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

Rachel Lobay, Maria Jahja, Ajitesh Srivastava, Ryan J. Tibshirani, Daniel J. McDonald

Version: March 26, 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 data-driven and methodological reasons. 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 estimate daily incident *infections* for each U.S. state. Rather than taking a model-based approach, our methods operate directly on data. We first deconvolve reported COVID-19 cases to their infection date using delay distributions estimated from the CDC linelist. We combine these deconvolved cases with serology data to scale up to 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. We also discuss some implications of our results: a disease burden that appears earlier and more extensively than previously quantified; differential infection-hospitalization ratio estimates. Our findings help to better understand the impact 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, likely much earlier. To see why, it is important to understand *whose* cases are being reported and what differentiates them from the 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 [4]. In addition, testing practices, availability, and uptake vary temporally and spatially [5–7]. Finally, cases provide a belated view of the pandemic's progression, because they are subject to delays due to the viral incubation period, the speed and severity of symptom onset, laboratory confirmation, test turnaround times, and eventual submission to public health [8;9]. For these reasons, reported cases are 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 as indicated by 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.

Explaining the course of the pandemic and investigating the effects of interventions, the burden facing various subgroups, and drawing insights for future pandemics is inhibited because the true 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 understates the true size of the pandemic. Regardless of these difficulties, it is important to the public and public health to perform a pandemic post-mortem and try to better explain its implications—to attempt to capture the true size and impact of the pandemic as much as we can. Estimates of daily incident infections are one such way to measure this and can guide understanding of the pandemic burden over space and time.

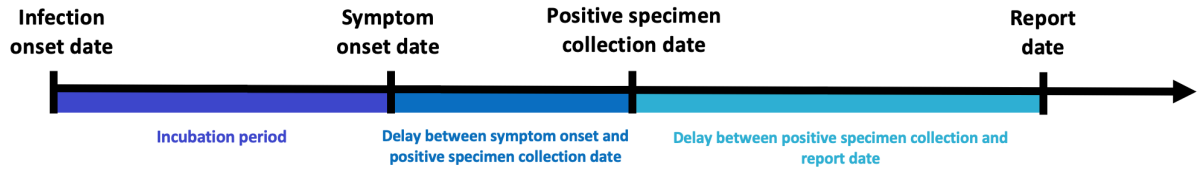

Figure 1: Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

In this work, we provide a statistically rigorous, data-first 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. Finally, the resulting deconvolved cases are adjusted to account for the unreported infections using seroprevalence and reinfection data to estimate adjust for the waning of antibody detectability over time. We examine some features of our infection estimates and the implications of using them rather than reported cases in assessing the impact of the pandemic. 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 the resulting publicly-available estimates) will prove an important benchmark for others to undertake retrospective analyses.

2 Results

2.1 Disease burden and viral transmission

From reconstructing the time series of COVID-19 infections per $100,000$ population 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.

Most states present at least two major spikes in infections - the first starts in the fall of 2020 and extends into the winter season, while the second starts in the late summer of 2021 and proceeds into the mid-fall. These represent major waves driven by the Ancestral and Delta variants. Similar patterns in the major surges of infections are observed in nearly all states, though to varying degrees. In general, greater similarities in the strength and magnitude of outbreaks are found to emerge in the clusters of states that border each other.

To avoid encroaching upon possible boundary issues with ending the estimation during a time of volatility (the period of the Delta-Omicron transition), we focus on the infection estimates prior to November 1, 2021. The largest observed outbreaks prior to this time were observed in the late summer or early fall of 2021 in Georgia, Louisiana, Idaho, Montana, and Wyoming which suggests a similar spread of the virus in small clusters of states that are in close geographic proximity. During this time, the two states that have the attain the highest rate of infections per 100,000 on single day are Georgia with about 451 infections per 100,000 on August 15, 2021 (95% confidence interval: $[334, 567]$) and Idaho with 451 on September 7, 2021 (95% confidence interval: $[312, 590]$). Prior to the Delta wave, the state that has the highest rate of infections per 100,000 on single day is Louisiana with about 358 infections per 100,000 on July 3, 2021 (95% confidence interval: $[177, 539]$) followed by Wyoming with 349 on November 13, 2020 (95% confidence interval: $[407, 546]$).

The period of lowest viral transmission is observed in the summer and fall of 2020. During this time, the state of New Hampshire achieves the lowest weekly rate of infections of 0.01 infections per 100,000 for the week of September 13, 2020. In the summer of 2020, Vermont maintains a rate under 10 infections per 100,000 from the week of June 1, 2020 to August 30, 2020, which is the longest continuous stretch observed for any state.

From a brief inspection of the geo-contiguous states, we can observe similar patterns in surges and periods of waning over time, suggesting that states who share similarities in climate and topography performed

similarly to each other. More precisely, we can observe neighboring states such as New Hampshire and Massachusetts or Idaho and Montana that present waves that mirror each other in amplitude and timing.

Interestingly, the two states that are geographically removed from the contiguous United States, Alaska and Hawaii, tend to perform quite differently from each other later in the pandemic. Alaska generally presents significantly greater rates of infections than Hawaii especially during the Delta era. This suggests that it is not so much the non-contiguity aspect as it is other distinguishing factors that lead to lower infection rates.

2.2 Infection estimates reveal waves missed by reported cases

In addition to the depicting the daily new infections per 100,000 population, Figure 2 compares these with the new reported cases. In general, the outbreaks in infections precede those in reported cases and are reliably larger in magnitude.

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 present in the spring of 2021 for North Dakota and South Dakota which is not visible in reported cases alone. For the specific date of Oct. 20, 2020, Figure 3 shows heightened case rates in the northern states like North Dakota, South Dakota, and Wisconsin relative to other states. However, the infection rates more clearly shows that the influx in infections is not limited to these states, but extends out to the surrounding states.

2.3 Underreporting of infections 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 across states and variants. For the major Delta wave, some of the greatest discrepancies between cases and infections are visible in the western states of Idaho and Montana, the southern states of Louisiana and Georgia, and the midwestern states of Iowa and Nebraska. In addition, we can see that the major Delta wave is only faintly detectable from cases in a number of eastern states such as Maryland and Connecticut. Now, moving to the Ancestral wave, we can see that this wave is poorly represented by cases in several midwestern states (take, for example, Illinois, Indiana, and Ohio). Earlier on in the pandemic, such discrepancies between cases and infections may be more attributable to failures in the reporting pipeline, while later on in the pandemic, they more likely due to the rise in asymptomatic infections across variants[10;11].

Infection underreporting appears to become considerably worse with time. While the main Delta wave is somewhat apparent from the case counts for all states (Figure 2), our infection estimates suggest that case counts tend to severely underestimate infections during this time for many states. The lowest of all states was New Jersey, where about 4.6% of the estimated infections were reported. This was followed by Maryland with 7.4%, Connecticut with 8.0%, and Florida with 8.7%. This underreporting issue extends to most states as in 39 states less than 30% of infections were reported during this time. Similar patterns were observed during the earlier period of Alpha domination, where Louisiana had the lowest percent of infections that were reported at 11.7% and was followed by California at 14.4%. Such patterns were comparatively less apparent during the earlier and larger period of Ancestral domination, where Ohio and Maryland held the lowest percentages of reported infections at 22.0% and 22.3%, respectively.

2.4 Spatial-temporal implications of underreporting

Figure 3 shows that for the earliest time of June 1, 2020, there is little discrepancy between case and infection rates across the states, while for the later times there are immense differences in the rates, in that case rates tend to underrepresent infections to a great extent. One example of this is clear on the choropleth maps for cases and infections on July 20, 2021. While the map of case rates shows counts close to zero for almost all states, the map of infection rates reveals that states like Texas, Louisiana, and Georgia are hotspots for infections at that time. Furthermore, it is interesting that many of the states that are impacted appear to have a high degree of geographic connectivity. This leads into another key takeaway: The spatial extent indicated by infections is often understated by cases. For example, on October 20, 2020, we can see that while case rates are singled-out to be higher in a few northern states (namely, North and South Dakota),

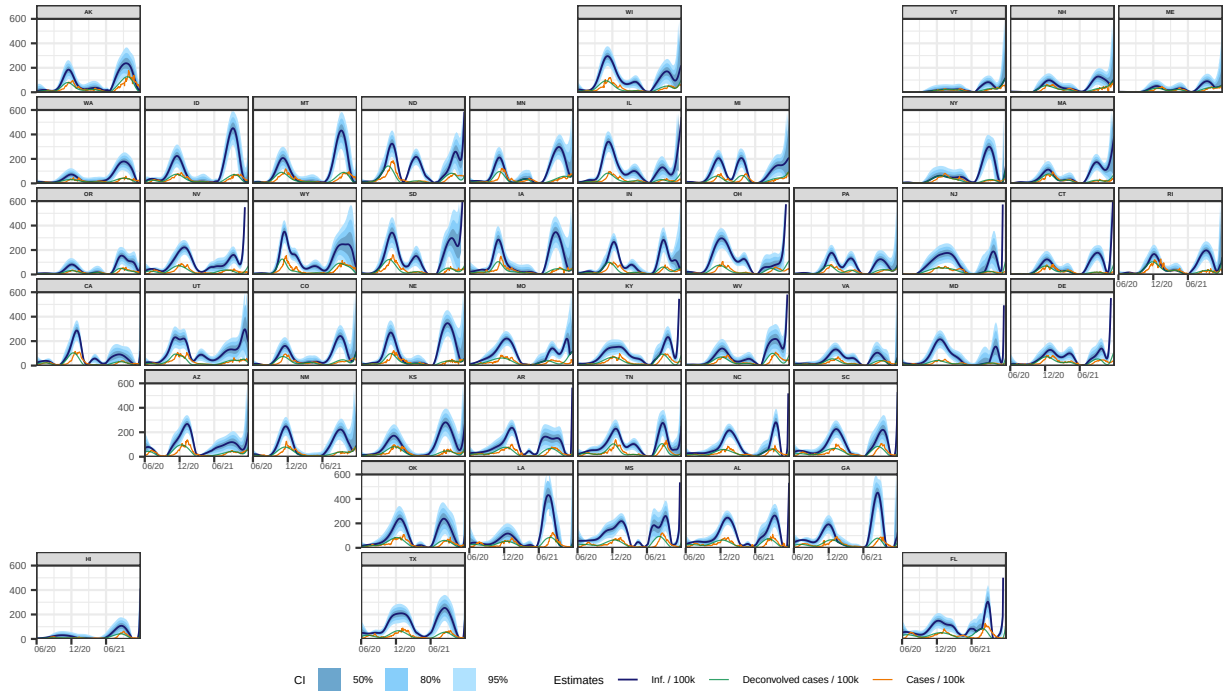

Figure 2: Estimates of the number of daily new infections per 100,000 population for each U.S. state from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% confidence intervals for the estimates, while the teal line represents the number of new daily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

they fail to reveal the impact on the surrounding states. In contrast, the corresponding plot of infection rates shows how higher rates of infections have also started to impact the surrounding states. Hence, cases provide a limited view of the spread of infections.

There is also the matter that the states that are primarily impacted according to cases can be discordant with those that are indicated by infections. For instance, if we consider August 27, 2021, it is clear that Montana and Idaho show some of the highest infection rates for that time. In contrast, the case map for this time does not indicate an influx in infections for these two states, but rather the higher rates tend to be restricted and localized to the southeastern states. To see an example of a converse situation, take the maps for December 17, 2020, where states like Tennessee and California are singled-out by cases, but not by infections. Hence, there are times when it can be misleading to look to infections for trends in cases and vice versa.

2.5 Earlier outbreaks are suggested by total infections and infections by variant

Figure 4 examines the infection estimates for a selection of states more closely. The top panel zooms in on the infection estimates for these states, while the bottom panel divides their estimated deconvolved cases into categories based on the circulating variant proportions at the time. From these plots, it is evident that there are times when the total infections and the infections broken down by the variant categories emphasize earlier outbreaks than indicated by cases. For instance, take the major Ancestral wave for California, Maryland, Idaho, Montana, or Ohio, where cases crest just after infections do. Such trends persist into Delta, though not necessarily for the same subset of states. Louisiana, Idaho and Montana are primes example of this during the period of Delta domination as we can clearly see that infections are a leading indicator of cases for the major Delta wave. The division by variant categories reveal the variant or variants that are behind these waves. The crest-trough patterns of these by-variant depictions align with infections rather than the cases by

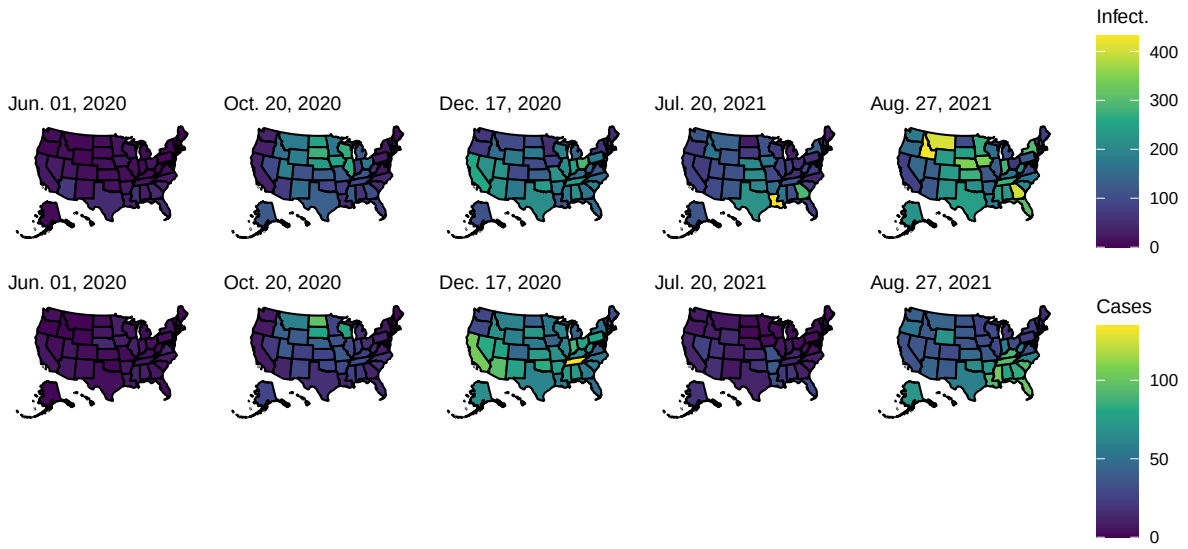

Figure 3: Choropleth maps of the state-level estimates of the number of daily new infections per 100,000 population (top row) and the daily new cases per 100,000 population (bottom row) for five times over the June 1, 2020 to November 29, 2021 period. Note that the first date was chosen as a baseline, while the other dates were chosen due to having large counts of infections across all states. In particular, we should note that the third and fifth dates present the largest number of infections across the 50 states from each year.

construction (as they are for deconvolved cases which are re-scaled to get the infection estimates).

2.6 The relationship between infections and hospitalizations is complicated

We systematically investigate the temporal relationship between infections and hospitalizations with Spearman's rank-correlation across different lags, shifting hospitalizations backward to align with infections. (Figure 5). The maximum average correlation across states is 0.513, occurring at a lag of 13 days. In contrast, we find that the greatest average Spearman correlation for cases is 0.691 and occurs at a lag of 1 day. That is, we find that case report rates are nearly contemporaneous to hospitalizations, while infection estimates clearly precede them.

The maximum correlation at a lag of 13 days is in similar to early estimates of the average time from infection to hospitalization of 9.7 days (95% CI: [5.4, 17.0]) for cases reported in January, 2020 in Wuhan, China as well as with estimates from across the pandemic in the UK that ranged from an average of 8.0 to 9.7 days[12].

While both the infection and deconvolved case case estimates are shown to be leading indicator of hospitalizations and their trajectories are similar, the average correlation they attain is vastly different. In particular, the correlation tends to be much greater for the deconvolved case estimates than for the infection estimates (with a difference of about 0.178 at the peaks). This finding is reasonably likely to stem from a difference in disease severity between the reported and unreported infections: unreported infections tend to be less severe and less likely to lead to hospitalization than those that are reported[13].

Figure 4: Top panel: Reported cases, deconvolved cases, and estimates of daily new infections (dark blue line) per 100,000 inhabitants. The blue shaded regions indicate the 50, 80, and 95% confidence bands, while the background is shaded to indicate the dominant variant in circulation at the time. Bottom panel: Deconvolved cases colored by variant per 100,000 inhabitants.

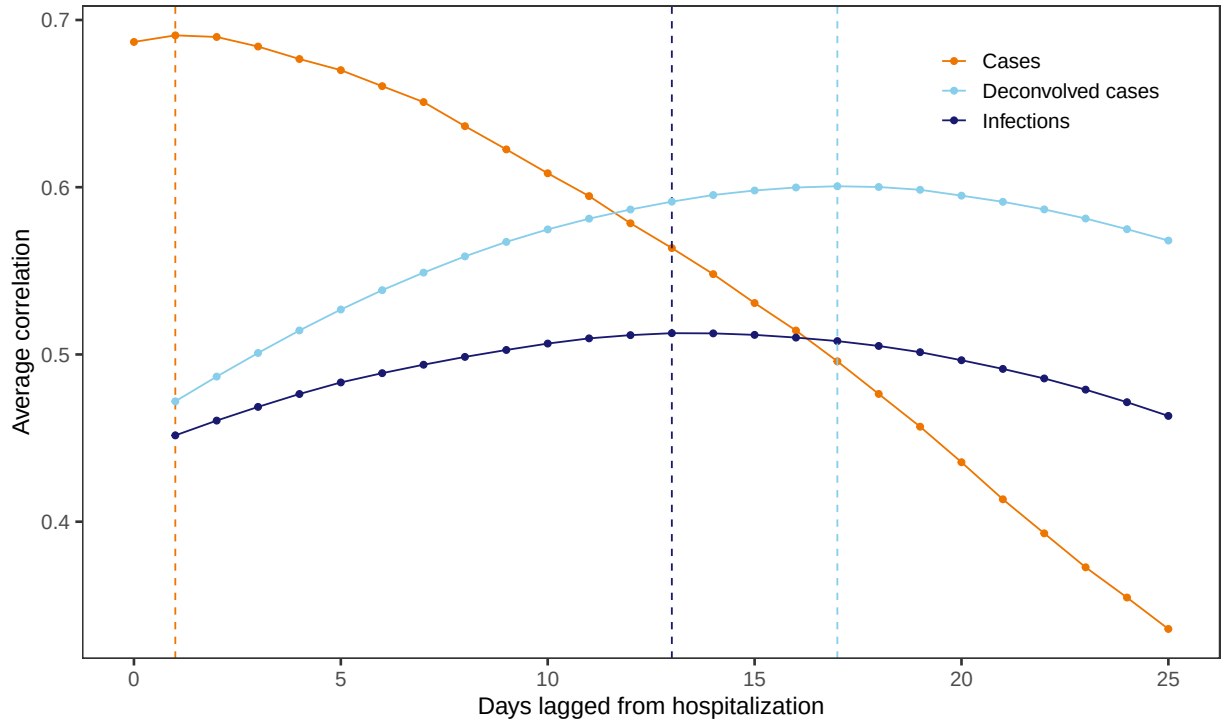

Figure 5: Spearman's correlation between each of the case, deconvolved case, and infection rates and hospitalization rates per 100,000. The correlations are calculated for each lag, state and rolling window of 61 days before averaging. The vertical dashed lines indicate the lags for which the highest average correlation is attained.

2.7 Estimating infection-hospitalization ratios

As a counterpart to the correlation analysis, we compute the time-varying infection-hospitalization ratios (IHRs) for each state using the correlation maximizing lag. We similarly compute the case-hospitalization ratios (CHRs) using their correlation maximizing lag for for comparison (Figure 6).

For each state, the CHRs tend to be larger and noisier relative to the IHRs. This supports our claim that the reported infections are more likely to require hospitalization than the unreported infections. Both the IHRs and CHRs exhibit similar geospatial and temporal trends as are noted for infections. Namely, states that are close in proximity (such as Ohio, Pennsylvania, and Virginia) tend to exhibit similar patterns in the IHRs and CHRs over time. In addition, there are similar spikes observed across many states during waves of infections that are driven by prominent new variants. For example, many states exhibit a striking spike in hospitalizations in mid-2021, which coincides with the rapid takeover of the Delta variant during that time[14]. This finding aligns with previous studies that found an increased risk in hospitalizations with Delta in comparison to other variants[15;16]. Similarly, during the fall of 2020 there tends to be another spike in the IHRs that rivals or surpasses that observed during the time of Delta (which is the case for states like New York, New Jersey, or Wyoming).

Overall, the relationship between infections and hospitalizations is complicated. We observe intermittent spikes that punctuate longer periods where the IHRs tend to stabilize somewhat, in that they tend to stay below 0.1 hospitalizations per infection. Unsurprisingly, these spikes tend to align with the emergence of new variants.

While we computed and compared CHRs and IHRs for all states, it is important to note that both likely to vary within states and depend on confounding variables such as age and the presence of major comorbidities[17]. 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[18].

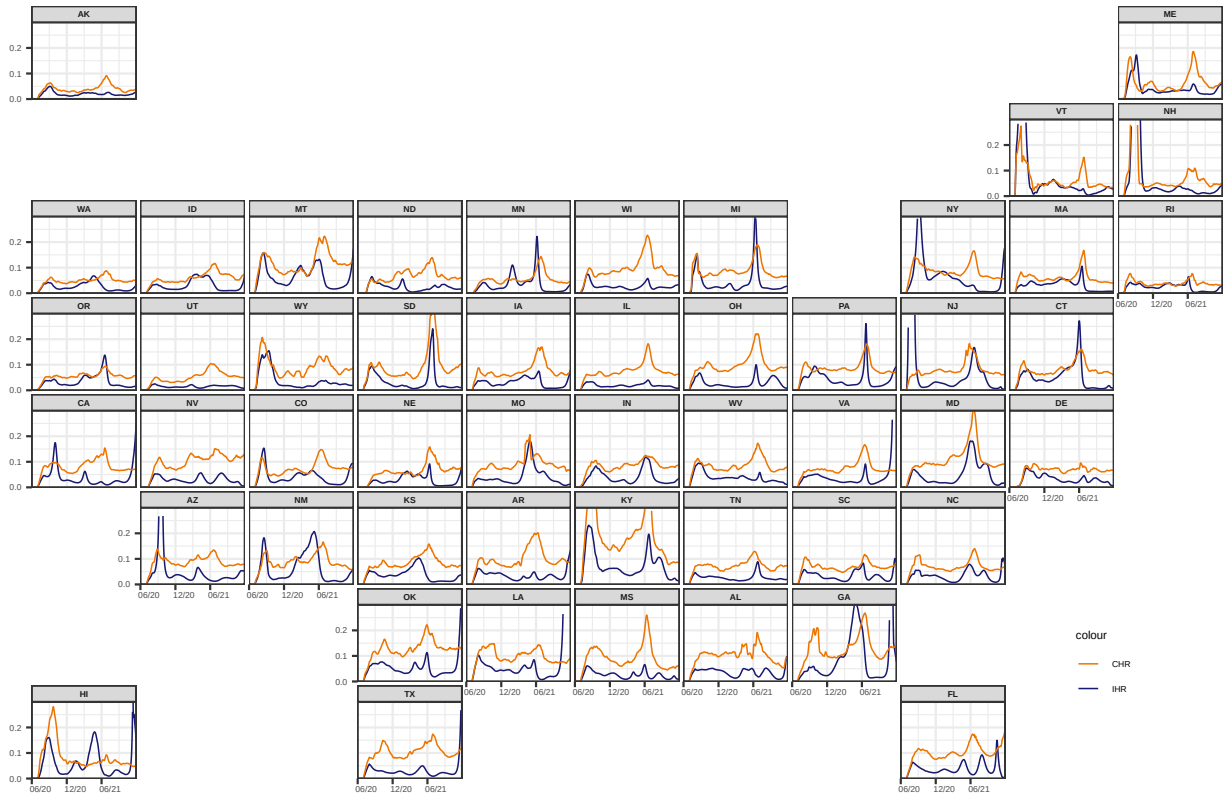

Figure 6: Time-varying IHR and CHR estimates for each state from June 1, 2020 to November 29, 2021, obtained using the corresponding optimal lag from the systematic lag analysis. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious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

3 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 both (a) that the pandemic impacted states earlier and at a larger scale than is indicated by cases and that (b) examining cases alone hides some spatio-temporal waves that become apparent by examining infections. We observe outbreaks in infections that are difficult to detect from cases alone such as the Delta wave in New Jersey, Connecticut, and Maryland. This suggests that cases paint an incomplete picture of the pandemic, especially when outbreaks are largely driven by unreported infections. Furthermore, since case reports generally follow symptom and infection onsets, cases have a built-in temporal bias. This is in addition to other biases from differences in reporting across states (such as temporary bottlenecks due influxes of data or more persistent processing issues that increase the average time from case detection to report [9;19]. Thus, while reported cases provide an indication of the trajectory of the pandemic, it is a delayed and incomplete version.

Our approach offers a number of advantages. For instance, we aim to incorporate as much state-specific information as possible when deriving our estimates. By using state-level case, line list, and variant circulation data, we are able to construct incubation and delay distributions that are specific to each state. Time-varying and state-specific seroprevalence data allows the reporting ratio estimates to similarly vary over space and time, a departure from existing work [20;21]. Existing approaches that use the delay distribution to generate infection estimates often only construct one delay distribution that is used for all states [22;23]. That is, our work avoids the assumption of geographic invariance, where it is assumed that all states have the same patterns of delay from symptom onset to case report. This assumption is unlikely to be true due to differences

in reporting pipelines, pandemic response, and variants in circulation, among other issues.

Another limitation of previous approaches to estimate latent infections is that they do not to account for reinfections. While reinfections represent a small fraction of total infections until later in the pandemic, ignoring them means that the infection-reporting ratio will tend to be underestimated with seroprevalence data alone. By accounting for these as well as the waning of seropositivity (See Methods Section 4.5), we more accurately estimate this ratio. However, we acknowledge that the extent to which each of these are accounted for could be improved upon in future work. Since the waning of immunity is likely to be variant-dependent [24], it follows that our model waning parameter may be better posed as a mixture of parameters for different variants with weights determined by the proportion of the variants circulating at the time in the state. Related to this is the issue that newer variants may escape detection [25;26]. While in a retrospective analysis where finalized data is used this is less likely to be an issue, this could very well pose a problem for real-time estimates of infections.

Regarding reinfections, a major reason why we chose an earlier end date of November 29, 2021 and ultimately decided to not tread into Omicron territory is because the Omicron variants come with substantial increase in the risk of reinfection in comparison to previous variants as Omicron has been shown to have an increased tendency towards immune escape [27–29]. So having quality reinfection data that is representative of each location under study is of the utmost importance for the Omicron era. Prior to the Omicron era, we find relative agreement between the infection data we have chosen and those from other potential sources [30–33]. While it would be ideal to use the reinfection rates over time for each U.S. state, most states do not publicly report reinfection data over the entire time period we considered or at all.

Using seroprevalence data to estimate the case-ascertainment ratio is subject to a number of issues, and precludes us from pushing the period of analysis past the Omicron wave in December 2021. While most state-level data suggests that reinfections tend to account for less than 20% of reported cases during Omicron [30–33], seropositivity rapidly reaches nearly 100% of the population, precluding its continued use. Due to these issues, alternative data sources for estimating the case-ascertainment ratio is necessary. For example, wastewater surveillance data is may be complementary to seroprevalence data, especially when testing is low [34]. However, viral detection is inconsistent across locations due to temperature, per-capita water use, and in-sewer travel time [34–36]. Sentinel surveillance streams for influenza-like illness or acute respiratory infection may provide decent proxies for COVID-19 incidence, especially when testing for mild cases of COVID-19 is diminishing or has ceased completely. Finally, alternative surveillance streams (potentially outside of public health) such as those from surveys, helplines, or medical records could potentially be integrated if they provide at least a rough indication of the disease intensity over time [6;37].

We adopt a relatively simple deconvolution-based approach and devote much of our efforts to tailoring our approach to the available data. A major result of this is the development of a way of to model the waning of detectable antibody levels and space-time-specific reporting ratios based on seroprevalence data. In a way, our approach is built for the data rather than trying to force the data to fit to an existing approach. However, our model is only as good as the quality and the quantity of the data provided to it. In our case, the lack of data is both a barrier to entry and a continual roadblock. The assumptions we are required to make as a consequence of this clearly limit the generalizability and call into question the reliability of the results. So while we highlight some interesting trends and numerical findings, these results are not definitive, but rather exploratory and intended to stimulate discussion on the challenging task of estimating infections. Despite these limitations, we are encouraged by the ability to use routine data to produce sensible estimates of infections in the United States and the plausibility of the apparent geospatial and temporal trends.

Well-informed, localized estimates of COVID-19 infections over time can help us to have a more clear and comprehensive understanding of the course of the pandemic. Such estimates contribute important information on the timing and magnitude of disease burden for each location and they highlight trends that may not be visible from case data alone. Therefore, our infection estimates provide key information for the ongoing debate 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.

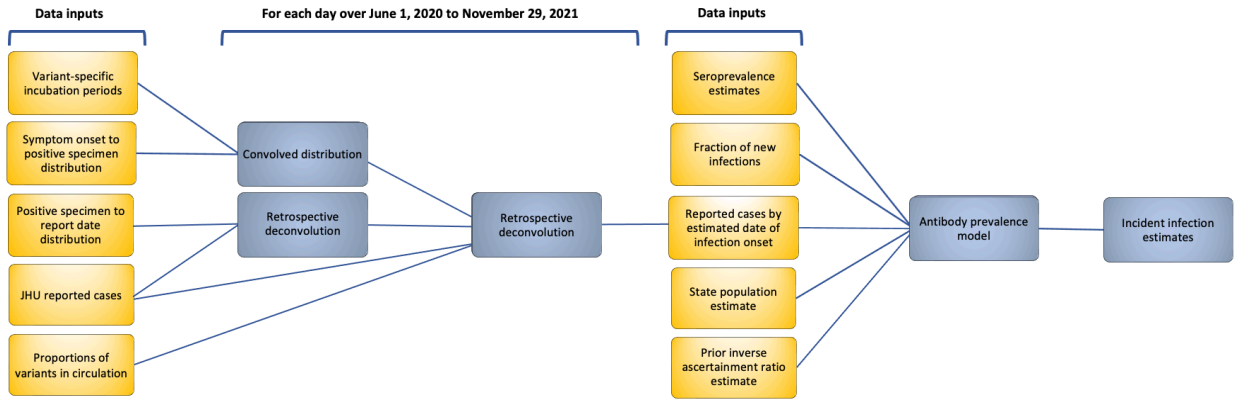

Figure 7: Flowchart of the inputted data and major analysis steps required to get from reported cases to incident infection estimates for each day over June 1, 2020 to November 29, 2021 for a state. Data sources are coloured in yellow, while data analysis steps are coloured in blue. The data sources that do not stem from an analysis step are literature estimates.

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 considered time period. Next, each incubation period and symptom onset to positive specimen delay distribution are joined using convolution to obtain variant-specific infection onset to positive specimen distributions for each time. Then 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. This lets us reach our ultimate goal of obtaining daily 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 CDC. Although there are both public and restricted versions of the dataset available containing the same patient records [38;39], 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 additional details and justifications are deferred to Supplementary Materials S1.1 and 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

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

Next we compute a similar national quantity $\tilde{p}_t(k) = \frac{1}{N_t} \sum_{s \in \mathcal{S}_t} \big(\#\ z_s \text{ with positive specimen at } s - k\big)$, without restricting to location ℓ. Next, let $\alpha_{\ell,t}$ be the ratio of $N_{\ell,t}$ to the number of cases reported by JHU CSSE[1] in the same window. Then, compute $p_{\ell,t}(k) = \alpha_{\ell,t}\tilde{p}_{\ell,t}(k) + (1 - \alpha_{\ell,t})\tilde{p}_t(k)$. 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 $\{\widehat{\pi}_{\ell,t}(k) : 0 < k \leq 60\}$ 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 Materials S1.3.

4.2 Estimating the incubation period distributions

One incubation period distribution is estimated for each variant under consideration. The variants we consider are Alpha, Beta, Gamma, Delta, and Omicron which are included because they are are designated to be variants of concern by WHO[40]. In addition, we include the Epsilon (California) and Iota (New York) variants because of their impact on those and the surrounding states[41;42]. We relegate all other variants to be in the Other category (which, for our purposes, is treated as a catch-all for all 2020 Ancestral variants observed in the U.S.) This decision to include an Other category is, in part, motivated by the lack of sequencing data for most states in 2020 as well as the presence of an Others category in the sequencing data for that time.

We obtain literature estimates of the variant-specific incubation periods for the eight aforementioned variant categories. For the Ancestral variant category, we use the literature estimates of the gamma distribution parameters[43]. For the Alpha, Beta, Gamma, Delta and Omicron variants, we use the mean and standard deviation of the number of days of incubation as reported in Tanaka et al.[44]; Grant et al.[45]; Ogata et al.[46]. Since the literature lacks reliable estimates for the incubation period of the Epsilon and Iota variants, 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 (as depicted in[14]).

We construct the incubation period distributions for the variants as gamma distributions. These distributions are the same for all states and based on literature estimates of the gamma parameters or the mean and standard deviation of the incubation period (in which case the method of moments is used like in Section 4.1 to fit a gamma density). Then, we discretize each resulting density to the support set, which is taken to be from 1 and 21 days. The implicit assumption for the lower bound is that there must be at least one day between infection and symptom onset (which follows the convention given in[47]). The assumption underlying the upper bound is that 21 days is the maximum number of days that the virus could be incubating in someone[48;49].

4.3 Variant circulation proportions

To estimate the daily proportions of the variants circulating in each state, we obtain the GISAID genomic sequencing data counts from CoVariants.org[14;50]. Since these counts are biweekly totals, we apply multinomial logistic regression using a third-order polynomial in time to get estimates of the daily proportions for the eight variant categories separately for each state. We then predict the probabilities over all days from March 1, 2020 to March 1, 2023 (which we justify in Section 4.4).

For clarity, we will present the general model form for one pair of variant categories, Other and Alpha, where Alpha is considered to be the reference level. The same form follows for the other variant categories.

Let the response, Y, be a categorical variable with eight levels (Alpha, Beta, Epsilon, Iota, Gamma, Delta, Omicron, Other). There are three predictors for the third-order polynomial in time. From these, the general form of the multinomial logistic regression model when the response is the logarithm of the relative odds ratio of Other with respect to Alpha is as follows:

$$\ln\left[\frac{p_{\text{Other}}}{p_{\text{Alpha}}}\right] = \beta_{0,1} + \beta_{1,1}T + \beta_{2,1}T^2 + \beta_{2,1}T^3$$

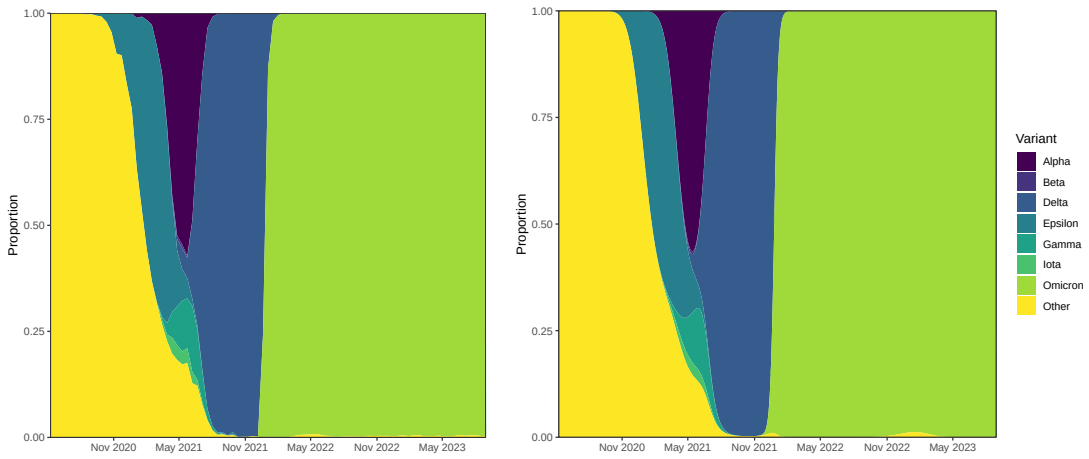

Figure 8: Left panel: Original biweekly proportions of the variants in circulation for California (pre-multinomial logistic regression). Right panel: Daily proportions of the variants in circulation for California (post-multinomial logistic regression and prediction).

where

$$p_{\text{Other}} = P(Y = \text{Other}|T)$$
$$p_{\text{Alpha}} = P(Y = \text{Alpha}|T).$$

Figure 8 shows the proportions by variant for California prior before and after being subject to the multinomial logistic regression model and subsequent prediction, from which it is evident that the estimates for the proportions have been smoothed.

4.4 Convolution and retrospective deconvolution to obtain deconvolved cases

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 9 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.

An important aspect of our methods is that deconvolution is not the same as a shift because simply shifting cases back in time and increasing them by some factor fails to capture the spatio-temporal dynamics of the pandemic. In our situation, reported cases are "pushed back" by the delays shown in Figure 1).

Now, the previous two steps enabled us to estimate one incubation period per variant and one symptom onset to positive specimen distribution for each state at each time under consideration. We proceed to convolve each such pair of distributions to get estimated infection to positive specimen distributions and, hence, estimated time-varying probabilities for the delay from infection onset to positive test specimen date for each state.

The main goal for the retrospective deconvolution stage is to estimate the daily number of new infections for each time using the dates that those cases were eventually reported. To this end, there are two types of deconvolution performed. The first is the deconvolution from report to positive specimen date and the second is from positive specimen date to infection onset date. We allocated the deconvolutions in this way to

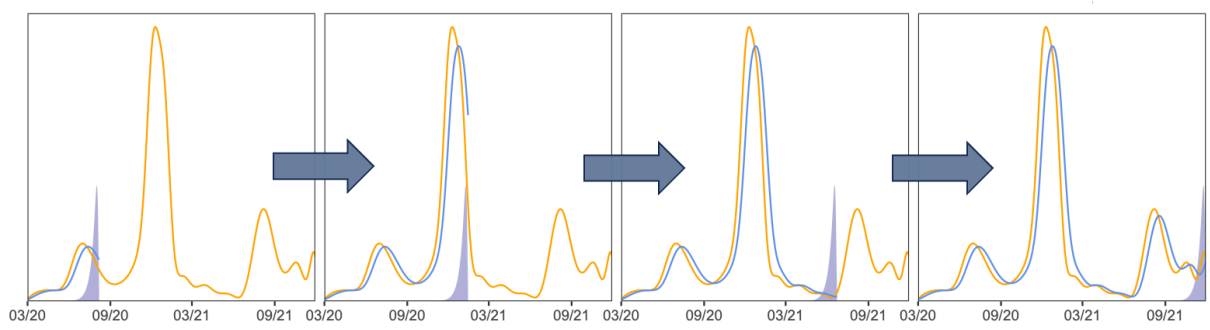

Figure 9: A general depiction of convolving smoothed cases (orange line) with the corresponding delay probabilities (shaded blue area) to get the convolved estimates (blue line) over four different times.

allow us to get the daily deconvolved case estimates by variant. The intermediate of positive specimen date was chosen because the variant proportion estimates are aligned to this date. So for each state at each time, nine deconvolutions are performed in total. The first is the deconvolution from report to positive specimen date, followed by the eight deconvolutions from positive specimen to infection onset for the eight different variant categories.

We will start by describing the first type of deconvolution performed from report to positive specimen date in detail and then describe the second type in terms of the changes made with respect to the first. For each state, we achieve the first goal to estimate positive specimen dates for the cases by solving an optimization problem. For this problem, let $\mathcal{T}$ represent the extended deconvolution period from March 1, 2020 to March 1, 2023, which was used to minimize the effect of boundary issues and to produce sufficient deconvolved case estimates for further analysis. Let $\hat{p}_t$ be probabilities from the estimated positive specimen to report date distribution for $t \in \mathcal{T}$, y_t the number of new JHU cases reported, and where $D^{(4)}$ is the discrete derivative matrix of order 4 such that $D^{(4)}x$ yields all 4^{th}-order differences of the vector x. From these, we estimate the deconvolved case counts across time by solving for the vector x in

$$\underset{x}{\text{minimize}} \sum_{t\in\mathcal{T}} \left(y_t - \sum_{k=1}^{d} \widehat{\pi}_t(k) x_{t-k} \right)^2 + \lambda \sum_{t\geq 5} \left| x_t - 4x_{t-1} + 6x_{t-2} - 4x_{t-3} + x_{t-4} \right|.$$

The above loss function decouples into two parts which trade data fidelity with desired smoothness (that encapsulate the classic bias-variance trade off). The first part represents minimizing the sum of squared errors between the JHU reported cases and the estimates, while the second part captures the smoothness of the estimates (smaller values being more smooth). The tuning parameter λ determines the relative importance of these competing goals. For details on the selection of λ and how this trend-filtering-regularized least squares deconvolution problem is solved, please see Supplementary Materials S1.4.

The COVIDcast API[37] is used to retrieve the daily number of new confirmed COVID-19 cases for each state that are based on reports from the John Hopkins Center for Systems Science and Engineering (JHU CSSE)[1]. From the same API, we also 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). Both datasets are updated as of June 6, 2022.

From this first type of deconvolution, we obtain case estimates by positive specimen date for each state. The second type of deconvolution, where the goal is to get estimates of the infection onset date for these cases, follows the form of first type, save for two key modifications to the inputs. Firstly, we utilize the results from the first deconvolution and, secondly, we must update the probabilities to be the convolutional estimates of the infection to positive specimen distributions. Thus, for a fixed variant category, y_t is the number of new cases deconvolved to positive specimen date multiplied by the estimated proportion of the variant in circulation in the state at t, and $\hat{p}_t$ are the probabilities from the estimated infection onset to positive specimen distribution for $t \in \mathcal{T}$. With these modifications to the inputs, the deconvolution proceeds in the exact same way as before. Since this deconvolution is done separately for each variant category, we ultimately obtain deconvolved case estimates by the date of infection onset that are separated by variant.

4.5 Inverse reporting ratio and the antibody prevalence model

The infection estimates from retrospective deconvolution are derived solely from the infection onset dates of the reported cases. To capture the unreported infections, it is necessary to adjust these deconvolved case estimates by a scaling factor that approximates the ratio of the true number of new infections to the new reported infections. We refer to this quantity as the inverse reporting ratio and denote it by a_t for day t. Our new goal is to estimate this quantity for every state at every time under consideration from June 1, 2020 to November 29, 2021.

The number of new reported infections is obtained from our deconvolved case estimates. As for the true infections, since seroprevalence of anti-nucleocapsid antibodies is used to estimate the percentage of people who have at least one resolving or past infection[51], we can use the change in subsequent seroprevalence measurements to capture new infections, accounting for those whose antibody levels fall below the detection threshold. We can adjust the retrospective deconvolution estimates using a model that is based on such seroprevalence estimates.

To estimate the proportion of the population in each state with evidence of previous infection across time, we use two major seroprevalence surveys that were led by the CDC: the 2020–2021 Blood Donor Seroprevalence Survey and the 2020–2022 Nationwide Commercial Lab Seroprevalence Survey[52;53]. A description of how we handled the dates in these datasets can be found in Supplementary Materials S1.5.

The daily fraction of new infections are estimated from the provided incidence of suspected reinfections over March 2020 to April 2022 in Clark County, which is based on surveillance work conducted by the Southern Nevada Health District (SNHD) and reported by Ruff et al.[30]. The proportion of new cases per week that are suspected reinfections are calculated by dividing the number of suspected reinfections by all new PCR-identified cases during the same week.

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.

For each state, let s_m be the seroprevalence estimate on m, w_m be the corresponding inverse variance weight, and $C_{m-1}^{m} = \sum_{t=m-1}^{m} c_t$ be the total reported infections from $m-1$ to m scaled by the state's population. To account for reinfections, we multiply the change in reported infections for m by the corresponding fraction of new infections, z_m.

Using these components, we construct the following model separately for each state

$$\begin{aligned} s_m &= (1-\gamma)s_{m-1} + a_m C_{m-1}^{m} z_m + \epsilon_m, \qquad && \epsilon_m \sim N(0, w_m\sigma_\epsilon^2) \\ a_{m+1} &= 3a_m - 3a_{m-1} + a_{m-2} + \eta_m, && \eta_m \sim N(0, \sigma_\eta^2) \end{aligned} \tag{1}$$

where γ is the percentage of people whose level of infection-induced antibodies falls below the detection threshold between time m and time $m+1$. Informally, we refer to γ as the waning parameter and we call this model the population antibody prevalence model.

We express the antibody prevalence model as a state-space model. This representation allows for convenient handling of missing data, extrapolation before and after the period of observed seroprevalence measurements, and maximum likelihood estimates of γ and σ_ϵ^2. Details of this methodology and the computation of the associated uncertainty measurements are deferred to the Supplementary Materials S1.6.

4.6 Lagged correlation to hospitalizations and time-varying IHRs

We use our infection estimates in a lagged correlation analysis with confirmed COVID-19 hospitalizations. Our primary 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 due to the virus. To remove day of the week effects, both the infection and hospitalization signals are subject to a 7-day moving average (center-aligned) before their conversion to rates.

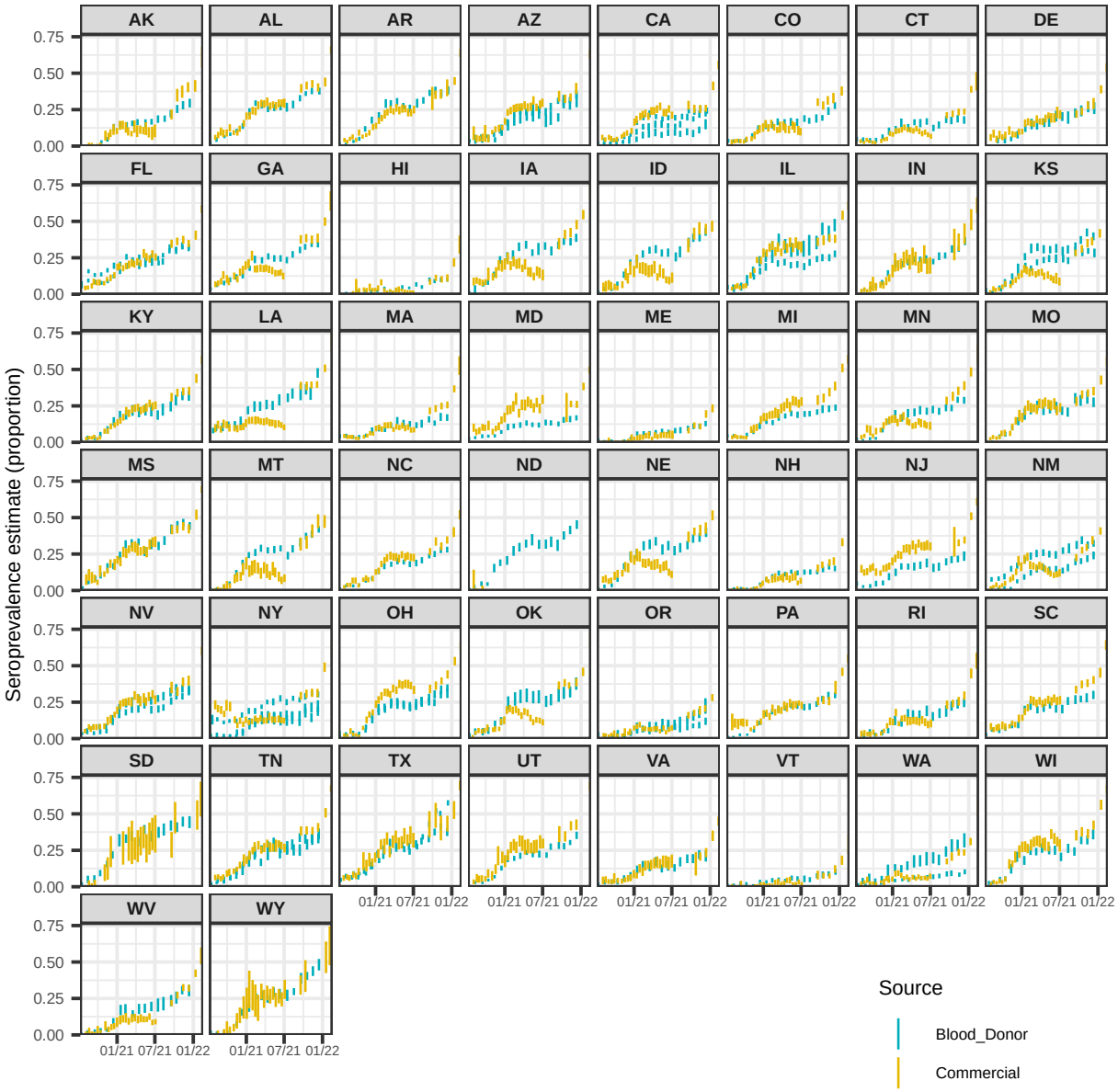

Figure 10: A comparison of the seroprevalence estimates from the Commercial Lab Seroprevalence Survey dataset (yellow) and the 2020–2021 Blood Donor Seroprevalence Survey dataset (blue). Note that the maximum and the minimum of the line ranges are the provided 95% confidence interval bounds to give a rough indication of uncertainty.

For each considered lag, we calculate the Spearman's correlation between the state infection and hospitalization rates for each observed day over the June 1, 2020 to November 29, 2021 time period with a center-aligned rolling window of 61 days for each such computation. We then calculate the average correlation 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. To compute this for a given day, we start with the standard computation where the number of individuals who are hospitalized due to COVID-19 on a day are divided those by the estimated total number who were infected on the lagged number of days before. To stabilize these lagged IHR estimates, we use 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

Code availability

References

[1] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):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] Virginia E. Pitzer, Melanie Chitwood, Joshua Havumaki, Nicolas A. Menzies, Stephanie Perniciaro, Joshua L. Warren, Daniel M. Weinberger, and Ted Cohen. The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. *American Journal of Epidemiology*, 190(9):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] Matt D.T. Hitchings, Natalie E Dean, Bernardo García-Carreras, Thomas J. Hladish, Angkana T. Huang, Bingyi Yang, and Derek A.T. Cummings. The usefulness of the test-positive proportion of severe acute respiratory syndrome coronavirus 2 as a surveillance tool. *American Journal of Epidemiology*, 190(7): 1396–1405, 2021.

[8] Lorenzo Pellis, Francesca Scarabel, Helena B. Stage, Christopher E. Overton, Lauren H.K. Chappell, Elizabeth Fearon, Emma Bennett, Katrina A. Lythgoe, Thomas A. House, Ian Hall, 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(1829):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] Nigel Garrett, Asa Tapley, Jessica Andriesen, Ishen Seocharan, Leigh H. Fisher, Lisa Bunts, Nicole Espy, Carole L. Wallis, April Kaur Randhawa, Nzeera Ketter, et al. High rate of asymptomatic carriage associated with variant strain Omicron. *MedRxiv*, 2022.

[12] Thomas Ward and Alexander Johnsen. Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLoS One*, 16(10):e0257978, 2021.

[13] Narjiss Sallahi, Heesoo Park, Fedwa El Mellouhi, Mustapha Rachdi, Idir Ouassou, Samir Belhaouari, Abdelilah Arredouani, and Halima Bensmail. Using unstated cases to correct for covid-19 pandemic outbreak and its impact on easing the intervention for qatar. *Biology*, 10(6):463, 2021.

[14] Emma Hodcroft. CoVariants: SARS-CoV-2 mutations and variants of interest. https://covariants.org, 2021.

[15] Katherine A. Twohig, Tommy Nyberg, Asad Zaidi, Simon Thelwall, Mary A. Sinnathamby, Shirin Aliabadi, Shaun R. Seaman, Ross J. Harris, Russell Hope, Jamie Lopez-Bernal, 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(1):35–42, 2022.

[16] Tommy Nyberg, Neil M. Ferguson, Sophie G. Nash, Harriet H. Webster, Seth Flaxman, Nick Andrews, Wes Hinsley, Jamie Lopez Bernal, Meaghan Kall, Samir Bhatt, 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(10332):1303–1312, 2022.

[17] Clark D. Russell, Nazir I. Lone, and J. Kenneth Baillie. Comorbidities, multimorbidity and COVID-19. *Nature Medicine*, 29(2):334–343, 2023.

[18] Spencer J. Fox, Emily Javan, Remy Pasco, Graham C. Gibson, Briana Betke, José L. Herrera-Diestra, Spencer Woody, Kelly Pierce, Kaitlyn E. Johnson, Maureen Johnson-León, et al. Disproportionate impacts of COVID-19 in a large US city. *PLOS Computational Biology*, 19(6):e1011149, 2023.

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

[20] H. Juliette T. Unwin, Swapnil Mishra, Valerie C. Bradley, Axel Gandy, Thomas A. Mellan, Helen Coupland, Jonathan Ish-Horowicz, Michaela A.C. Vollmer, Charles Whittaker, Sarah L. Filippi, et al. State-level tracking of COVID-19 in the United States. *Nature Communications*, 11(1):6189, 2020.

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

[22] Melanie H. Chitwood, Marcus Russi, Kenneth Gunasekera, Joshua Havumaki, Fayette Klaassen, Virginia E Pitzer, Joshua A. Salomon, Nicole A. Swartwood, Joshua L. Warren, Daniel M. Weinberger, 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(8):e1010465, 2022.

[23] Maria Jahja, Andrew Chin, and Ryan J. Tibshirani. Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. *Statistical Science*, 37(2):207–228, 2022.

[24] Nick Pooley, Salim S. Abdool Karim, Behazine Combadière, Eng Eong Ooi, Rebecca C. Harris, Clotilde El Guerche Seblain, Masoumeh Kisomi, and Nabila Shaikh. Durability of vaccine-induced and natural immunity against COVID-19: A narrative review. *Infectious Diseases and Therapy*, 12(2):367–387, 2023.

[25] National Institutes of Health. Assessing how SARS-CoV-2 mutations might affect rapid tests. https://www.nih.gov/news-events/nih-research-matters/assessing-how-sars-cov-2-mutations-might-affect-rapid-tests, 2022.

[26] U.S. Food and Drug Administration. SARS-CoV-2 viral mutations: Impact on COVID-19 tests. https://www.fda.gov/medical-devices/coronavirus-covid-19-and-medical-devices/sars-cov-2-viral-mutations-impact-covid-19-tests, 2023.

[27] Jia Wei, Nicole Stoesser, Philippa C Matthews, Tarnjit Khera, Owen Gethings, Ian Diamond, Ruth Studley, Nick Taylor, Tim EA Peto, A Sarah Walker, et al. Risk of sars-cov-2 reinfection during multiple omicron variant waves in the uk general population. *Nature Communications*, 15(1):1008, 2024.

[28] Juliet RC Pulliam, Cari van Schalkwyk, Nevashan Govender, Anne von Gottberg, Cheryl Cohen, Michelle J Groome, Jonathan Dushoff, Koleka Mlisana, and Harry Moultrie. Increased risk of sars-cov-2 reinfection associated with emergence of omicron in south africa. *Science*, 376(6593):eabn4947, 2022.

[29] Elias Eythorsson, Hrafnhildur Linnet Runolfsdottir, Ragnar Freyr Ingvarsson, Martin I Sigurdsson, and Runolfur Palsson. Rate of sars-cov-2 reinfection during an omicron wave in iceland. *JAMA Network Open*, 5(8):e2225320–e2225320, 2022.

[30] Jeanne Ruff, Ying Zhang, Matthew Kappel, Sfurti Rathi, Kellie Watkins, Lei Zhang, and Cassius Lockett. Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. *Emerging Infectious Diseases*, 28(10):1977, 2022.

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

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

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

[34] Oliver McManus, Lasse Engbo Christiansen, Maarten Nauta, Lene Wulff Krogsgaard, Naja Stolberg Bahrenscheer, Lene von Kappelgaard, Tobias Christiansen, Mikkel Hansen, Nicco Claudio Hansen, Jonas Kähler, et al. Predicting COVID-19 incidence using wastewater surveillance data, Denmark, October 2021–June 2022. *Emerging Infectious Diseases*, 29(8):1589, 2023.

[35] Olga E. Hart and Rolf U. Halden. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities and challenges. *Science of the Total Environment*, 730:138875, 2020.

[36] Xuan Li, Shuxin Zhang, Samendra Sherchan, Gorka Orive, Unax Lertxundi, Eiji Haramoto, Ryo Honda, Manish Kumar, Sudipti Arora, Masaaki Kitajima, et al. Correlation between SARS-CoV-2 RNA concentration in wastewater and COVID-19 cases in community: A systematic review and meta-analysis. *Journal of Hazardous Materials*, 441:129848, 2023.

[37] Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, Sumit Agrawal, Wael Al Saeed, Taylor Arnold, Amartya Basu, Jacob Bien, et al. An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51):e2111452118, 2021.

[38] 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.

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

[41] Shangxin Yang, Peera Hemarajata, Evann E. Hilt, Travis K. Price, Omai B. Garner, and Nicole M. Green. 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(5):649–652, 2022.

[42] Ralf Duerr, Dacia Dimartino, Christian Marier, Paul Zappile, Guiqing Wang, Jennifer Lighter, Brian Elbel, Andrea B Troxel, Adriana Heguy, 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(18):e152702, 2021.

[43] Lauren C. Tindale, Jessica E. Stockdale, Michelle Coombe, Emma S. Garlock, Wing Yin Venus Lau, Manu Saraswat, Louxin Zhang, Dongxuan Chen, Jacco Wallinga, and Caroline Colijn. Evidence for transmission of COVID-19 prior to symptom onset. *eLife*, 9:e57149, 2020.

[44] Hideo Tanaka, Tsuyoshi Ogata, Toshiyuki Shibata, Hitomi Nagai, Yuki Takahashi, Masaru Kinoshita, Keisuke Matsubayashi, Sanae Hattori, and Chie Taniguchi. Shorter incubation period among COVID-19 cases with the BA. 1 Omicron variant. *International Journal of Environmental Research and Public Health*, 19(10):6330, 2022.

[45] Rebecca Grant, Tiffany Charmet, Laura Schaeffer, Simon Galmiche, Yoann Madec, Cassandre Von Platen, Olivia Chény, Faïza Omar, Christophe David, Alexandra Rogoff, 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.

[46] Tsuyoshi Ogata, Hideo Tanaka, Fujiko Irie, Atsushi Hirayama, and Yuki Takahashi. Shorter incubation period among unvaccinated delta variant coronavirus disease 2019 patients in Japan. *International Journal of Environmental Research and Public Health*, 19(3):1127, 2022.

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

[48] Nazar Zaki and Elfadil A. Mohamed. The estimations of the COVID-19 incubation period: A scoping reviews of the literature. *Journal of Infection and Public Health*, 14(5):638–646, 2021.

[49] Jordi Cortés Martínez, Daewoo Pak, Gabriela Abelenda-Alonso, Klaus Langohr, Jing Ning, Alexander Rombauts, Mireia Colom, Yu Shen, and Guadalupe Gómez Melis. 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):1–7, 2022.

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

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

[52] 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.

[53] 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.

[54] 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.

[55] Aaditya Ramdas and Ryan J. Tibshirani. Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858, 2016.

[56] Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

[57] Ryan J. Tibshirani. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. *Foundations and Trends in Machine Learning*, 15(6):694–846, 2022.

[58] James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*, volume 38. OUP Oxford, 2012.

[59] Jouni Helske. KFAS: Exponential family state space models in R. *Journal of Statistical Software*, 78(10): 1–39, 2017.

Acknowledgements

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[50], on which this research is based.

ATTN: Grants, Delphi, etc.

Author contributions

Competing interests

The authors declare no competing interests.

Online Supplement

S1 Additional information about datasets or estimation methodology

S1.1 Additional details on the date fields in the CDC linelist

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 date Jahja et al. [23]. 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 S1 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 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 S1, from which it is clear the fraction of missing cases is substantial for many states, often surpassing 50% [23]. 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).

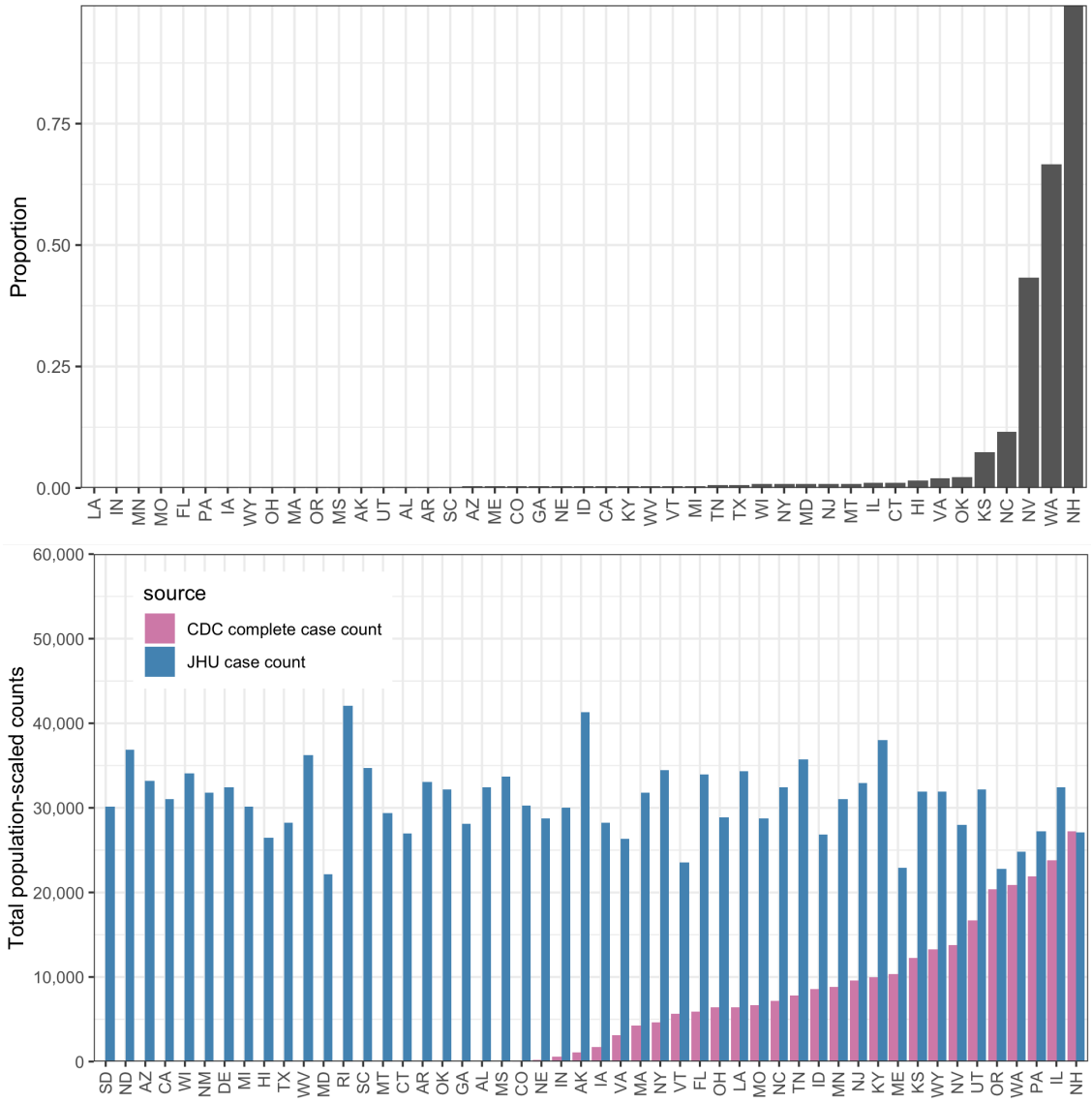

Figure S1: Top panel: Proportion of complete cases with zero delay between positive specimen and report date in the restricted CDC line list dataset. Bottom panel: Complete case counts by state in the CDC line list versus the cumulative complete case counts from JHU CSSE as of June 6, 2022. All counts have been scaled by the 2022 state populations as of July 1, 2022 from U.S. Census Bureau, Population Division [54].

S1.2 Table on the percent pairwise occurrence of events in the CDC line list

S1.3 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 [23]):

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

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).

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 two 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. As in Jahja et al.[23], we update the conditional expectation formula to reflect these two assumptions:

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

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

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

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.

S1.4 How we solve the the trend-filtering problem and select λ

We solve the trend-filtering-regularized least squares deconvolution problem by employing the ADMM algorithm from Ramdas and Tibshirani[55] that is described in Appendix A of Jahja et al.[23]. The solution to

the problem is an adaptive piecewise cubic polynomial[56;57].

The tuning parameter, λ, is selected by using 3-fold cross validation similar to Jahja et al.[23] in which every third infection count is reserved for testing. The tuning parameter that results in the smallest mean squared error is selected.

S1.5 Details on the construction of date variables in the seroprevalence datasets

The date variables that come with the two seroprevalence datasets are not the same 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.

S1.6 State space representation of the antibody prevalence model

The antibody prevalence model from Equation 1 is conceptualized as a Gaussian state space model (as in[58;59]).

In general, for $t = 1, \ldots, n$, 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

$$y_t = Z\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, H_t) \tag{2}$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim N(0, Q_t) \tag{3}$$

where $\alpha_1 \sim N(a_1, P_1)$ and there is independence amongst α_1, ϵ_t and η_t[58;59]. For notational compactness, we let $\alpha = \left(\alpha_1^\top, \ldots, \alpha_n^\top\right)$ and $y = \left(y_1^\top, \ldots, y_n^\top\right)$.

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 nature[58].

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 forward[58]. 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

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

with covariance,

$$P_{t+1} = \mathrm{Var}(\alpha_{t+1} \mid y_t, \ldots, y_1).$$

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

$$\hat{a}_t = \mathbb{E}[\alpha_t \mid y_n, \ldots, y_1] \tag{4}$$

$$V_t = \mathrm{Var}(\alpha_t \mid y_n, \ldots, y_1). \tag{5}$$

The filtering and smoothing steps are based on recursions that are described in Appendix A of Helske[59] 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 2 and 3 as follows:

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad Z = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad H_m = \begin{bmatrix} w_{m,c}\sigma_o^2 & 0 \\ 0 & w_{m,b}\sigma_o^2 \end{bmatrix}$$

$$\alpha_m = \begin{bmatrix} s_m \\ a_m \\ a_{m-1} \\ a_{m-2} \end{bmatrix} \qquad T_m = \begin{bmatrix} \gamma & C_{m-1}^m z_m & 0 & 0 \\ 0 & 3 & -3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad Q = \begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_a^2 \end{bmatrix}$$

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

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 Unwin et al. [20] 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 $\sigma_s^2 = 0.000003$, $\gamma = 0.99$, 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 4 and 5. Forwards and backwards extrapolation is then used to estimate the ratios and covariance outside of the observed seroprevalence range [58], 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 S1.7 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

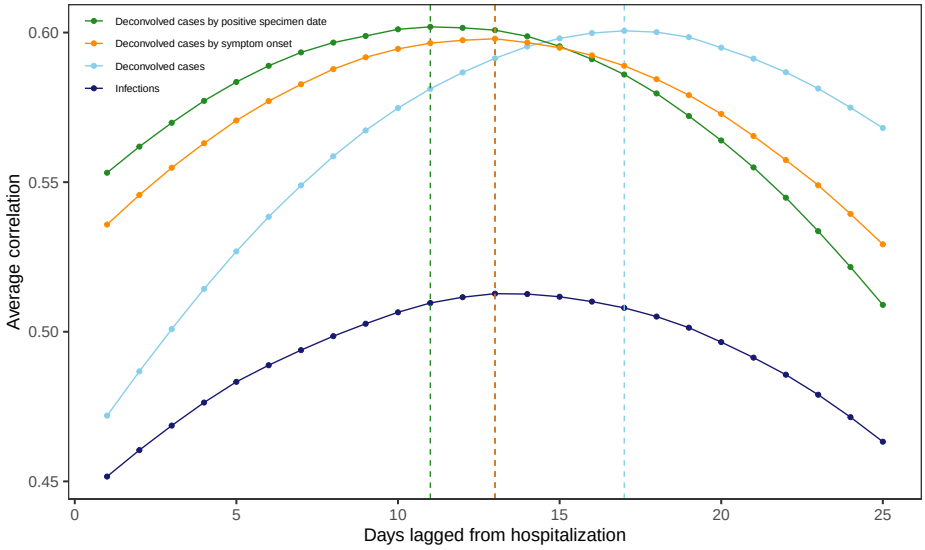

Figure S2: Lagged Spearman's correlation between the infection and hospitalization rates per 100,000 averaged for each lag across U.S. states and days over June 1, 2020 to November 29, 2021, and taken over a rolling window of 61 days. The infection rates are based on the counts for the deconvolved case 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.

additionally enforce that the lower bound must be at least the deconvolved case estimate for the time under consideration.

S1.7 Bayesian specification of the antibody prevalence model

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

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

and likelihood

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

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

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

and variance

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

S1.8 Scaling by population

Annual estimates of the resident state populations as of July 1 of 2020 and 2021 are taken from the December 2022 press release on the U.S. Census Bureau website[54]. Unless otherwise specified, we use the July 1, 2020 estimates.

S1.9 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 S2. From this, we can see that the deconvolved case or 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 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 about 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 17 days.