier on average, with Massachusetts attaining the maximum difference of 26 days. Such trends are also observed in the major Delta wave in the states that present a prominent peak during this time such as

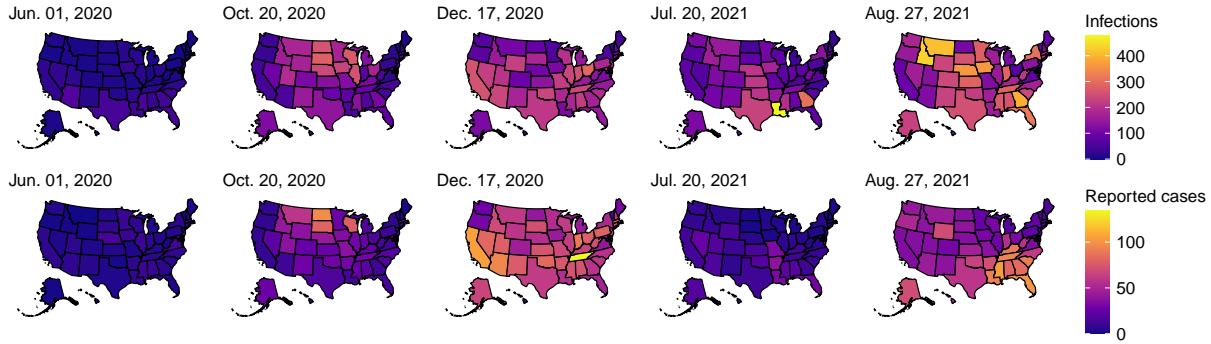


Figure 4: Choropleth maps of the state-level estimates of the number of daily new infections per 100,000 population (top row) and the daily new cases per 100,000 population (bottom row) for five select dates between June 1, 2020 and November 29, 2021. Note that the first date was chosen as a baseline, while the other dates were chosen because they show 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. Note that the colors are scaled differently for infections and cases to enable relative comparisons.

Montana and Louisiana, where infections lead cases by about 41 and 24 days, respectively. The division by variant categories reveals the variant(s) that are behind these waves. For example, in California alone, the Epsilon wave appears to coincide with a second Ancestral wave. To give another example, we can see a major increase in Alpha in Massachusetts over the spring of 2021. To a lesser extent, this trend is apparent for all of the other states, save for California, where Alpha is not a major driver of infections in comparison to the other variants in circulation around that time.

## 2.4 Infections lead hospitalizations according to cross-correlations

We systematically investigate the temporal relationship between infections and hospitalizations with Spearman’s rank-correlation across different lags (Figure 6). The maximum average correlation across states is 0.48, occurring at a lag of 13 days. 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.

With respect to previous literature, the maximum correlation being attained at a lag of 13 days is fairly consistent with estimates of the average time from infection to hospitalization for cases reported in January, 2020 in Wuhan, China (9.7 days) as well as with estimates from across the pandemic in the UK (ranging from 8.0 to 9.7 days)<sup>12,13</sup>. Importantly, our 13 day lag for the U.S. also includes the impact of the reporting pipeline, a delay omitted from the international estimates.

The average correlation is consistently larger for cases than infections (with a difference of about 0.21 at the peaks). This increase is likely due to two reasons. First, many cases are detected contemporaneously with hospitalization: people may first test positive only when they go to the hospital for treatment. Second, unreported infections tend to be less severe and less likely to lead to hospitalization than those that are reported<sup>14</sup>.

## 2.5 IHR estimates tend to be smaller and exhibit less pronounced spikes

As a counterpart to the correlation analysis, we compute the time-varying infection-hospitalization ratios (IHRs) for each state using the correlation maximizing lag (13 days). We similarly compute the case-hospitalization ratios (CHRs) using their correlation maximizing lag for comparison (1 day, Figure 7). For each state, the CHRs tend to be larger in comparison to the IHRs. This is consistent with the claim that reported infections are more likely to require hospitalization than unreported infections.

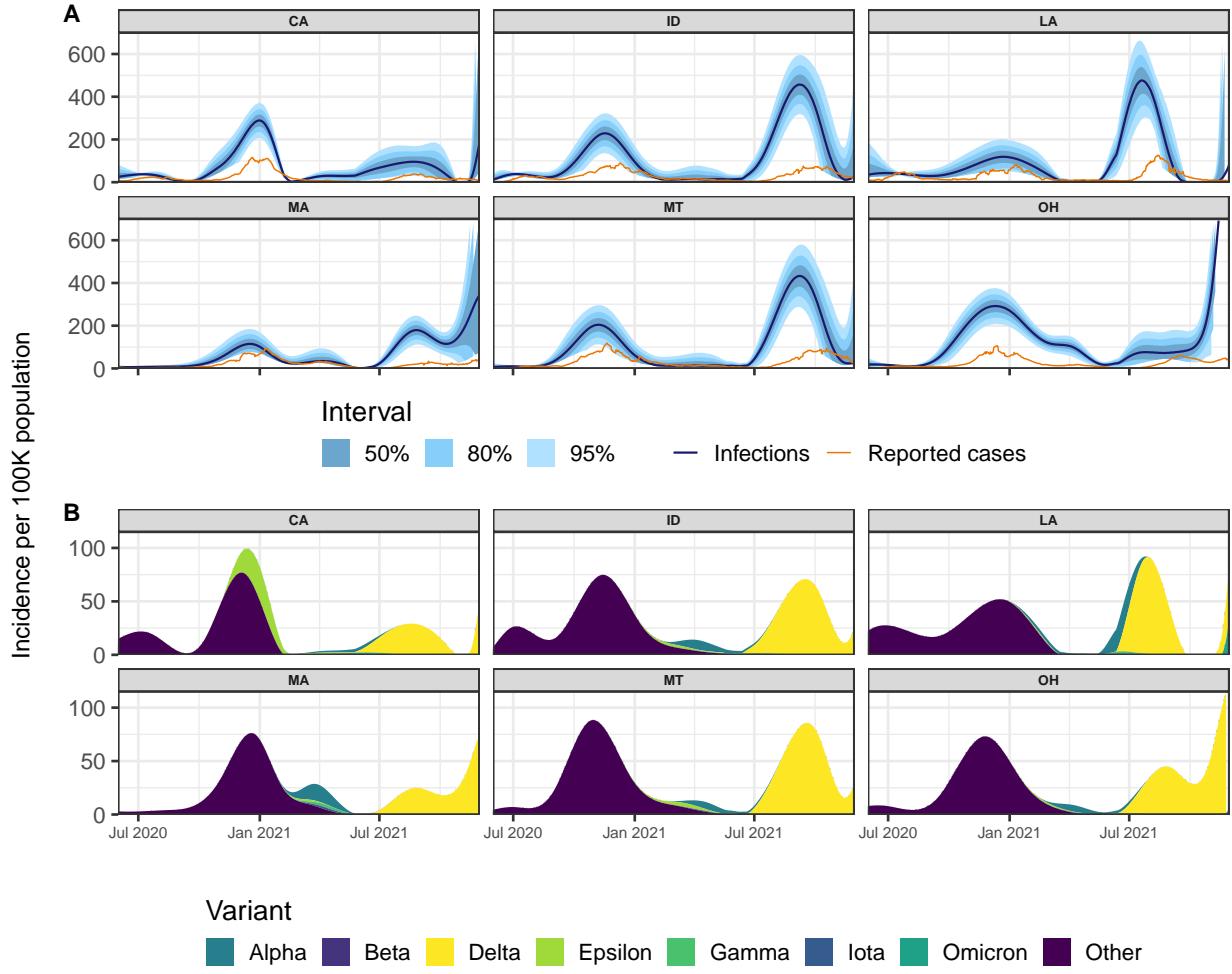


Figure 5: Panel A: Reported cases (orange) and estimates of daily new infections (dark blue) per 100K inhabitants. The blue shaded regions indicate 50, 80, and 95% confidence bands. Panel B: Deconvolved cases colored by variant per 100K inhabitants.

Both IHRs and CHRs exhibit similar geospatial and temporal 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, some states exhibit a striking increase in hospitalizations in mid-2021, which coincides with the rapid takeover of the Delta variant<sup>15</sup>. This finding aligns with previous studies that found an increased risk in hospitalization due to Delta<sup>16,17</sup>. Interestingly, when the Ancestral variant dominates in 2020, there is a spike in the IHRs that rivals or sometimes even surpasses that which is observed during Delta. This situation is most readily observable in New England as well as in some Western states such as Arizona, New Mexico, and Wyoming.

Overall, the relationship between infections and hospitalizations is complicated. We observe intermittent spikes that punctuate longer periods where the IHRs are relatively stable, remaining below 0.1 hospitalizations per infection. While we computed the IHRs and CHRs for all states, it is important to note that both likely vary within states and depend on confounding variables such as age and the presence of major comorbidities<sup>18</sup>. Therefore, it would be beneficial to account for such variables in their calculations by, for example, stratifying infections and hospitalizations by age to produce age-specific estimates of the IHRs for each state<sup>19</sup>.

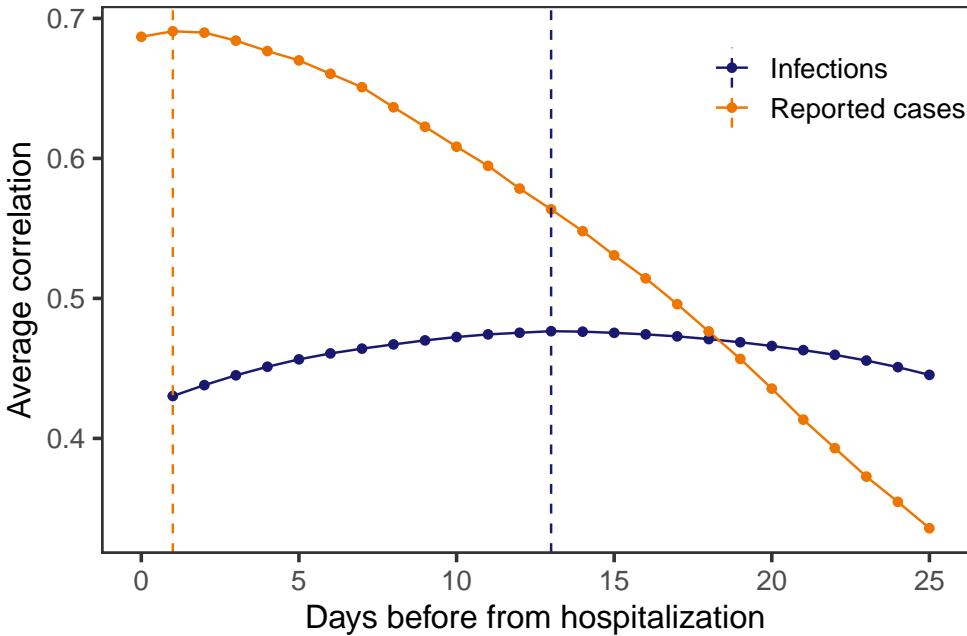


Figure 6: Spearman’s rank correlation between each of cases and infections with hospitalizations per 100,000. These are calculated for each lag, state, and rolling window of 61 days before averaging. The vertical dashed lines indicate the lags for which the highest average correlation is attained.

### 3 Discussion

We retrospectively estimated daily incident infections for each U.S. state over the period June 1, 2020 to November 29, 2021. Our estimates suggest, confirming 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 bias 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<sup>9,20</sup>. 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 state- and time-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<sup>21,22</sup>. Unlike previous approaches that use a single delay distribution to generate estimates for all states<sup>23–25</sup>, 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<sup>25</sup>, 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<sup>26–28</sup>.

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

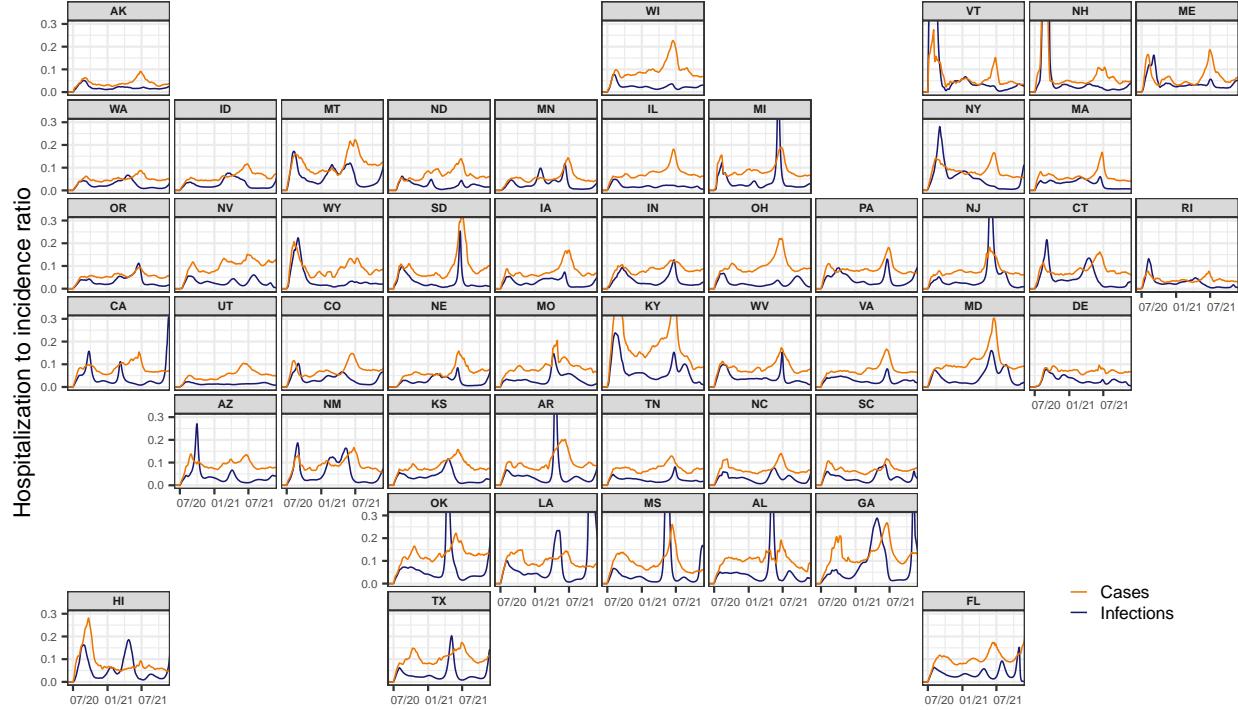


Figure 7: Time-varying IHR and CHR estimates for each state from June 1, 2020 to November 29, 2021, obtained using the respective correlation maximizing lag from Section 2.4. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious day of the week effects. Also note that the different starting points across states are due to the availability of the hospitalization data.

seropositivity (Section 4.3), we more accurately estimate this ratio. However, future work could refine this analysis. Because the waning of immunity is likely to be variant-dependent<sup>29</sup>, 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<sup>30–32</sup>. 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<sup>33–36</sup>, 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<sup>37</sup>. In follow-up work, we extend our construction of daily incident infections for the states where sufficient data is available to track reinfections and to incorporate wastewater data. An alternative approach could use surveillance streams such as those from surveys, helplines, or medical records could potentially be integrated if they offer a sufficiently strong signal of the disease intensity over time<sup>6,38</sup>.

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 process 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.

## 4 Methods

In what follows, we provide details on how we estimate the daily incident infections for each state over the considered time period of June 1, 2020 to November 29, 2021 and the data we used to achieve this. [Figure 2](#) provides a visual summary of the major analysis tasks, which can be summarized as follows: First, we estimate the time-varying delays from positive specimen to report date using a line list and use these to deconvolve reported cases to the date of the positive specimen collection. Then, we estimate the delay from symptoms to positive specimen, combine this with variant specific infection-to-symptoms delays, and use this delay to push back to the date of infection. The resulting infection onset estimates are aggregated across the variant categories and adjusted to account for the unreported infections by using state-specific, time-varying seroprevalence data in an antibody prevalence model.

### 4.1 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), as estimated from de-identified patient-level line list data on COVID-19 cases from the CDC<sup>[39](#)</sup>. Simply shifting cases back in time would fail to reflect the fact that some cases take much longer to be reported than others (see Supplementary Methods Section [S1.1](#) for further details).

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 line list was used to estimate the necessary delay distribution. 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  $\ell$  at time  $t$ , as disseminated by the John Hopkins Center for Systems Science and Engineering (JHU CSSE)<sup>[1](#)</sup> and retrieved with the COVIDcast API<sup>[38](#)</sup>. Let  $\pi_{\ell,t}(k)$  be the probability that cases with positive specimen collection on time  $t-k$  are reported at  $t$ . Then, we model  $y_{\ell,t}$  as Gaussian with mean

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

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

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

The two parts of this optimization problem trade data fidelity (the sum of squared errors) with smoothness in the resulting estimates (the fourth order differences of  $\mathbf{x}$ ). The tuning parameter  $\lambda$  determines the relative importance of these competing goals. The solution to this minimization problem is an adaptive piecewise cubic polynomial<sup>[40,41](#)</sup> and can be accurately computed with ease<sup>[24,42](#)</sup>. We select  $\lambda$  with cross-validation to minimize the out-of-sample reconvolution error.

To estimate the  $\pi_{\ell,t}(k)$  for all states  $\ell$ , times  $t$ , and delays  $k$ , we use the CDC line list<sup>[39,43](#)</sup>. The line list contains three key dates of interest for many cases that will eventually appear in case reports: symptom onset, positive specimen collection, and report to the CDC. Handling missingness and imputation in these dates requires careful attention, and so additional details and justifications are provided in Supplementary

Methods Section S1.2. Define  $z_{\ell,t}$  to be a case report occurring at time  $t$  in location  $\ell$ , and let  $\pi_{\ell,t}(k)$  be the probability that  $z_{\ell,t}$  has a positive specimen collected  $k$  days earlier. We assume that all positive specimen results 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 the time of interest  $t$ . Then, we compute the observed probability mass function (pmf)

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

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  $\ell$ . Next, let  $\alpha_{\ell,t}$  be the ratio of  $N_{\ell,t}$  to the number of cases reported by JHU CSSE<sup>1</sup> 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 of the JHU cases reports appear in the CDC line list (and vice versa). 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  $\pi_{\ell,t}(k)$ . Additional minor details and justification for the delay distribution calculations are deferred to Supplementary Methods Section S1.3.

## 4.2 From positive specimen collection to infection onset

To continue, pushing positive specimen collection time back to infection onset (Step 2 in Figure 2), we will 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. This means that both the probabilities and the observations must be replaced with variant-specific quantities.

For the observations, we use our estimates from Section 4.1,  $\hat{x}_{\ell}$ , 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<sup>15,44</sup>, and estimate a multinomial logistic regression model. This procedure is now standard<sup>45–47</sup>, so we defer details to Supplementary Materials Section S1.4. 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 4.1), with variant-specific incubation periods. The convolution of these yields a distribution  $\tau_{j\ell,t}(k)$ . Details on the convolution and its inputs are deferred to Supplementary Methods Sections S1.5.1 to S1.5.3.

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

$$\mathbb{E}[\hat{v}_{j\ell,t} \hat{x}_{\ell,t} \mid u_{j\ell,s}, s \leq t] = \sum_k \tau_{j\ell,t-k}(k) u_{j\ell,t-k} \quad (4)$$

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

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

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

## 4.3 Inverse reporting ratio and the antibody prevalence model

To capture the unreported infections, it is necessary to adjust these deconvolved case estimates by the ratio of the true number of new infections to the new 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<sup>48</sup>, so we use the change in subsequent seroprevalence measurements to estimate *all* new infections rather than just those eventually appearing as cases.

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<sup>49,50</sup>. See Supplementary Methods Section S1.6 for additional details. Each of these provides seroprevalence estimates along with confidence intervals. In order 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 modelled as Gaussian,

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

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

with source-specific measurement errors that scale proportional to the reported confidence intervals, respectively  $w_{\ell,m}^1$  and  $w_{\ell,m}^2$ .

To complete the model, we suppose that latent seroprevalence is modeled as a Guassian 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  $m$ :

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

where  $\hat{u}_{\ell,t}$  are deconvolved cases (from Section 4.2),  $z_m$  is the fraction of reinfections, and  $a_{\ell,m}$  is the inverse reporting ratio. Note that  $\gamma$  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<sup>33</sup>, and these estimates are broadly similar to those in other locations with available data<sup>33–36</sup>. Finally, we specify the time-varying evolution of the inverse reporting ratio as Gaussian with expectation,

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

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  $\mathbf{a}_\ell$ . Writing it in this way allows for convenient handling of missing or irregularly-spaced survey measurements, extrapolation of the estimated latent quantities before and after the period of observed seroprevalence measurements, and maximum likelihood estimates of the latent variables and all unknown parameters. Details of this methodology and the computation of the associated uncertainty measurements are deferred to Supplementary Methods Section S1.7.

#### 4.4 Lagged correlation to hospitalizations and time-varying IHRs

From the COVIDcast API<sup>38</sup>, 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. To that end, we consider a wide range of possible lag values ranging from 1 to 25 days. Zero and negative lags are not considered because COVID-19 infection onset must precede hospitalization. To remove day of the week effects, both the infection and hospitalization signals are averaged over a 7-day, center-aligned, moving window before their conversion to rates.

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.

## Data availability

The Nationwide Commercial Laboratory Seroprevalence Survey and 2020-2021 Nationwide Blood Donor Seroprevalence Survey data are both freely available at <https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv> and <https://data.cdc.gov/Laboratory-Surveillance/2020-2021-Nationwide-Blood-Donor-Seroprevalence-Su/mtc3-kq6r>. The COVID-19 Case Surveillance Restricted Access Detailed Data requires a Restriction Information and Data Use Restrictions Agreement. More information on the dataset and the agreement can be found [https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t/about\\_data](https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t/about_data). The JHU reported case and hospital admission data are publicly accessible through the COVIDcast API, <https://cmu-delphi.github.io/delphi-epidata/api/covidcast.html>. The data used to construct the incubation periods was obtained from <https://elifesciences.org/articles/57149>, <https://pubmed.ncbi.nlm.nih.gov/35627870/>, <https://pubmed.ncbi.nlm.nih.gov/34849500/>, and <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8834809/>. The data used to estimate the variant proportions in circulation is freely available at <https://github.com/hodcroftlab/covariants>. The prior inverse ascertainment ratio estimates were extracted from Table 1 in <https://www.nature.com/articles/s41467-020-19652-6>. The fraction of new infections are based on estimates from [https://wwwnc.cdc.gov/eid/article/28/10/22-1045\\_article](https://wwwnc.cdc.gov/eid/article/28/10/22-1045_article). Finally, the state population estimates are from a U.S. Census Bureau press release at <https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html>.

## 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] Ontario Agency for Health Protection and Promotion. COVID-19 variant of concern Omicron (B.1.1.529): Risk assessment. [https://www.publichealthontario.ca/-/media/documents/ncov/voc/2022/01/covid-19-omicron-b11529-risk-assessment-jan-6.pdf?sc\\_lang=en](https://www.publichealthontario.ca/-/media/documents/ncov/voc/2022/01/covid-19-omicron-b11529-risk-assessment-jan-6.pdf?sc_lang=en) (2022).
- [11] Garrett, N. *et al.* High rate of asymptomatic carriage associated with variant strain Omicron. *MedRxiv* (2022).
- [12] Linton, N. M. *et al.* Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine* **9**, 538 (2020).
- [13] Ward, T. & Johnsen, A. Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLoS One* **16**, e0257978 (2021).
- [14] Sallahi, N. *et al.* Using unstated cases to correct for COVID-19 pandemic outbreak and its impact on easing the intervention for Qatar. *Biology* **10**, 463 (2021).
- [15] Hodcroft, E. CoVariants: SARS-CoV-2 mutations and variants of interest. <https://covariants.org> (2021).
- [16] Twohig, K. A. *et al.* Hospital admission and emergency care attendance risk for SARS-CoV-2 Delta (B. 1.617. 2) compared with Alpha (B. 1.1. 7) variants of concern: A cohort study. *The Lancet Infectious Diseases* **22**, 35–42 (2022).
- [17] Nyberg, T. *et al.* Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 Omicron (B. 1.1. 529) and Delta (B. 1.617. 2) variants in England: A cohort study. *The Lancet* **399**, 1303–1312 (2022).
- [18] Russell, C. D., Lone, N. I. & Baillie, J. K. Comorbidities, multimorbidity and COVID-19. *Nature Medicine* **29**, 334–343 (2023).

- [19] Fox, S. J. *et al.* Disproportionate impacts of COVID-19 in a large US city. *PLOS Computational Biology* **19**, e1011149 (2023).
- [20] Dunkel, S. COVID-19 case numbers: Why the delay in reporting? <https://www.tpchd.org/Home/Components/Blog/Blog/21448> (2020).
- [21] Unwin, H. J. T. *et al.* State-level tracking of COVID-19 in the United States. *Nature Communications* **11**, 6189 (2020).
- [22] Center for the Ecology of Infection Diseases. COVID-19 portal. <https://www.covid19.uga.edu/nowcast.html> (2020).
- [23] 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).
- [24] Jahja, M., Chin, A. & Tibshirani, R. J. Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. *Statistical Science* **37**, 207–228 (2022).
- [25] Miller, A. C. *et al.* Statistical deconvolution for inference of infection time series. *Epidemiology* **33**, 470–479 (2022).
- [26] 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).
- [27] 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).
- [28] 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).
- [29] 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).
- [30] 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).
- [31] Pulliam, J. R. *et al.* Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science* **376**, eabn4947 (2022).
- [32] 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).
- [33] Ruff, J. *et al.* Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. *Emerging Infectious Diseases* **28**, 1977 (2022).
- [34] New York State COVID-19 reinfection data. <https://coronavirus.health.ny.gov/covid-19-reinfection-data> (2021).
- [35] Hawaii Department of Health COVID-19 reinfection data. [https://health.hawaii.gov/coronavirusedisease2019/files/2022/09/reinfection\\_report\\_2022-09-28.pdf](https://health.hawaii.gov/coronavirusedisease2019/files/2022/09/reinfection_report_2022-09-28.pdf) (2022).
- [36] Reported COVID-19 reinfections in Washington State. <https://doh.wa.gov/sites/default/files/2022-02/421-024-ReportedReinfections.pdf> (2022).
- [37] McManus, O. *et al.* Predicting COVID-19 incidence using wastewater surveillance data, Denmark, October 2021–June 2022. *Emerging Infectious Diseases* **29**, 1589 (2023).

- [38] Reinhart, A. *et al.* An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences* **118**, e2111452118 (2021).
- [39] 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).
- [40] Tibshirani, R. J. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42**, 285–323 (2014).
- [41] 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).
- [42] Ramdas, A. & Tibshirani, R. J. Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics* **25**, 839–858 (2016).
- [43] Centers for Disease Control and Prevention. COVID-19 case surveillance public use data. <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf> (2020).
- [44] Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).
- [45] Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).
- [46] 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).
- [47] Figgins, M. D. & Bedford, T. SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. *MedRxiv* 2021–12 (2021).
- [48] Centers for Disease Control and Prevention. COVID Data Tracker. <https://covid.cdc.gov/covid-data-tracker/#national-lab> (2020).
- [49] 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).
- [50] 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).
- [51] 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).
- [52] World Health Organization. Tracking SARS-CoV-2 variants. <https://www.who.int/activities/tracking-SARS-CoV-2-variants> (2021).
- [53] 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).
- [54] 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).
- [55] Tindale, L. C. *et al.* Evidence for transmission of COVID-19 prior to symptom onset. *eLife* **9**, e57149 (2020).

- [56] 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).
- [57] 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).
- [58] 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).
- [59] 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).
- [60] 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).
- [61] Bajema, K. L. *et al.* Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. *JAMA Internal Medicine* **181**, 450–460 (2021).
- [62] 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<sup>44</sup>, 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.

# Supplementary Information: Retrospective estimation of latent COVID-19 infections before Omicron in the U.S.

## S1 Supplementary Methods

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

### S1.1 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

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

Figure S1 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.

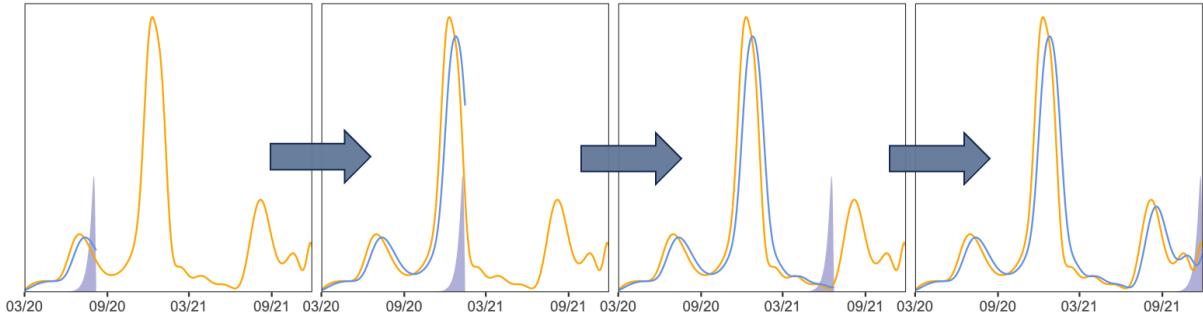


Figure S1: A general depiction of convolving smoothed cases (orange line) with the corresponding delay probabilities (shaded blue area) to get the convolved estimates (blue line) over four different times.

### S1.2 Additional details on the date fields in the CDC line list

Because the 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 S1 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% 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<sup>24</sup>. So it is possible

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 S1: 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).

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 CDC restricted line list contains 74,849,225 cases (rows) in total compared to 84,714,805 cases reported by the JHU CSSE; that is, line list is missing about 10 million cases. The extent that this issue impacts each state is shown in Figure S2, from which it is clear the fraction of missing cases is substantial for many states, often surpassing 50%<sup>24</sup>. 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.

### S1.3 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  $\ell$  with positive specimen collection date  $t$  and delay  $d$ . For each  $j = 1, 2, \dots$ , define  $z_{\ell b j} = \sum_{d=50j}^{50(j+1)} \sum_t x_{\ell t}(d)$ . Then, locate the collection of potentially problematic combinations

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

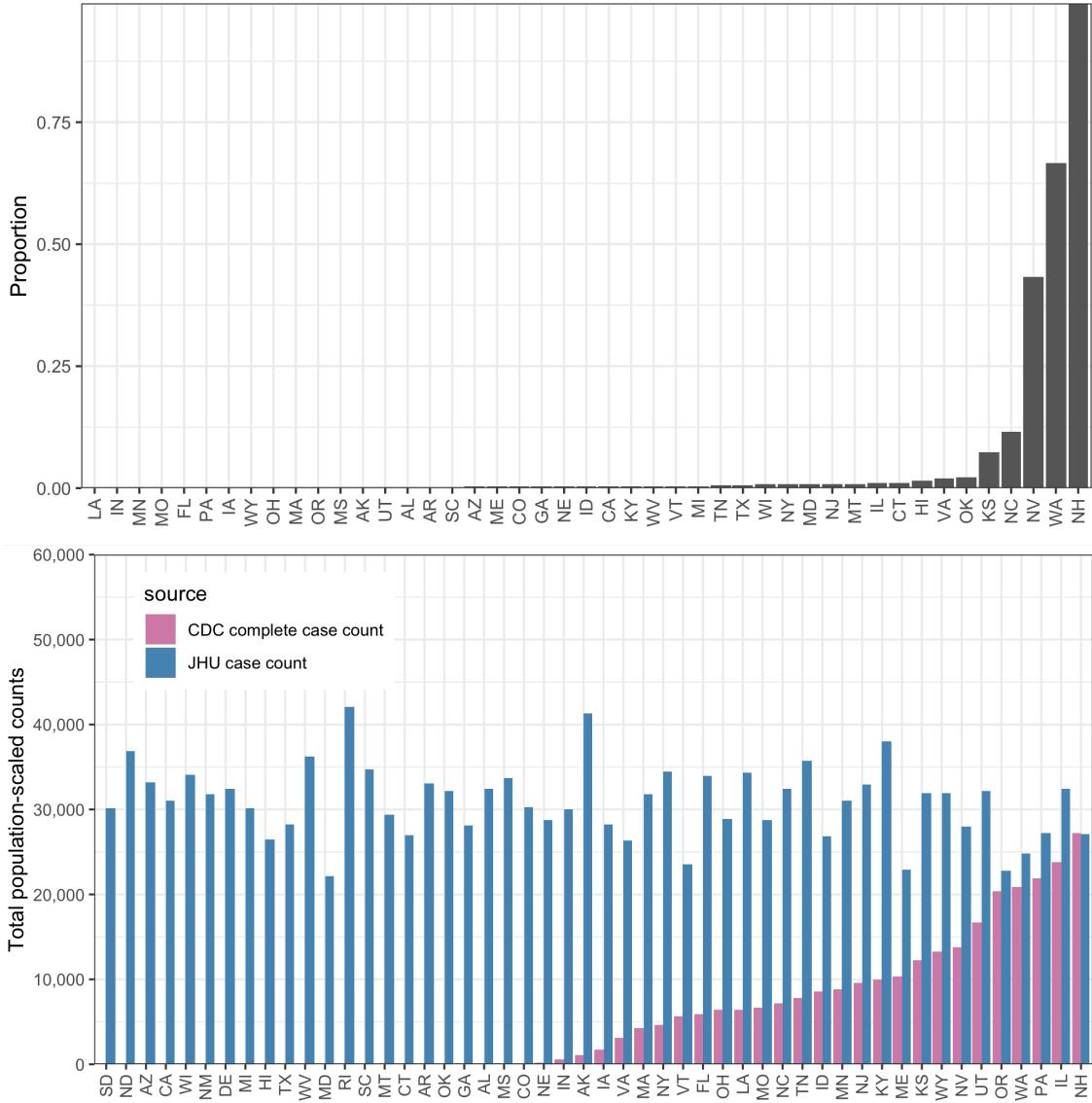


Figure S2: Top panel: Proportion of complete cases with zero delay between positive specimen and report date in the restricted CDC line list dataset. Bottom panel: Complete case counts by state in the CDC line list versus the cumulative complete case counts from JHU CSSE as of June 6, 2022. All counts have been scaled by the 2022 state populations as of July 1, 2022 from <sup>51</sup>.

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 4.1)  $\tilde{p}_{\text{NH},t}(k)$  would be a point mass at  $k = 0$  and the weight  $\alpha_{\text{NH},t} = 1$  resulting  $\hat{\pi}_{\text{NH},t}(k)$  also being a point mass at  $k = 0$ . In this specific case, we force  $\alpha_{\text{NH},t} = \min_{\ell \neq \text{NH}} \alpha_{\ell,t}$ .

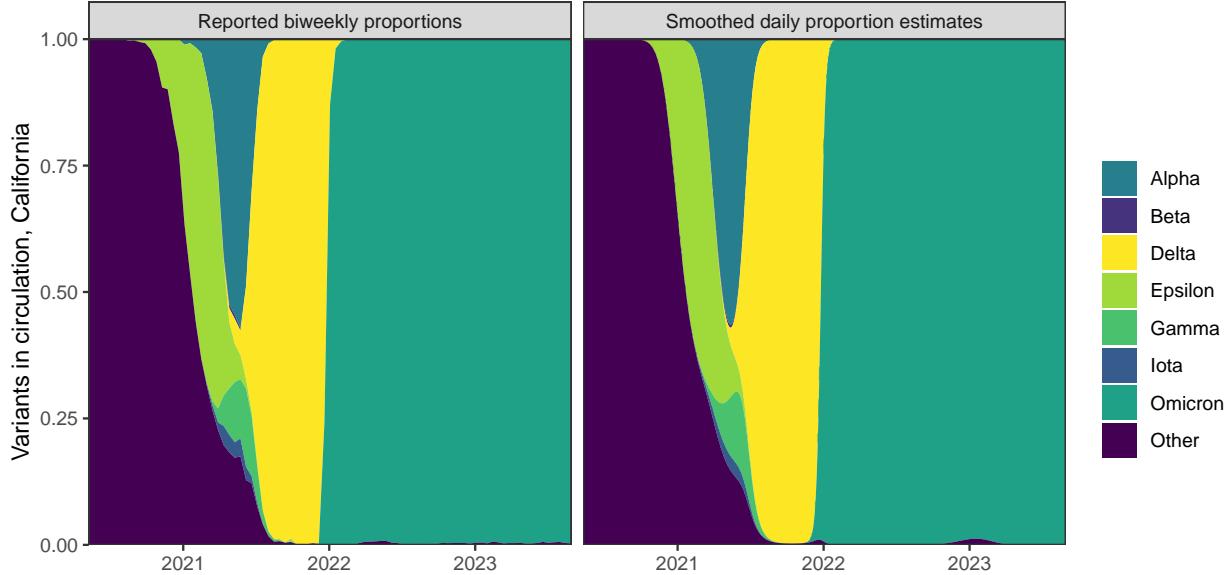


Figure S3: Left: Original biweekly proportions of the variants in circulation for California. Right: Daily proportions of the variants in circulation for California.

#### S1.4 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<sup>15,44</sup>. 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<sup>45–47</sup>.

We let  $V_{j\ell,t}$  to be the probability of a new cases at time  $t$  in location  $\ell$  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

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

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 S3 shows the proportions by variant for California before (left) and after (right) the smoothing procedure.

#### S1.5 Constructing the delay from infection to test

The result of Step 1 (Section 4.1) is  $\hat{x}_{\ell,t}$ , 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.

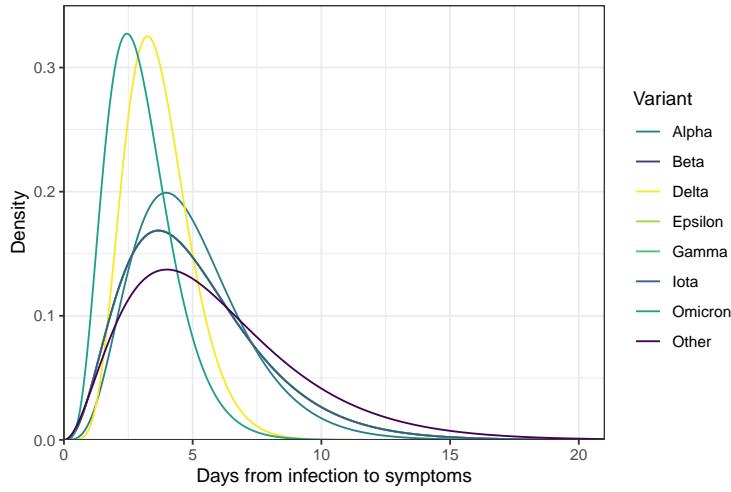


Figure S4: Gamma density for the incubation period of each of the eight variant categories. Note that the Ancestral variant uses reported shape and scale parameters<sup>55</sup>, while the remaining variants convert reported estimates for the mean and variance<sup>26,27,56</sup> using the method of moments to produce the gamma parameters.

### S1.5.1 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 S3), 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<sup>52</sup>, while we include the Epsilon (California) and Iota (New York) variants because of large impact on those and neighbouring states<sup>53,54</sup>.

The incubation period of the Ancestral variant has been modelled as a gamma distribution<sup>55</sup>, 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<sup>26,27,56</sup>. Therefore, we use method of moments to match the mean and variance to estimate the gamma parameters, using the moment equations given in Section 4.1. Then, we discretize each resulting density shown in Figure S4 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<sup>57</sup> and that an asymptomatic infection will resolve within 21 days<sup>58,59</sup>.

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<sup>15</sup>. 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.

### S1.5.2 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 4.1 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 S5. Unlike the delay from positive specimen collection to report, the delay 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

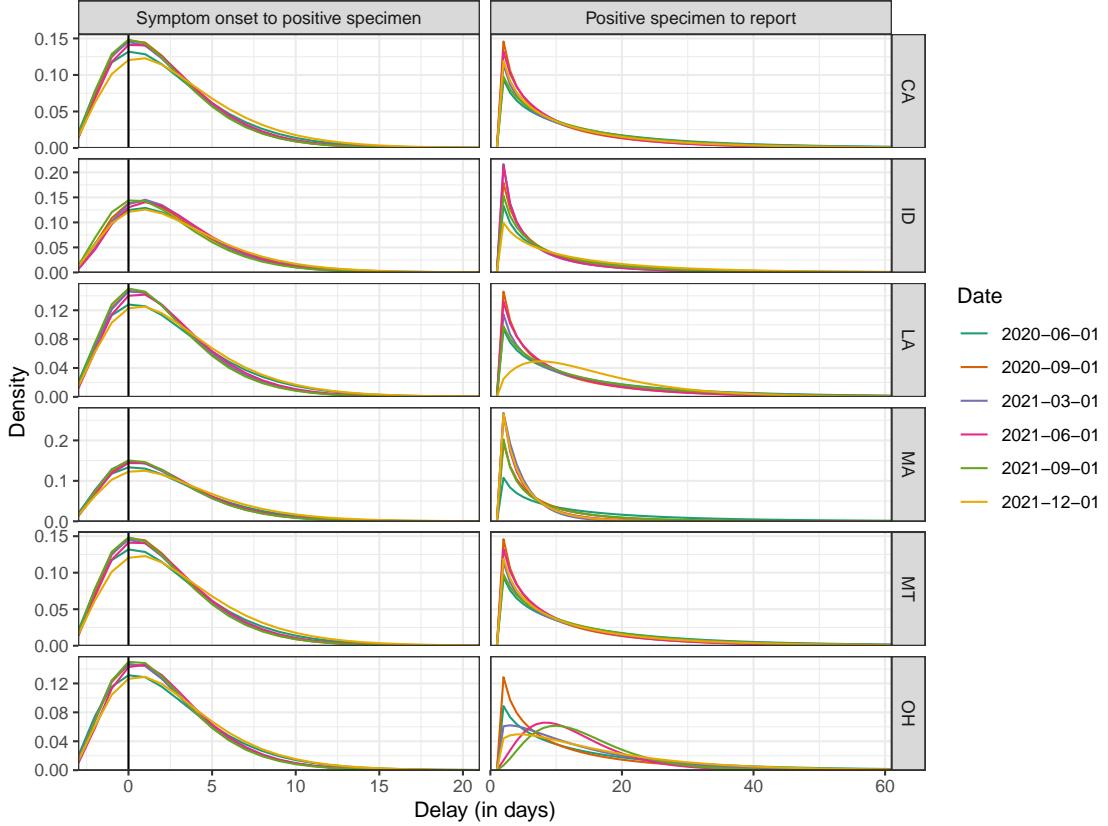


Figure S5: Depictions of the estimated delay from symptom onset to positive specimen date (left) and from positive specimen date to report date (right) for a sample of six states over several dates.

of symptoms. Required regular testing for jobs in health care settings, construction, or the film industry could also produce negative delays.

### S1.5.3 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 Section S1.5.1, denoted by  $\{i_j(k) : 0 < k \leq 21\}$  with the location-time-specific symptom-to-positive-test distributions from Section S1.5.2, 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 S5 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 S4.

## S1.6 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<sup>49,50</sup>. In the former, the CDC collaborated with 17 blood collection organizations in the largest nationwide COVID-19 seroprevalence survey to date<sup>49</sup>. The blood donation samples were used to construct monthly seroprevalence estimates for nearly all states from July 2020 to December 2021<sup>60</sup>. 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<sup>61</sup>) for the antibodies to the virus. The resulting dataset contains seroprevalence estimates for a

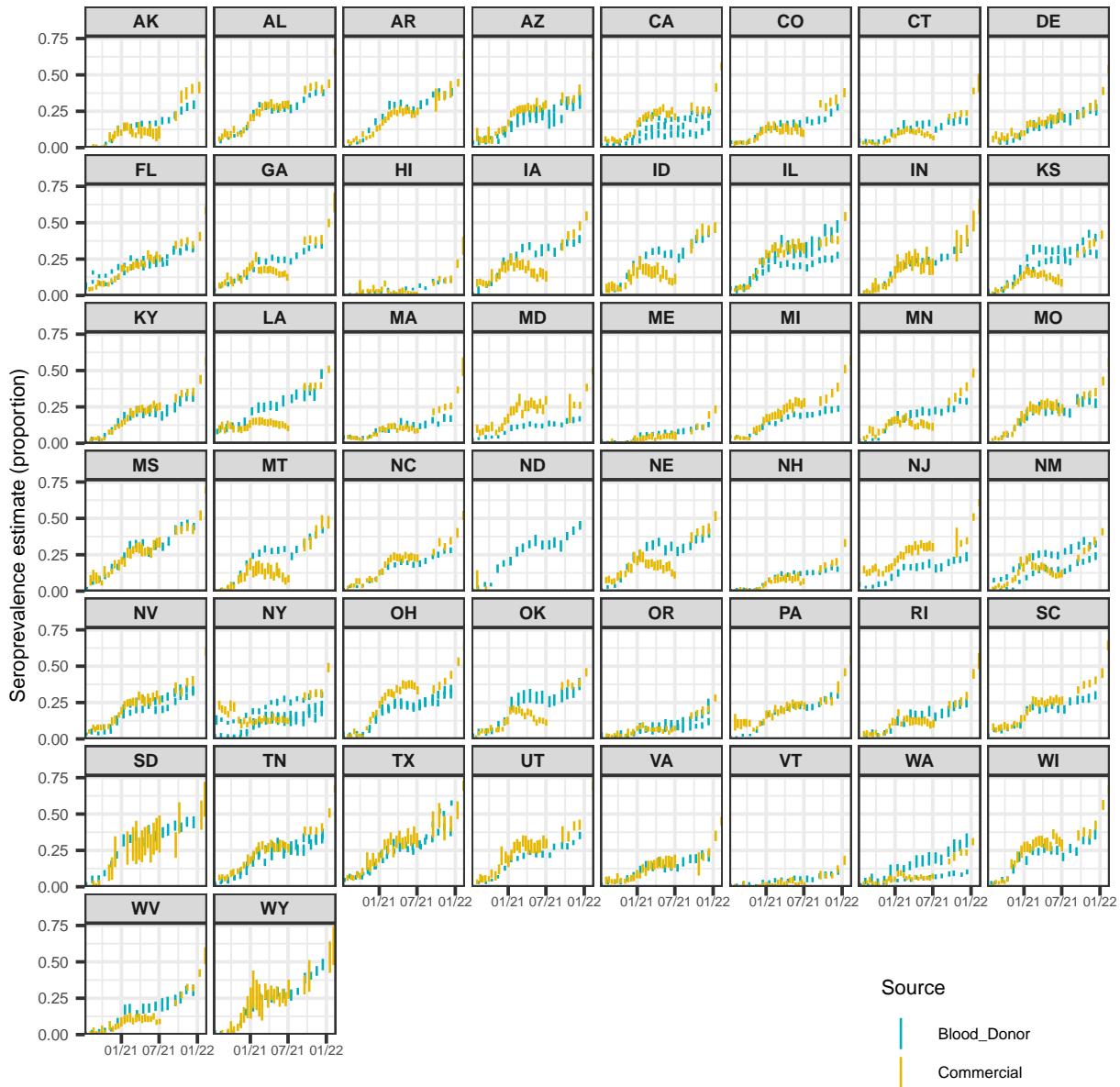


Figure S6: A comparison of the seroprevalence estimates from the Commercial Lab Seroprevalence Survey dataset (yellow) and the 2020–2021 Blood Donor Seroprevalence Survey dataset (blue). Note that the maximum and the minimum of the line ranges are the provided 95% confidence interval bounds to give a rough indication of uncertainty.

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<sup>48,60,61</sup>. 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 S6). 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.

## S1.7 State space representation of the antibody prevalence model

The antibody prevalence model described in Section 4.3 can be expressed as a linear Gaussian state space model<sup>62</sup>. For  $m = 1, \dots, M$ , let  $\alpha_m$  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

$$y_m = Z\alpha_m + \sigma_r^2 \epsilon_m, \quad \epsilon_m \sim N(0, H_m) \quad (12)$$

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

where  $\alpha_1 \sim N(a_1, P_1)$  and  $\epsilon_m$  and  $\eta_m$  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):

$$\begin{aligned} Z &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} & H_m &= \begin{bmatrix} w_m^1 & 0 \\ 0 & w_m^2 \end{bmatrix} \\ \alpha_m &= \begin{bmatrix} s_m \\ a_m \\ a_{m-1} \\ a_{m-2} \end{bmatrix} & T_m &= \begin{bmatrix} (1 - \gamma) & \hat{u}_m^\Sigma (1 - z_m) & 0 & 0 \\ 0 & 3 & -3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} & R &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \\ Q &= \begin{bmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{bmatrix} & a_1 &= \begin{bmatrix} \tilde{s}_1 \\ \tilde{a}_1 \\ \tilde{a}_1 \\ \tilde{a}_1 \end{bmatrix} & P_1 &= \begin{bmatrix} \sigma_{\tilde{s}_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\tilde{a}_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\tilde{a}_1}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\tilde{a}_1}^2 \end{bmatrix} \end{aligned}$$

where  $\sigma_r^2$  is the variance of observations,  $\sigma_\epsilon^2$  is the variance of the seroprevalence estimates,  $\sigma_\eta^2$  is the trend variance, and  $\hat{u}_m^\Sigma$  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  $\sigma_\eta^2$  is a multiple of  $\sigma_\epsilon^2$ .

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 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  $\alpha_1$  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,  $\sigma_{\tilde{s}_1}^2$ , is taken to be the mean of the standard errors of the two seroprevalence estimates. For the initial values of the trend components, we use the inverse of the ascertainment ratio estimate for each as of June 1, 2020 from Table 1 in<sup>21</sup>: the mean for  $\tilde{a}_1$  and half the width of the confidence interval converted to the inverse for  $\sigma_{\tilde{a}_1}^2$ .

The initial  $\sigma_r^2$  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  $\sigma_\epsilon^2$  and  $\gamma$  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<sup>62</sup>, 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<sup>51</sup>.