

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

Rachel Lobay^{a,1}, Maria Jahja^b, Ajitesh Srivastava^c, Ryan J. Tibshirani^d, and Daniel J. McDonald^a

^aDepartment of Statistics, The University of British Columbia

^bDepartment of Statistics & Data Science, Carnegie Mellon University

^cDepartment of Computer and Electrical Engineering, University of Southern California

^dDepartment of Statistics, The University of California, Berkeley

Version: May 17, 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. Our results cover all states at the daily frequency, incorporate variant-specific incubation periods, and account for 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^{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 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 omitted⁴. 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 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.

¹To whom correspondence should be addressed. E-mail: rachel.lobay@stat.ubc.ca

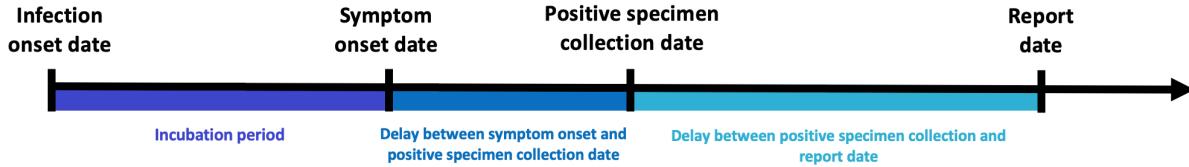


Figure 1: Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

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 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. We then 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. These estimates as well as the R and Python code used to produce them can be found at <https://github.com/cmu-delphi/latent-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 2 to 4). 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, 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 2 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

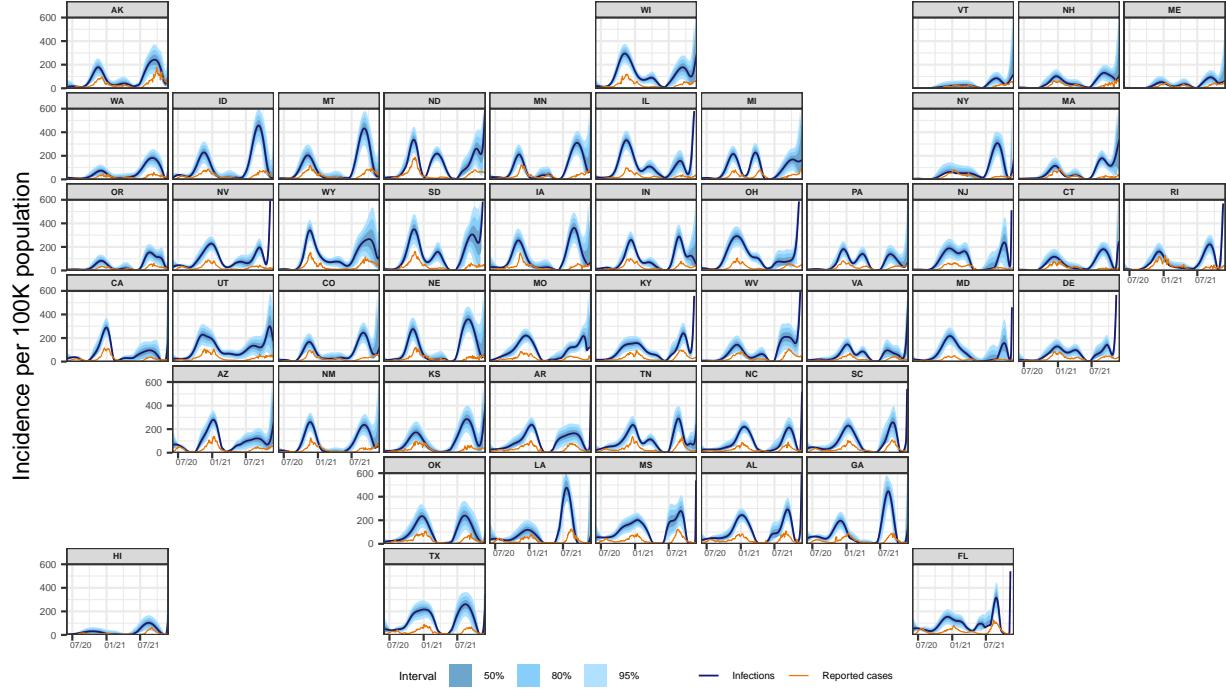


Figure 2: Estimates of the number of daily new infections per 100,000 population for each U.S. state from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% intervals for the estimates, while the orange line represents the trailing 7-day average of reported cases per 100,000.

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 2 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^{10,11}.

The ratio of cases to infections decreases with time. While the Delta wave is somewhat apparent from

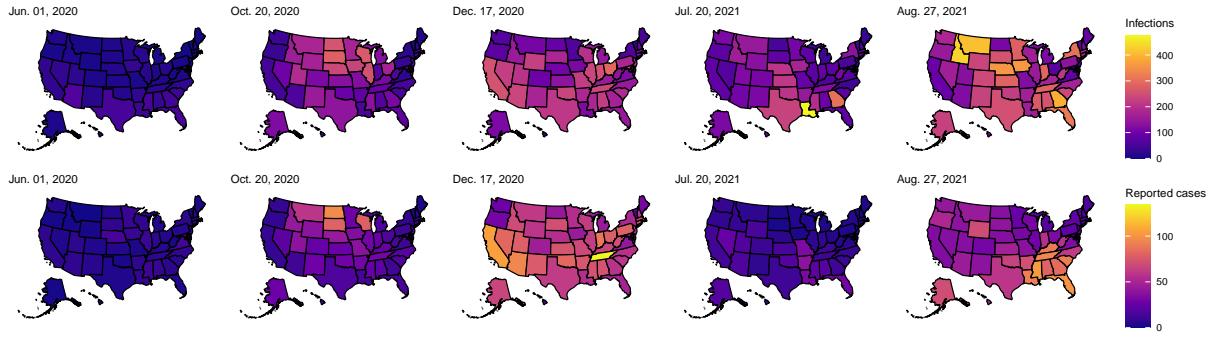


Figure 3: Choropleth maps of the state-level estimates of the number of daily new infections per 100,000 population (top row) and the daily new cases per 100,000 population (bottom row) for five select dates between June 1, 2020 and November 29, 2021. Note that the first date was chosen as a baseline, 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.

the case counts for all states (Figure 2), 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 case-report ratio was larger in earlier waves, and its effects 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 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 4 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

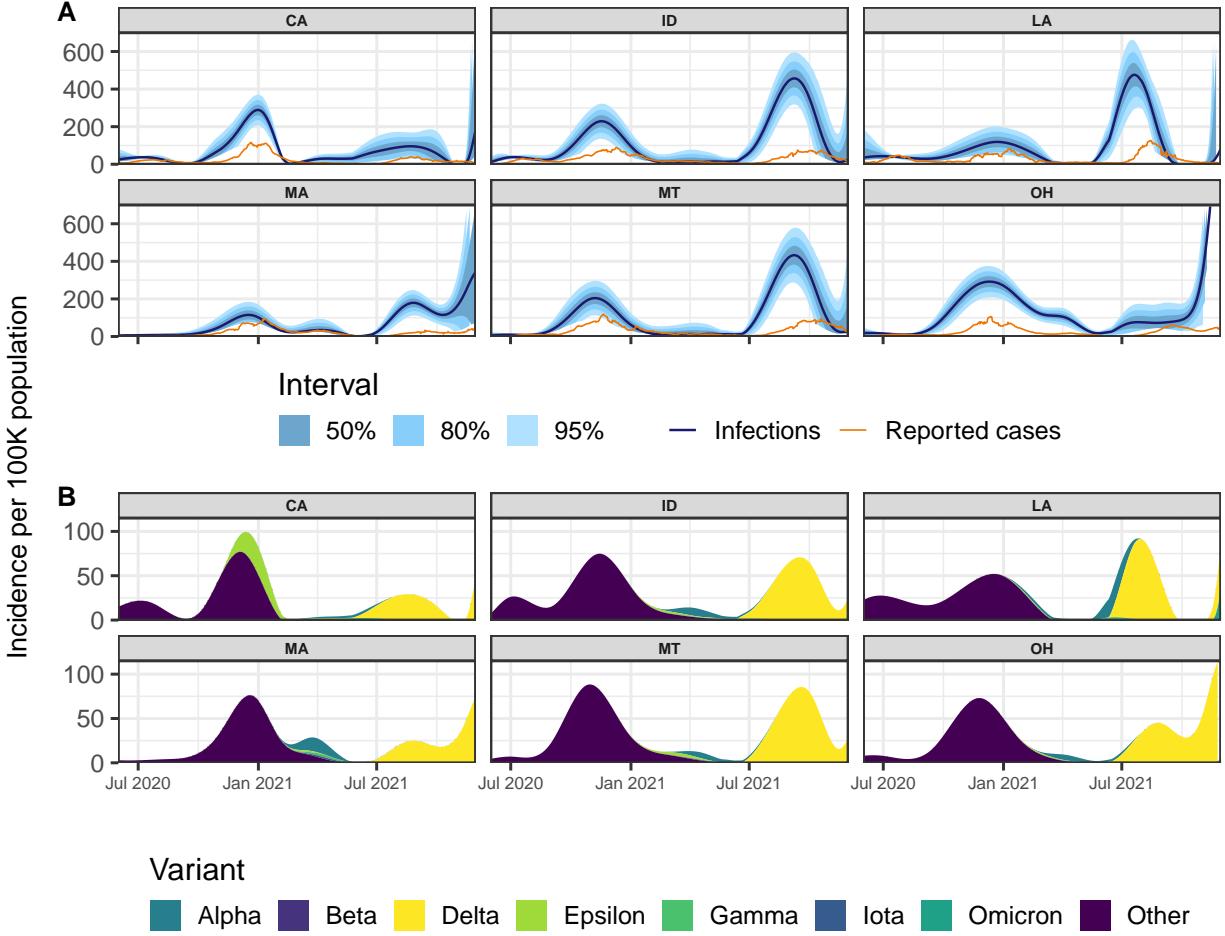


Figure 4: Panel A: Reported cases (orange) and estimates of daily new infections (dark blue) per 100K inhabitants. The blue shaded regions indicate 50, 80, and 95% confidence bands. Panel B: Deconvolved cases colored by variant per 100K inhabitants.

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 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 5). 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,

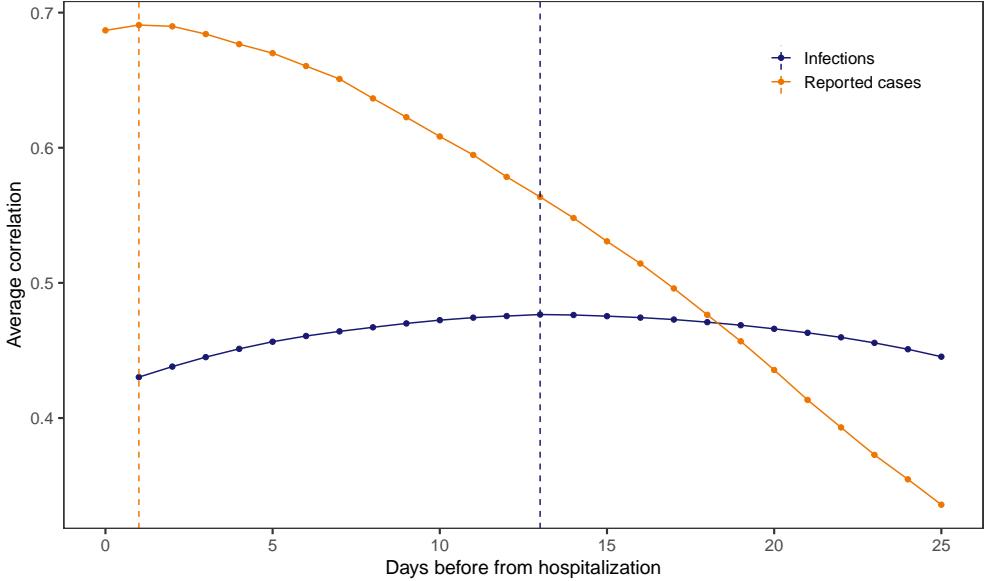


Figure 5: Spearman’s rank correlation between each of cases and infections with hospitalizations per 100,000. These are calculated for each lag, state, and rolling window of 61 days before averaging. The vertical dashed lines indicate the lags for which the highest average correlation is attained.

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)^{12,13}. 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 reported¹⁴.

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 6). 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.

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 variant¹⁵. This finding aligns with previous studies that found an increased risk in hospitalization due to Delta^{16,17}. 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 comorbidities¹⁸. Therefore, it would be beneficial to account for such variables in their calculations by, for example, stratifying

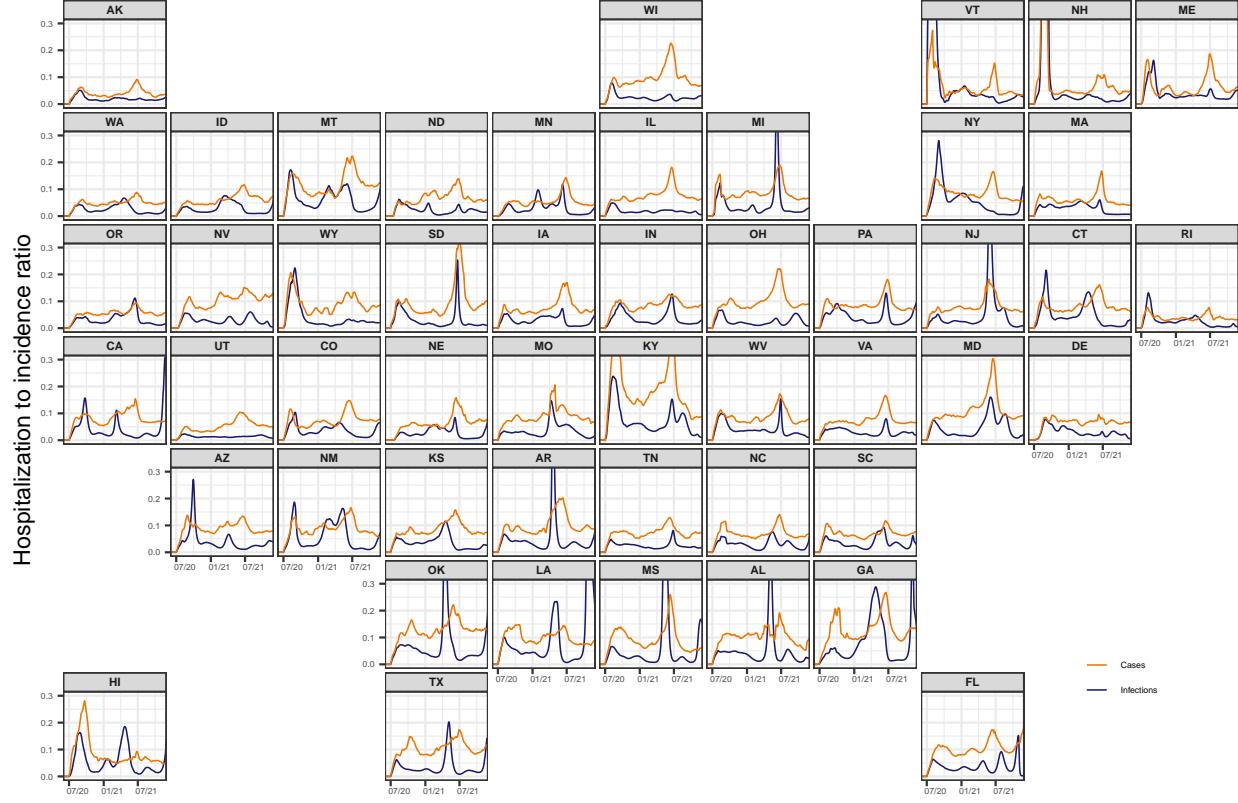


Figure 6: Time-varying IHR and CHR estimates for each state from June 1, 2020 to November 29, 2021, obtained using the respective correlation maximizing lag from Section 2.4. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious day of the week effects. Also note that the different starting points across states are due to the availability of the hospitalization data.

infections and hospitalizations by age to produce age-specific estimates of the IHRs for each state¹⁹.

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 leads to erroneous conclusions about trends in infections. More importantly, we observe outbreaks in infections that are missed by cases alone: for example, 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^{9,20}. 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^{21,22}. Unlike previous approaches that use a single delay distribution to generate estimates for all states^{23,24}, our work avoids this assumption of geographic invariance,

an assumption that is far from realistic due to differences in reporting pipelines, pandemic response, and variants in circulation, among other issues.

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 (Section 4.5), we more accurately estimate this ratio. However, future work could refine this analysis. Because the waning of immunity is likely to be variant-dependent²⁵, 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^{26–28}. 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^{29–32}, 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 unavailable³³. In follow-up work, we extend our construction of daily incident infections for some 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^{6,34}.

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 7 provides a visual summary of the data, analysis tasks, and the relationships between them. The major analysis tasks this figure aims to convey are as follows. First, we estimate variant-specific incubation periods and two types of delay distributions for each day over the time period under consideration. Next, each day-location-specific incubation period and symptom onset to positive specimen delay distribution are combined using convolution, resulting in variant-specific infection onset to positive specimen distributions for each time and location (for a brief description of convolution, please see Supplementary Methods subsection S1.1). Subsequently, two types of deconvolution are performed. We first deconvolve from case report to positive specimen date. We then deconvolve from positive specimen to report date by variant. The resulting infection 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.

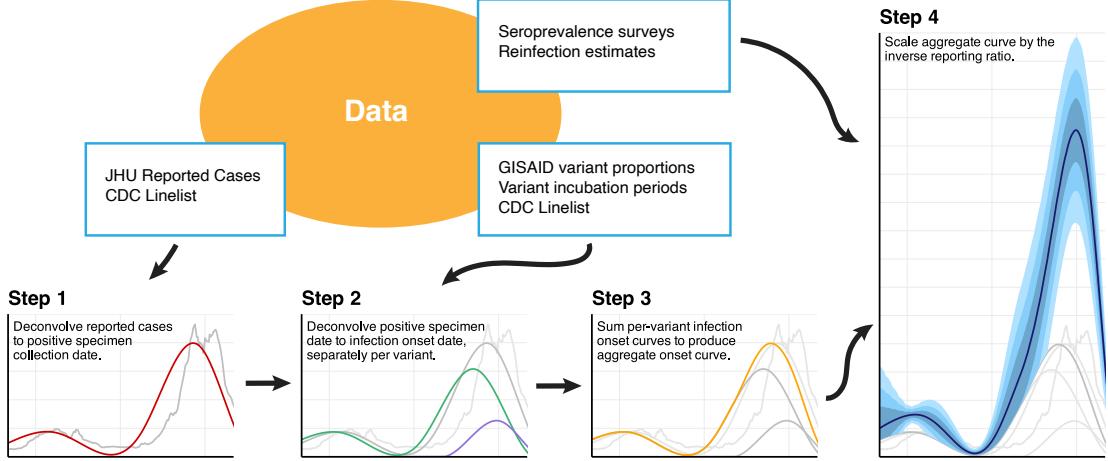


Figure 7: Flowchart of the data and major analysis steps required to get from reported cases to incident infection estimates.

4.1 Estimating delay distributions from private line lists

We obtain de-identified patient-level line list data on COVID-19 cases from the CDC³⁵. Although there are both public and restricted versions of the dataset available containing the same patient records^{35,36}, only the restricted dataset contains information on the state of residence. The three key dates of interest are those for symptom onset, positive specimen collection, and report to the CDC. Handling missingness and imputation in these dates is somewhat complicated and so additional details and justifications are deferred to subsection S1.2.

We use the line list to estimate the delay distribution for the pairs symptom onset to positive specimen and positive specimen to report. We provide the full procedure for the latter, before giving a brief description below for the former. First, define $z_{\ell,t}$ to be a case report occurring at time t in location ℓ , and let $\pi_{\ell,t}(k)$ to be the probability that $z_{\ell,t}$ has a positive specimen collected k days earlier. We assume that all positive specimens will be reported within 60 days and that no test will be reported on the same date as it was collected, that is, $\pi_{\ell,t}(0) = 0$ and $\pi_{\ell,t}(k) = 0$ whenever $k > 60$. Let $N_{\ell,t}$ be the number of $z_{\ell,s}$ with $s \in [t - 75 + 1, t + 60] = \mathcal{S}_t$ and positive specimen date greater than $s - 60$. Then, we first compute

$$\tilde{\pi}_{\ell,t}(k) = \frac{1}{N_{\ell,t}} \sum_{s \in \mathcal{S}_t} (\# z_{\ell,s} \text{ with positive specimen at } s - k). \quad (1)$$

We compute a similar national quantity $\tilde{p}_t(k) = \frac{1}{N_t} \sum_{s \in \mathcal{S}_t} (\# z_s \text{ with positive specimen at } s - k)$, 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 CSSE¹ in the same window. Then, compute $p_{\ell,t}(k) = \alpha_{\ell,t} \tilde{\pi}_{\ell,t}(k) + (1 - \alpha_{\ell,t}) \tilde{p}_t(k)$. This construction allows for more reliance on the state estimate when there are more CDC cases relative to JHU (and vice versa). We calculate the mean $m_{\ell,t}$ and variance $v_{\ell,t}$ of $\{p_{\ell,t}(k) : 0 < k \leq 60\}$ and estimate a 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 support set of 1 to 60 days to produce an estimate $\{\hat{\pi}_{\ell,t}(k) : 0 < k \leq 60\}$ of the delay distribution $\pi_{\ell,t}$.

Estimating the delay from symptom onset to positive specimen date follows the same procedure with a few minor adjustments. First, 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. This is reasonable because 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. Additional minor details are discussed in Supplementary Methods subsection S1.4. Moreover, we show both types of delays for a sample of states over several dates in Supplementary Methods subsection S1.5.

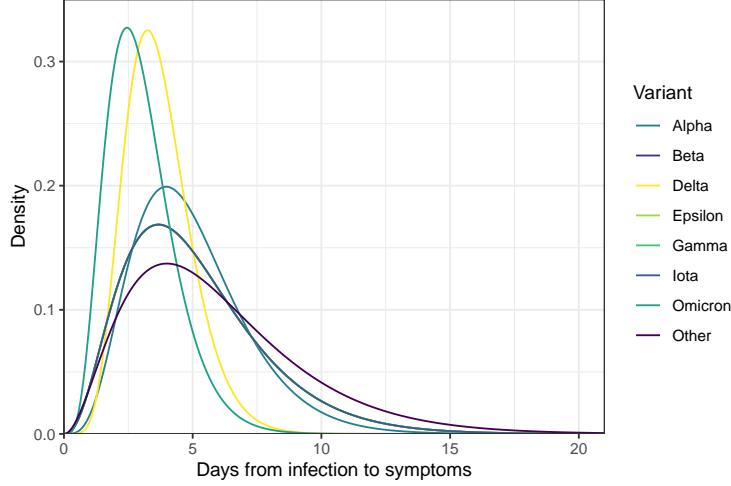


Figure 8: Gamma density for the incubation period of each of the eight variant categories. Note that the Ancestral variant directly utilizes the available gamma shape and scale parameters, while the remaining variants use the method of moments to estimate the gamma parameters.

4.2 Estimating the incubation period distributions

ATTN: Daniel, This section could be moved to the supp mat if we're looking for stuff to cut from the main body before introducing the models. It's up to you.

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, which dominated at various points during our study period: Ancestral/Other, Alpha, Beta, Epsilon, Iota, Gamma, Delta, and Omicron. Alpha, Beta, Delta, Gamma, and Omicron are all variants of concern³⁷, while we include the Epsilon (California) and Iota (New York) variants because of large impact on those and neighbouring states^{38,39}.

The Ancestral variant has been modelled as a gamma distribution⁴⁰, so we simply use the reported shape and scale parameters. For the Alpha, Beta, Gamma, Delta and Omicron variants, we use the reported mean and standard deviation of the number of days of incubation^{41–43}. To match these moments to the gamma distribution, we solve the same moment equations described in subsection 4.1. Then, we discretize each resulting density shown in Figure 8 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 develop⁴⁴ and that an asymptomatic infection will resolve within 21 days^{45,46}.

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 Clades¹⁵. 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 (which is one reason why we may refer to it as the Other category).

4.3 Variant circulation proportions

ATTN: Daniel, This section could also be moved to the supp mat if we're looking for stuff to cut from the main body before introducing the models.

To estimate the daily proportions of the variants circulating in each state, we obtain the GISAID genomic sequencing data from CoVariants.org^{15,47}. 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

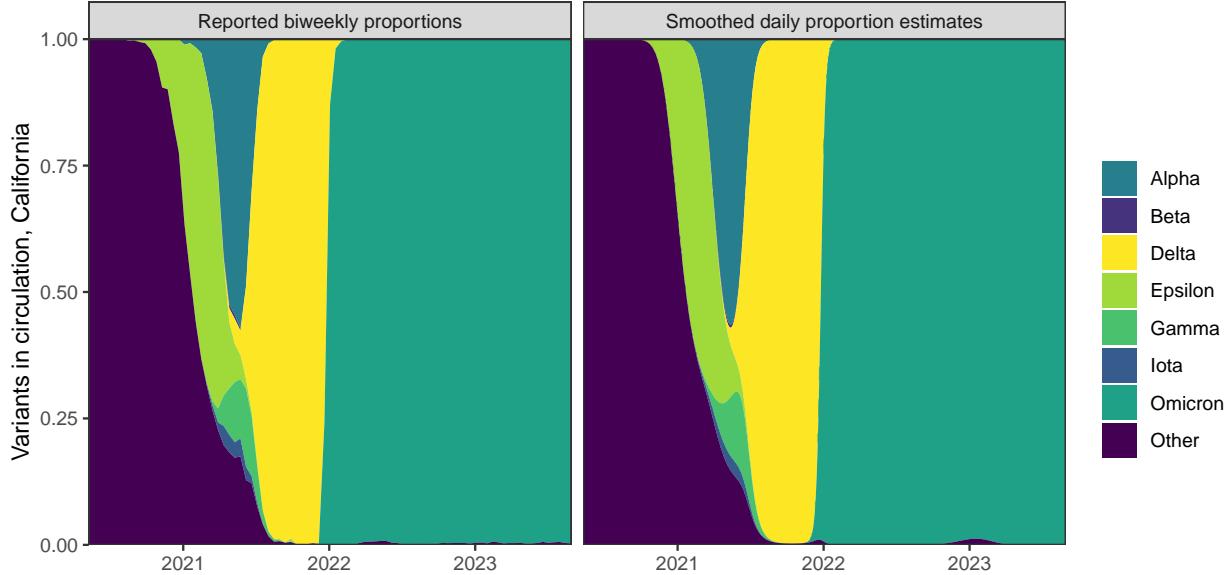


Figure 9: Left: Original biweekly proportions of the variants in circulation for California. Right: Daily proportions of the variants in circulation for California.

for each state. Multinomial logistic regression is a standard technique to model the frequency of SARS-CoV-2 variants^{48–50}.

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 the nonparametric multinomial logistic regression model is given as the system

$$\log \left(\frac{V_{j\ell,t}}{1 - V_{j\ell,t}} \right) = f_{j\ell}(t), \quad j = 1, \dots, J, \quad \text{subject to } \sum_{j=1}^J \exp\{f_{j\ell}(t)\} = 1, \quad \forall t. \quad (2)$$

The constraint ensures that the estimated proportions will sum to 1 across all J variants. To encourage smoothness of the estimated proportions, we specify $f_{j\ell}(t)$ as a third-order polynomial in time: that is $f_{j\ell}(t) = \beta_{j\ell,0} + \beta_{j\ell,1}t + \beta_{j\ell,2}t^2 + \beta_{j\ell,3}t^3$, computed such that the resulting matrix of covariates is orthogonal. Figure 9 shows the proportions by variant for California before (left) and after (right) the smoothing procedure.

4.4 Retrospective deconvolution: from cases to infections

Retrospective deconvolution estimates the daily number of new infections corresponding to each variant for each time and location, “pushing back” the dates that those cases were eventually reported to the time of infection. Because the circulating variant proportions in subsection 4.3 correspond to the positive specimen date, this requires two stages. The first is the deconvolution from report to positive specimen date, and the second is from positive specimen date to infection onset date.

An important aspect of our methods is that deconvolution is not the same as a simple shift. Shifting cases back in time and increasing them by some fixed factor fails to capture the temporal dynamics of the pandemic. Therefore, in our situation, we opt for deconvolution to “push back” the reported cases by using probability distributions of the delays shown in Figure 1.

We will start by describing the first type of deconvolution performed from report to positive specimen date in detail. For this problem, let $t = 1, \dots, T'$ index the extended deconvolution period from March 1, 2020 to March 1, 2023, extended to minimize the effects of boundary issues. **ATTN: Ryan, thoughts on this notation? The sum over t is awkward.** 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 API³⁴. Recall that $\widehat{\pi}_{\ell,t}(k)$ is the associated probability that these reported cases were collected k days earlier.

We estimate the deconvolved cases by positive specimen date by solving the following optimization problem:

$$\underset{\mathbf{x}}{\text{minimize}} \sum_{t=1}^{T'} \left(y_{\ell,t} - \sum_{k=1}^{60} \widehat{\pi}_{\ell,t}(k) x_{t-k} \right)^2 + \lambda \sum_{t=4}^{T'} |x_t - 4x_{t-1} + 6x_{t-2} - 4x_{t-3} + x_{t-4}|. \quad (3)$$

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 λ determines the relative importance of these competing goals. The solution to the problem is an adaptive piecewise cubic polynomial^{51,52} and can be accurately computed with ease⁵³. We select λ with 3-fold cross validation²⁴ in which every third day is reserved for testing, and the value that results in the smallest out-of-sample mean squared error is chosen.

The result of this first deconvolution is $\widehat{x}_{\ell,t}$, case estimates by positive specimen date for each state. To continue, pushing back to infection estimates, we first need the variant-specific delays from infection to positive specimen collection. These are calculated by convolving the location-time-specific symptom-to-test distributions from subsection 4.1, denoted by $\{q_{\ell,t}(k) : -3 \leq k \leq 21\}$, with the variant-specific incubation periods from subsection 4.2, denoted by $\{i_j(k) : 0 < k \leq 21\}$. The convolution of these yields a distribution $\mathbf{q}_{\ell,t} * \mathbf{i}_j = \{\tau_{j\ell,t}(k) : -3 \leq k \leq 42\}$. However, only a fraction of $\widehat{x}_{\ell,t}$ corresponds to each variant, so we must weight them by the variant proportions $\widehat{v}_{j\ell,t}$ estimated in subsection 4.3. The analogous optimization problem is therefore:

$$\underset{\mathbf{u}}{\text{minimize}} \sum_{t=1}^{T'} \left(\widehat{v}_{j\ell,t} \widehat{x}_{\ell,t} - \sum_{k=-3}^{42} \tau_{j\ell,t}(k) u_{t-k} \right)^2 + \lambda \sum_{t=4}^{T'} |u_t - 4u_{t-1} + 6u_{t-2} - 4u_{t-3} + u_{t-4}|. \quad (4)$$

We call the solution $\widetilde{\mathbf{u}}_{j\ell}$ the *variant-specific deconvolved cases* and emphasize that these are those cases that will eventually be reported to public health. Because this deconvolution is done separately for each location and variant category, we ultimately obtain deconvolved case estimates by the date of infection onset that are separated by variant. Finally, we denote the total deconvolved cases at location ℓ as $\widehat{\mathbf{u}}_\ell = \sum_j \widetilde{\mathbf{u}}_{j\ell}$.

4.5 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.

Because seroprevalence of anti-nucleocapsid antibodies represents the percentage of people who have at least one resolving or past infection⁵⁴, we can use the change in subsequent seroprevalence measurements to estimate *all* new infections, rather than just those eventually appearing as cases. This intuition suggests modelling reported seroprevalence at time $t+1$ as a fraction $1-\gamma$ of the previous seroprevalence measurement at t plus the reinfection-adjusted deconvolved cases multiplied by the inverse reporting ratio at time t :

$$s_{\ell,t+1} = (1 - \gamma_\ell) s_t + a_{\ell,t} (1 - z_{\ell,t}) \widehat{u}_{\ell,t} + \epsilon_{\ell,t}, \quad (5)$$

where $\widehat{u}_{\ell,t}$ is deconvolved cases, $z_{\ell,t}$ is the fraction of reinfections, $a_{\ell,t}$ is the inverse reporting ratio, and $\epsilon_{\ell,t}$ represents noise. Note that γ_ℓ is the fraction of people whose level of infection-induced antibodies falls below the detection threshold between time t and time $t+1$. Informally, we refer to γ as the waning parameter. Unfortunately, population seroprevalence is not directly observed, but is estimated through various surveys that target different subsets of the population.

We use two major seroprevalence surveys: the 2020–2021 Blood Donor Seroprevalence Survey and the Nationwide Commercial Lab Seroprevalence Survey^{55,56} to estimate the proportion of the population with evidence of previous infection in each state over time. See Supplementary Methods subsection S1.6 for additional details. Each of these provides seroprevalence estimates along with confidence intervals. The daily fraction of new infections are based on surveillance work conducted by the Southern Nevada Health District²⁹. These results are broadly similar to those in other locations with available data^{29–32}.

In order to account for different surveys occurring on different dates with noisy estimates, we estimate the model on the weekly frequency, observed on Monday, and treat $s_{\ell,t}$ as a latent variable. Therefore, we write,

$$r_{\ell,m}^1 = s_{\ell,m} + \tau_{\ell,m}, \quad \tau_{\ell,m} \sim N(0, w_{\ell,m}^1 \sigma_{\ell,r}^2), \quad (6)$$

$$r_{\ell,m}^2 = s_{\ell,m} + \varphi_{\ell,m}, \quad \varphi_{\ell,m} \sim N(0, w_{\ell,m}^2 \sigma_{\ell,r}^2), \quad (7)$$

$$s_{\ell,m+1} = (1 - \gamma_{\ell})s_{\ell,m} + a_{\ell,m}(1 - z_{\ell,m})\hat{u}_{\ell,m}^{\Sigma} + \epsilon_{\ell,m}, \quad \epsilon_{\ell,m} \sim N(0, \sigma_{\ell,\epsilon}^2), \quad (8)$$

where r^1 and r^2 correspond to the two different seroprevalence surveys. These surveys each have measurement errors with variance σ_r^2 that scale proportional to the observed confidence intervals for the estimates, respectively $w_{\ell,m}^1$ and $w_{\ell,m}^2$. We denote $\hat{u}_{\ell,m}^{\Sigma} = \sum_{t=m}^{m+1} \hat{u}_{\ell,t}$. Finally, to ensure that \mathbf{a}_{ℓ} is smooth over time, we complete the model with an additional equation that enforces smoothness,

$$a_{\ell,m+1} = 3a_{\ell,m} - 3a_{\ell,m-1} + a_{\ell,m-2} + \eta_{\ell,m}, \quad \eta_{\ell,m} \sim N(0, \sigma_{\eta}^2). \quad (9)$$

This antibody prevalence model is a state-space model with latent variables \mathbf{s}_{ℓ} and \mathbf{a}_{ℓ} and unknown parameters γ_{ℓ} , σ_r^2 , σ_{ϵ}^2 , and σ_{η}^2 . This model allows for convenient handling of missing data, extrapolation before and after the period of observed seroprevalence measurements, and maximum likelihood estimates of the errors. Details of this methodology and the computation of the associated uncertainty measurements are deferred to Supplementary Methods subsection S1.7.

4.6 Lagged correlation to hospitalizations and time-varying IHRs

From the COVIDcast API³⁴, 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. 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. 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 (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] 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).
- [26] 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).
- [27] Pulliam, J. R. *et al.* Increased risk of sars-cov-2 reinfection associated with emergence of omicron in south africa. *Science* **376**, eabn4947 (2022).
- [28] 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).
- [29] Ruff, J. *et al.* Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. *Emerging Infectious Diseases* **28**, 1977 (2022).
- [30] New York State COVID-19 reinfection data. <https://coronavirus.health.ny.gov/covid-19-reinfection-data> (2021).
- [31] Hawaii Department of Health COVID-19 reinfection data. https://health.hawaii.gov/coronavirusedisease2019/files/2022/09/reinfection_report_2022-09-28.pdf (2022).
- [32] Reported COVID-19 reinfections in Washington State. <https://doh.wa.gov/sites/default/files/2022-02/421-024-ReportedReinfections.pdf> (2022).
- [33] McManus, O. *et al.* Predicting COVID-19 incidence using wastewater surveillance data, Denmark, October 2021–June 2022. *Emerging Infectious Diseases* **29**, 1589 (2023).
- [34] Reinhart, A. *et al.* An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences* **118**, e2111452118 (2021).
- [35] 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).
- [36] 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).
- [37] World Health Organization. Tracking SARS-CoV-2 variants. <https://www.who.int/activities/tracking-SARS-CoV-2-variants> (2021).

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

- [56] 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).
- [57] 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).
- [58] 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).
- [59] Bajema, K. L. *et al.* Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. *JAMA Internal Medicine* **181**, 450–460 (2021).
- [60] Durbin, J. & Koopman, S. J. *Time Series Analysis by State Space Methods*, vol. 38 (OUP Oxford, 2012).
- [61] Helske, J. KFAS: Exponential family state space models in R. *Journal of Statistical Software* **78**, 1–39 (2017).

Acknowledgements

The authors are thankful to the Delphi team for their continuous support for this study. 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 Initiative⁴⁷, on which this research is based. RL is supported by the National Sciences and Engineering Research Council of Canada and the University of British Columbia.

ATTN: Daniel add the other Grants etc.

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.

Online Supplement

S1 Supplementary Methods

This section contains additional information about the datasets used and the estimation 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 distribution of probabilities. In the 2D and discrete context, it is simply the elementwise multiplication of the signal for some time by a forward-facing distribution of probabilities, which are then summed to get a value for the outcome. Figure S1 presents a depiction of the convolution procedure for the signal of smoothed cases (orange line). Essentially, to push the cases forward in time, we take the appropriately aligned (forward-in-time) delay distribution and convolve or multiply the case counts by it to get the distribution of convolved case 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 points, but rather it utilizes the most relevant probabilities to propagate the data points forward in time. Deconvolution proceeds in the same fashion, but in the opposite direction to go back in time.

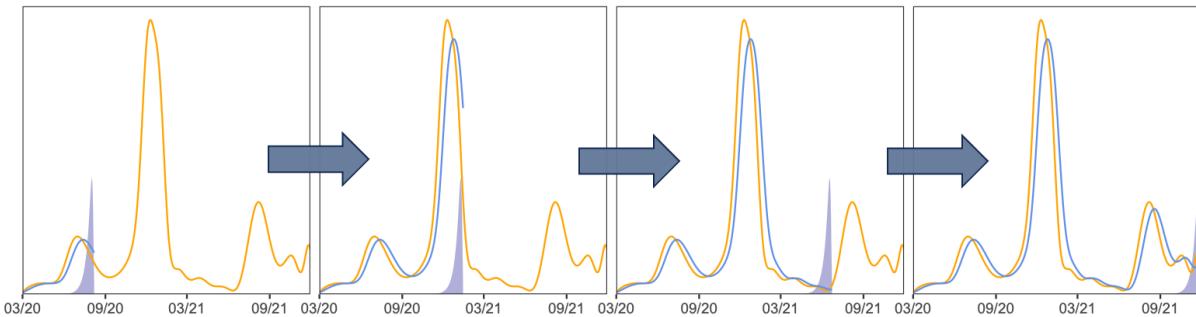


Figure S1: A general depiction of convolving smoothed cases (orange line) with the corresponding delay probabilities (shaded blue area) to get the convolved estimates (blue line) over four different times.

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

Since the restricted dataset 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 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.

We observe that the line list is prone to high percentages of 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. Relatedly, cases with missing report or positive specimen dates may be filled with their symptom onset date²⁴. So it is possible that all three variables may be imputed with the same date for a case. However, we only actually deal with select pairs of events; we do not use all three at once in our construction of the delay distributions or anywhere else in our analysis. Therefore, we restrict our investigation of missingness to the pairs of events. Figure S2 suggests that this issue impacts states differentially due to the inconsistent proportions of zero delay between positive specimen and report date across states.

Due to the contamination in the zero delay cases (the true extent of which is unknown to us), we omit all such cases where the positive specimen and report dates have zero delay from our analysis. We choose to

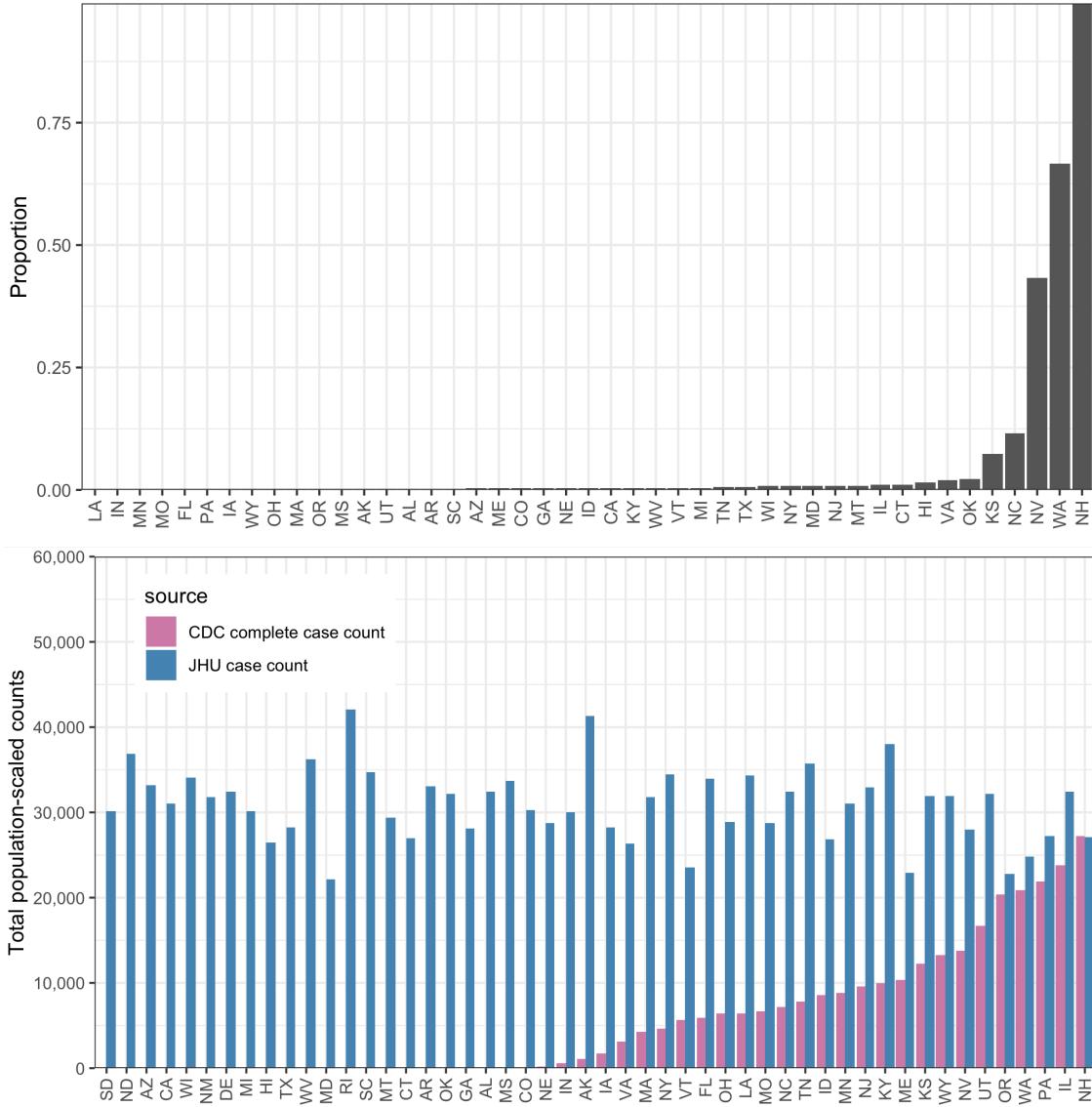


Figure S2: Top panel: Proportion of complete cases with zero delay between positive specimen and report date in the restricted CDC line list dataset. Bottom panel: Complete case counts by state in the CDC line list versus the cumulative complete case counts from JHU CSSE as of June 6, 2022. All counts have been scaled by the 2022 state populations as of July 1, 2022 from⁵⁷.

allow for zero and negative delay for symptom onset to report because correspondence with the CDC confirms the distinct 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.

For the same release date, the 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%²⁴. 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.

In the line list, we observe unusual jarring spikes in reporting in 2020 compared to 2021. Upon plotting by report date, we find that a few states are contributing unusually large case counts on isolated days very late

in the reporting process (usually well beyond 50 days). We strongly suspect that these large accumulations of cases over time are due breakdowns of the reporting pipeline (which may be expected to occur more frequently in the year following its instantiation than later in time). Such anomalies are not likely to be reliable indicators of the delay from positive specimen to case report. Therefore, we devise a simple, ad hoc approach to detect and prune these reporting backlogs.

First, we obtain the part of the line list intended for the positive specimen to case report delay estimation, where both such dates are present and where zero and negative delay cases have been omitted. Then, for each of the three dates of June 1, September 1, and December 1, 2020, we bin the reporting delays occurring from 50 days up to the maximum observed delay. For each bin, we obtain the total delay count for each state. We check whether each count on the log scale is at least the median (for the bin) plus 1.5 times the interquartile range and retain only those that exceed this criterion as potential candidates for pruning. Next, we compute the counts by report date for each candidate state. If there is a report date with a count greater than or equal to the pre-specified threshold, then we remove those cases from the line list. Based on inspection and intuition, we set the threshold to 2000 for the first two bins, and then lower it to 500 for the remaining bins. A similar trial and error approach is used to set the bin size (to 50 days).

S1.3 Table on the percent pairwise occurrence of events in the CDC line list

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 we built the current 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	Nothing because current handling of 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).

S1.4 Justifications for delay distribution calculations

Let y_t denote the count of new cases reported at time t and x_t denote the count of deconvolved cases with positive specimen at t . For all cases in the line list that had both a positive specimen and a report date, we can count the those that are reported at time t by enumerating them according to positive specimen date (similar to how symptom onset date was used in ²⁴):

$$y_t = \sum_{s=1}^t \sum_{i=1}^{x_s} \mathbf{1} (\text{the } i^{\text{th}} \text{ positive specimen at } s \text{ gets reported at } t).$$

Taking the conditional expectation of the above yields

$$\mathbb{E}(y_t \mid x_s, s \leq t) = \sum_{s=1}^t \pi_t(s)x_s,$$

where $\pi_t(s) = \mathbb{P}(\text{case report at } t \mid \text{positive specimen date at } s)$ for each $s \leq t$ are the delay probabilities and the $\{\pi_t(s) : s \leq t\}$ sequence comprises the delay distribution at time t . Notice that there are no time restrictions placed on the positive specimen date, except that it must have been between the start of the pandemic and the report date, inclusive. This is unlikely to be a realistic assumption to make as t moves farther away from s .

Thus, we make three key assumptions about these distributions. First, positive specimen tests that are reported to the CDC are always reported within $d = 60$ days, which is true for the majority of the reported cases. Second, the probability of zero delay is zero, which stems from the contamination of zero-delay in the line list. We update the conditional expectation formula to reflect these two assumptions as in ²⁴:

$$\mathbb{E}(y_t \mid x_s, s \leq t) = \sum_{k=1}^{60} p_t(k)x_{t-k}$$

where for $k = 1, \dots, 60$,

$$p_t(k) = \mathbb{P}(\text{case report at } t \mid \text{positive specimen at } t - k).$$

Thirdly, there are times where the empirical probability was observed to be precisely 1 at zero delay and the proportion of CDC relative to JHU cases used for the weight was also 1. Since we believe that having zero delay for all cases is unrealistic and unlikely to be representative of all cases for the state, we inject a small amount of variance manually by setting the the CDC-to-JHU proportion to be the minimum shrinkage proportion observed for the affected state (such instances were isolated to the state of New Hampshire). Aside from these modifications, the construction of the delay distribution proceeds in precisely the same manner as for positive specimen to report date.

S1.5 Estimated delays for a sample of six states

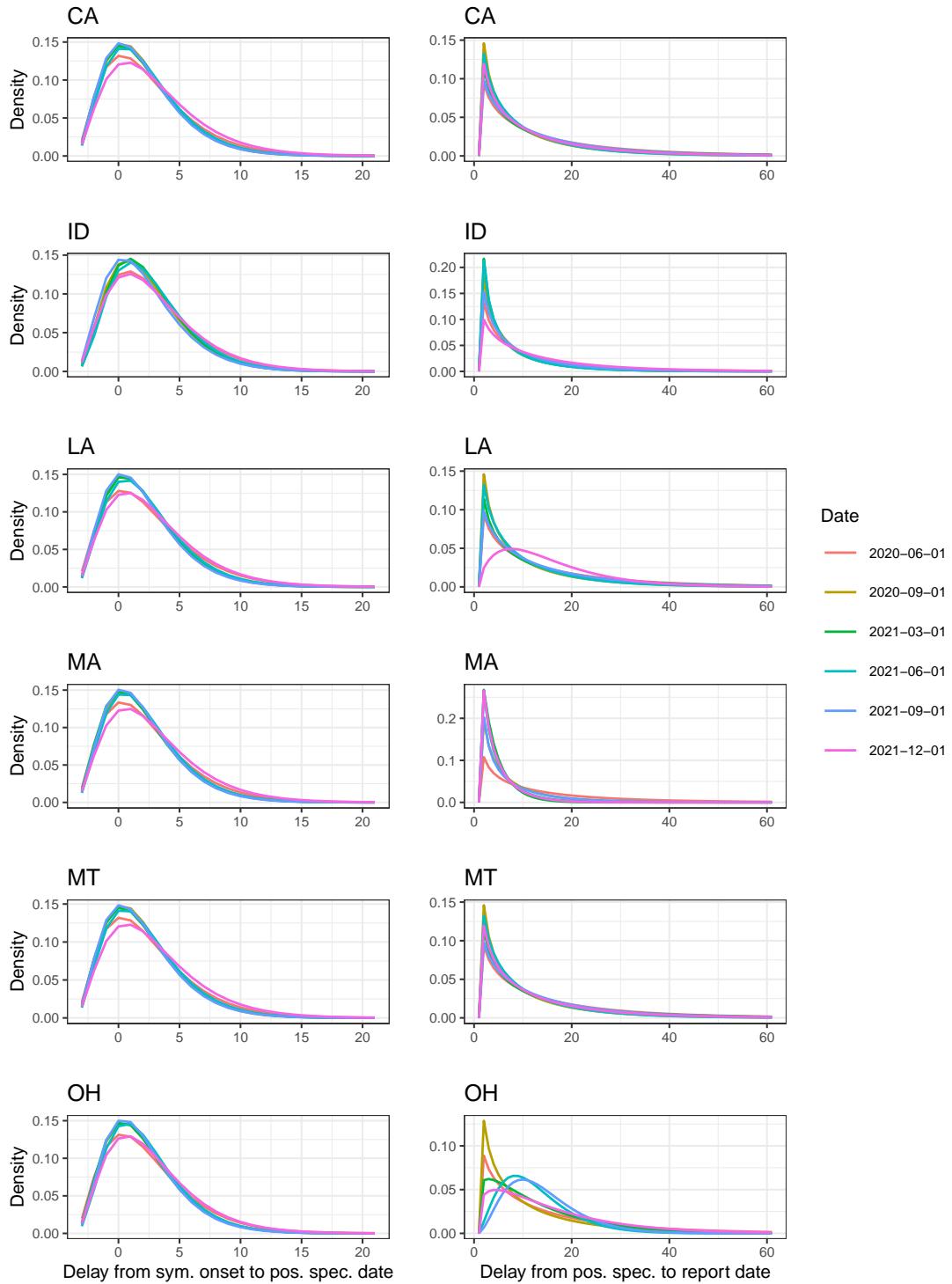


Figure S3: Depictions of the estimated discretized delay from symptom onset to positive specimen date (left) and from positive specimen date to report date (right) for a sample of six states.

S1.6 Details about seroprevalence data

In the former, the CDC collaborated with 17 blood collection organizations in the largest nationwide COVID-19 seroprevalence survey to date⁵⁵. The blood donation samples were used to construct monthly seroprevalence estimates for nearly all states from July 2020 to December 2021⁵⁸. In the latter survey, the CDC collaborated with two private commercial laboratories and used blood samples to test for the antibodies to the virus from people that were in for routine or clinical management (presumably unrelated to COVID-19⁵⁹). 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 aimed, at least in part, to 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^{54,58,59}. 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 S4). Such limitations indicate that reliance upon only one seroprevalence survey is inadvisable. 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.

The date variables that come with the two seroprevalence datasets are different and so the date variables that we are able to construct from them are not the same. For the commercial dataset, we use the midpoint of the provided specimen collection date variable. 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 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.

To adapt to the sparseness in the seroprevalence data, we convert our daily data to weekly by summing the reported infections and shifting the observed seroprevalence measurements to the nearest Monday. If there are multiple measurements in a week from a seroprevalence source, then the average is used. We denote these changes by changing the time-based subscript from t to m where m indicates the Monday relative to our June 1, 2020 start date. Since we operate with weekly data where the weeks are designated by Monday, we set the end date to be November 29, 2021.

S1.7 State space representation of the antibody prevalence model

The antibody prevalence model is conceptualized as a Gaussian state space model (as in^{60,61}).

In general, for $t = 1, \dots, n$, let α_t be the $m \times 1$ vector of latent state processes at time t and y_t be the $p \times 1$ vector of observations at time t . Under the assumption that η is a $k \times 1$ vector, the form of the linear Gaussian state space model is

$$y_t = Z\alpha_t + \epsilon_t, \quad \epsilon_t \sim N(0, H_t) \quad (10)$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim N(0, Q_t) \quad (11)$$

where $\alpha_1 \sim N(a_1, P_1)$ and there is independence amongst α_1 , ϵ_t and η_t ^{60,61}. For notational compactness, we let $\alpha = (\alpha_1^\top, \dots, \alpha_n^\top)^\top$ and $y = (y_1^\top, \dots, y_n^\top)^\top$.

The observation equation can be viewed as a linear regression model with the time-varying coefficient α_t , while the second equation is a first-order autoregressive model, which is Markovian in nature⁶⁰.

The underlying idea behind the two equations is that we are assuming that the system evolves according to α_t (as in the second equation), but since those states are not directly observed, we turn to the observations y_t and use their relationship with α_t (as in the first equation) to drive the system forward⁶⁰. So the objective of state space modeling is to obtain the latent states α based on the observations y and this is achieved through Kalman filtering and smoothing.

Kalman filtering gives the following one-step-ahead predictions of the states

$$a_{t+1} = \mathbb{E}[\alpha_{t+1} | y_t, \dots, y_1]$$

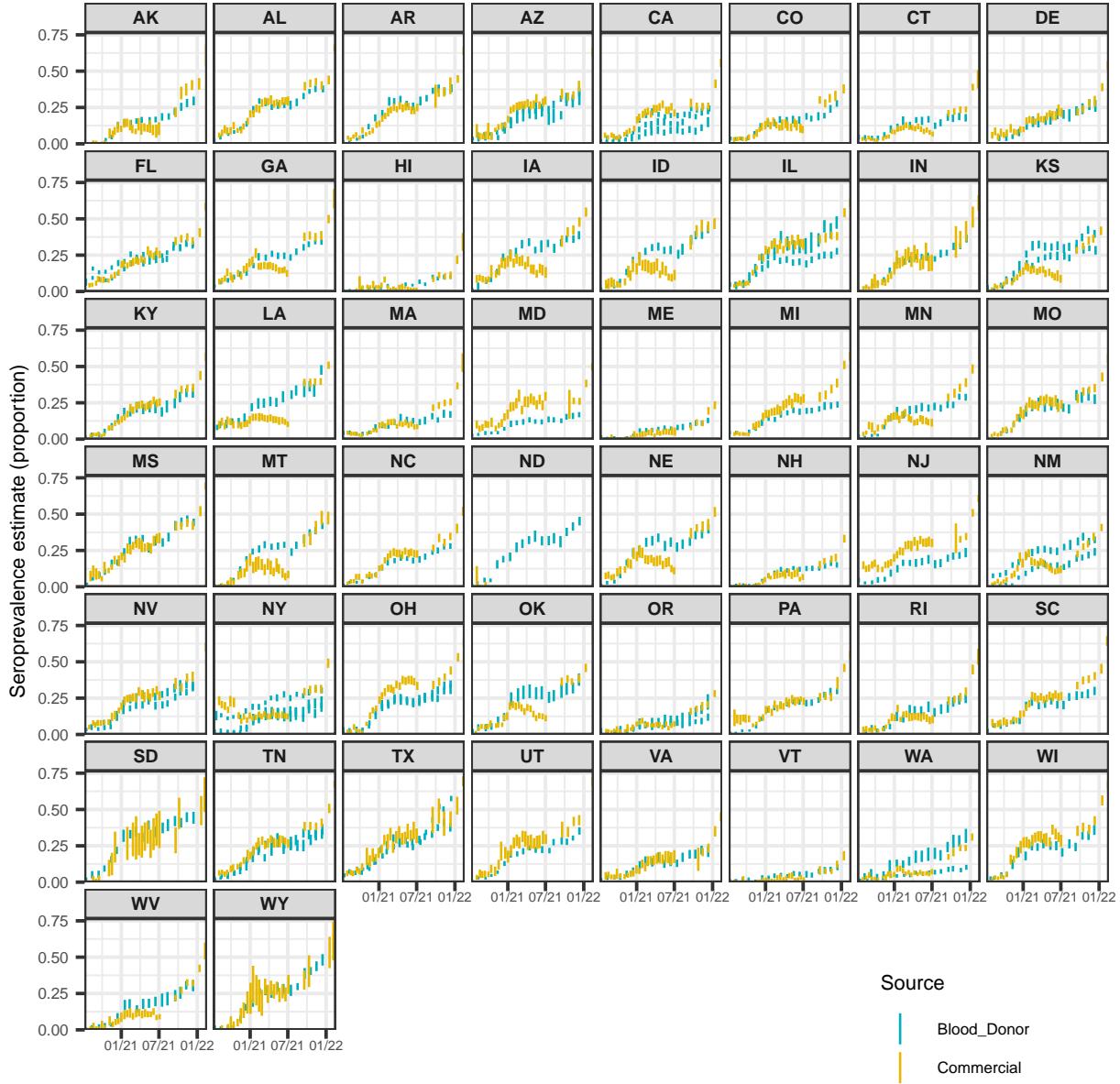


Figure S4: A comparison of the seroprevalence estimates from the Commercial Lab Seroprevalence Survey dataset (yellow) and the 2020–2021 Blood Donor Seroprevalence Survey dataset (blue). Note that the maximum and the minimum of the line ranges are the provided 95% confidence interval bounds to give a rough indication of uncertainty.

with covariance,

$$P_{t+1} = \text{Var}(\alpha_{t+1} | y_t, \dots, y_1).$$

Then, the Kalman smoother works backwards to the first time to give

$$\hat{a}_t = \mathbb{E}[\alpha_t | y_n, \dots, y_1] \quad (12)$$

$$V_t = \text{Var}(\alpha_t | y_n, \dots, y_1). \quad (13)$$

The filtering and smoothing steps are based on recursions that are described in Appendix A of⁶¹ as we use the R package KFAS to estimate our model.

To express the antibody prevalence model in state space form, we define the components in Equations 10 and 11 as follows:

$$\begin{aligned}
R &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} & Z &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} & H_m &= \begin{bmatrix} w_{m,c}\sigma_o^2 & 0 \\ 0 & w_{m,b}\sigma_o^2 \end{bmatrix} \\
\alpha_m &= \begin{bmatrix} s_m \\ a_m \\ a_{m-1} \\ a_{m-2} \end{bmatrix} & T_m &= \begin{bmatrix} \gamma & C_{m-1}^m z_m & 0 & 0 \\ 0 & 3 & -3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} & Q &= \begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_a^2 \end{bmatrix} \\
a_1 &= \begin{bmatrix} \tilde{s}_1 \\ \tilde{a}_1 \\ \tilde{a}_1 \\ \tilde{a}_1 \end{bmatrix} & P_1 &= \begin{bmatrix} \sigma_{\tilde{s}_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\tilde{a}_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\tilde{a}_1}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\tilde{a}_1}^2 \end{bmatrix}
\end{aligned}$$

where σ_o^2 is the variance of observations, σ_s^2 is the variance of the seroprevalence estimates, and σ_a^2 is the trend variance. Since we expect the inverse ratios to be more variable than the seroprevalence estimates, we enforce that the estimate of σ_a^2 is a multiple of σ_s^2 . Letting the subscripts b and c denote the blood donor and commercial datasets, $w_{m,c}$ and $w_{m,b}$ 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 standard formula for the standard error of a proportion. These weights are then re-scaled so they sum to the number of observed seroprevalence measurements for the source. All days that are unobserved (i.e., lack seroprevalence measurements) are given weights of one. Finally, the ratio of the average observed weights for the sources is used as a multiplier to scale all of the weights for one source. For example, if the average weight of the commercial source is double the average weight of the blood donor source (for an arbitrary state), then we scale all of the weights in the commercial source (including the ones) by two. The main purpose of this step is to ensure that the source with a greater sample size contributes more weight in the model on average.

The prior distribution for α_1 is estimated using both data-driven constraints and externally sourced information. To obtain the initial value of the seroprevalence component, \tilde{s}_1 , we extract the first observed seroprevalence measurement from each source, round down to two decimal places, and take the average to be \tilde{s}_1 . 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 all of the initial values of the trend components, we use the inverse of the ascertainment ratio estimate as of June 1, 2020 for each state from Table 1 in ²¹ and denote this by \tilde{a}_1 . The initial variance estimate of $\sigma_{\tilde{a}_1}^2$ is based on the variance implied by the given inverse ascertainment ratio distribution.

The initial σ_o^2 is taken to be the average of the estimated variances from the linear models for the sources where the observed seroprevalence measurements are regressed on the enumerated dates. The initial value of the multiplier is set to be 100 for all states. The σ_s^2 and γ values are fixed and from averaging the estimated values for all states on the real line (obtained under the starting conditions $\sigma_s^2 = 3 \times 10^{-6}$, $\gamma = 0.99$, and σ_o^2 as described).

Following the maximum likelihood estimation of the two non-fixed parameters we use the Kalman filtering and smoothing to obtain the smoothed estimates of the weekly inverse reporting ratios and their covariance matrices as shown in Equations 12 and 13. Forwards and backwards extrapolation is then used to estimate the ratios and covariance outside of the observed seroprevalence range ⁶⁰, followed by linear interpolation to fill-in estimates for each day in our considered time period. After we obtain one vector of inverse reporting ratios for each state in this way, we take each inverse reporting ratio and multiply it by the corresponding deconvolved case estimate (that has undergone linear interpolation to correct instances of 0 reported infections) to obtain an estimate of new infections. We are able to convert these numbers of infections to infections per 100,000 population by simple re-scaling (enabled by the fact that normality is preserved under linear transformations).

The 50, 80, and 95% confidence intervals are constructed by taking a Bayesian view of the antibody prevalence model (refer to S1.8 for the Bayesian specification of the model). That is, for each time, t , we

obtain an estimate of the posterior variance of a_t , apply the deconvolved case estimate as a constant multiplier, and then use resulting variance to build a normal confidence interval about the infection estimate. We additionally enforce that the lower bound must be at least the deconvolved case estimate for the time under consideration.

S1.8 Bayesian specification of the antibody prevalence model

In brief, the antibody prevalence model where we let $\beta = \{\gamma, a_1, \dots, a_t\}$ and X be the design matrix, corresponds to a Bayesian model with prior

$$\beta \sim N \left(0, \frac{\sigma^2}{\lambda} (A^T D^T D A)^{-1} \right)$$

and likelihood

$$s | X, \beta \sim N(X\beta, \sigma^2 W^{-1}),$$

where A is indicator matrix save for the first column of 0s (corresponding to γ), D represents the discrete derivative matrix of order 3, and W is the inverse variance weights matrix. Then, the posterior on a_t is normally distributed with mean

$$(X^T W X + \lambda A^T D^T D A)^{-1} X^T W s$$

and variance

$$\sigma^2 (X^T W X + \lambda A^T D^T D A)^{-1}.$$

S1.9 Scaling by population

Annual estimates of the resident state populations as of July 1 of 2020 and 2021 are taken from the Dec. 2022 press release from the U.S. Census Bureau⁵⁷. Unless otherwise specified, we use the July 1, 2020 estimates.

S1.10 Ablation analysis of infection-hospitalization correlations

To better understand the contribution of the intermediate steps to the lagged correlation analysis, we carry out a brief ablation study in which we calculate the lagged correlation 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 S5](#). 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.

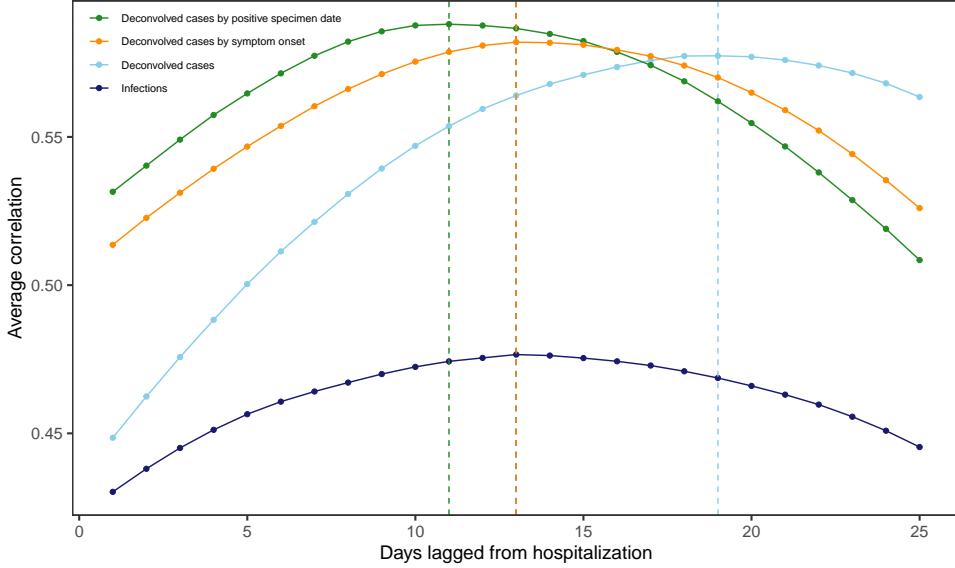


Figure S5: Lagged Spearman's correlation between the infection and hospitalization rates per 100,000 averaged for each lag across U.S. states and days over June 1, 2020 to November 29, 2021, and taken over a rolling window of 61 days. The infection rates are based on the counts for the deconvolved case 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.