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]

[a]Department of Statistics, The University of British Columbia
[b]Department of Statistics & Data Science, Carnegie Mellon University
[c]Department of Computer and Electrical Engineering, University of Southern California
[d]Department of Statistics, The University of California, Berkeley

Version: March 27, 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.

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

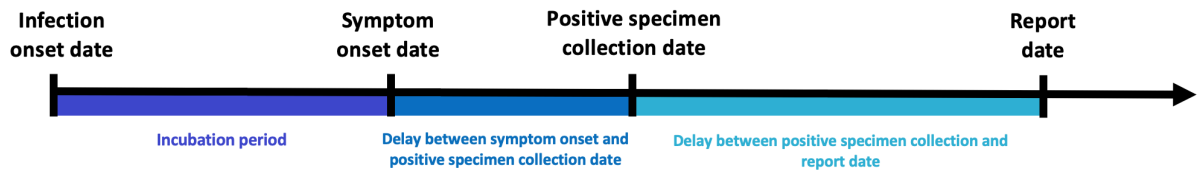

Figure 1: Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

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.

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

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 2–4). Outbreaks in infections precede those in reported cases and are reliably larger in magnitude. But simply shifting cases back in time and increasing them by a constant multiplicative factor fails to capture the spatio-temporal dynamics of the pandemic.

The largest per-capita outbreaks prior to Omicron were observed in the late summer or early fall of 2021 in Georgia, Louisiana, Idaho, Montana, and Wyoming, matching the intuition of similar viral spread in clusters of geographically proximate states. During this time, the two states that have the highest rate of infections on single day are Georgia (451 infections per 100K, on August 15, 2021) and Idaho (also 451 infections per 100K, on September 7, 2021). The period of lowest viral transmission is observed in the summer and fall of 2020. During this time, New Hampshire saw only about 1 new infection per million residents per week ATTN: really?. From June 2020 to the end of August, Vermont saw only about 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, and deconvolved cases—reported cases "pushed back" by the delays shown in Figure 1.

Most states exhibit 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, respectively. Similar patterns of these major surges 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.

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. Additionally, a wave of infections is present in the spring of 2021 for North Dakota and South Dakota, preceeding the Alpha wave in other jurisdictions. For the specific date of Oct. 20, 2020, Figure 3 shows increased case reports in a cluster of Northern-Midwest states like North Dakota, South Dakota, and Wisconsin relative to others. However, the infection rates more clearly emphasize that this phenomenon is not limited to these states, but extends out to the surrounding states as well. In the fall of 2021, 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. It is clear that cases fail to adequately capture the rise and fall in infections during this time.

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 across states and variants. For the 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 Delta wave is only faintly detectable from cases in a number of Eastern states (e.g., Maryland and Connecticut). The Ancestral wave is poorly represented ATTN: what does "poorly represented" mean by cases in several midwestern states (for example, Illinois, Indiana, and Ohio). 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 between cases and infections decreases with time. While the Delta wave is somewhat apparent from the case counts for all states (Figure 2), infection estimates suggest that case counts severely underestimate infections during this time for many states, moreso than in earlier waves. The most extreme was New Jersey, where about 4.6% of estimated infections were eventually reported as cases. Similarly low are Maryland (7.4%), Connecticut (8.0%), and Florida (8.7%). This issue extends to most states: in 39 states fewer than 30% of infections would eventually appear in case reports. This ratio was less extreme in earlier waves, and its effects most apparent in different regions. During Alpha, Louisiana had the lowest ratio of infections to cases (11.7%) followed by California (14.4%). Such patterns are even less apparent during the Ancestral wave, where Ohio and Maryland had the lowest ratio of reported cases to infections at 22.0% and 22.3%, respectively.

Figure 3 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. For example, on July 20, 2021, while the map of case rates shows low and geographically consistent spread, infection rates reveal that Texas, Louisiana, Georgia, and their neighbors are hotspots at that time. Generally, the spatial extent of infections is often 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), they fail to reveal the impact on the surrounding states: infection rates are similarly elevated in all of the surrounding states.

By focusing on wtates 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 rates tend to be localized to the Southeastern states. However, the opposite occurs as well: on December 17, 2020, Tennessee and California have the highest case rates but minimal infections relative to other states.

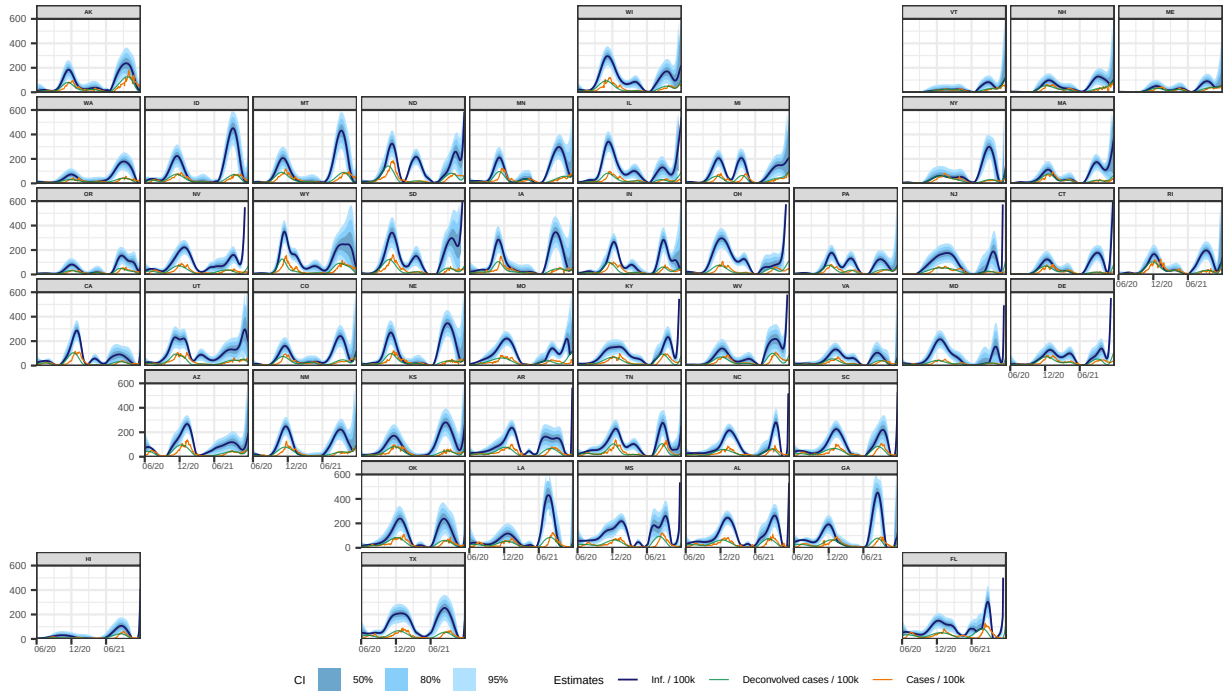

Figure 2: Estimates of the number of daily new infections per 100,000 population for each U.S. state from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% confidence intervals for the estimates, while the teal line represents the number of new daily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

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 separates their estimated deconvolved cases based on the circulating variant proportions at the time. These figures show times when the total infections and the deconvolved cases broken down by the variant categories emphasize earlier outbreaks than would be indicated by cases alone. For example the major Ancestral wave in California, Maryland, Idaho, Montana, and Ohio peaks earlier for infections than cases. Such trends are similar with Delta, though more obviously in Louisiana, Idaho and Montana than California, Maryland and Ohio. The division by variant categories reveals 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 construction (as they are for deconvolved cases which are re-scaled to get the infection estimates). ATTN: These last two sentences are not sufficiently direct: they need to say something interesting, without too much mealy-mouthedness.

2.4 The relationship between infections and hospitalizations is messy.

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 0.51, occurring at a lag of 13 days. In contrast, we find that the greatest 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. The maximum correlation at a lag of 13 days is in keeping with estimates of the average time from infection to hospitalization for cases reported in January, 2020 in Wuhan, China (9.7 days) as well as with estimates from across the pandemic in the UK (rangeing from 8.0 to 9.7 days) [12]. Importantly, the 13 day lag in the U.S. also includes the impact of the reporting pipeline, a delay omitted from the international estimates.

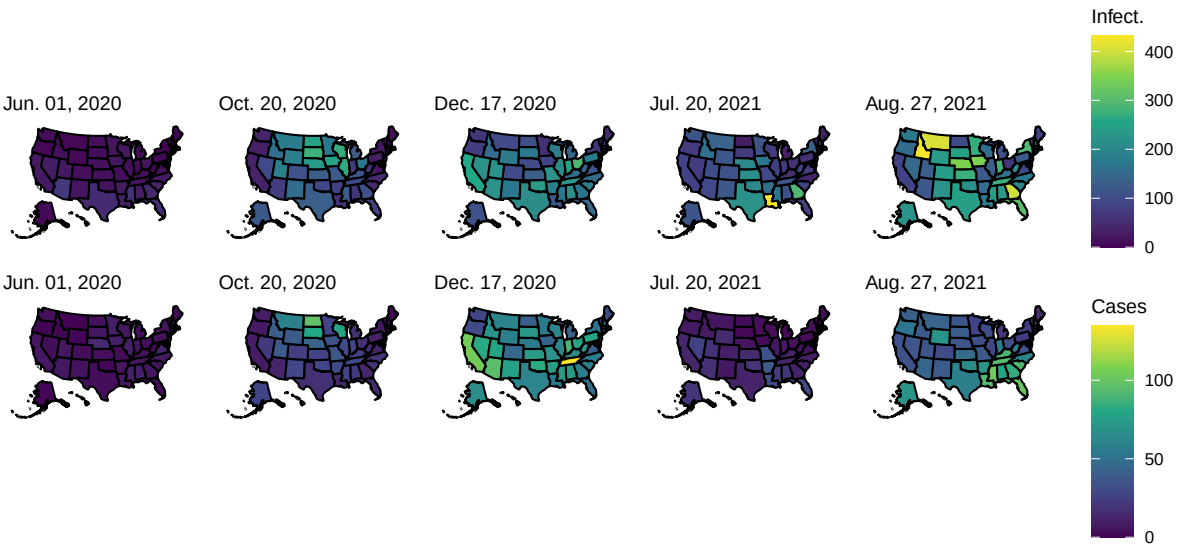

Figure 3: Choropleth maps of the state-level estimates of the number of daily new infections per 100,000 population (top row) and the daily new cases per 100,000 population (bottom row) for five dates between June 1, 2020 to November 29, 2021. Note that the first date was chosen as a baseline, while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

While both the infections and deconvolved cases are leading indicators of hospitalizations and their trajectories are similar, the average correlation they attain is different. In particular, the correlation is larger for deconvolved cases than for infections (with a difference of about 0.18 at the peaks). The increase is likely due to 2 issues. First, many cases are detected contemporaneously with hospitalization: people first test positive once they go to the emergency room for treatment. Second, unreported infections tend to be less severe and less likely to lead to hospitalization than those that are reported [13].

2.5 Infection-hospitalization ratios are smaller and more stable.

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 noiser relative to IHRs. This supports our 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 above. Namely, states that are proximate (for example, Ohio, Pennsylvania, and Virginia) tend to exhibit similar temporal patterns in IHRs and CHRs. In addition, similar spikes are evident across many states during waves of infections that are driven by variants of concern. For example, many states exhibit a striking increase in hospitalizations in mid-2021, coinciding with the rapid takeover of the Delta variant [14]. This finding aligns with previous studies that found an increased risk in hospitalization due to Delta [15,16]. Similarly, during the fall of 2020 a spike in IHRs rivals or surpasses that observed during the time of Delta (which is the case for states like New York or Wyoming). ATTN: what is the "Similarly" referring to here?

Overall, the relationship between infections and hospitalizations is complicated. We observe intermittent spikes that punctuate longer periods where the IHRs are relatively stabile, remaining below 0.1 hospitalizations

Figure 4: Top panel: Reported cases, deconvolved cases, and estimates of daily new infections (dark blue line) per 100K 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 100K inhabitants.

per infection. While we computed and compared CHRs and IHRs for all states, it is important to note that

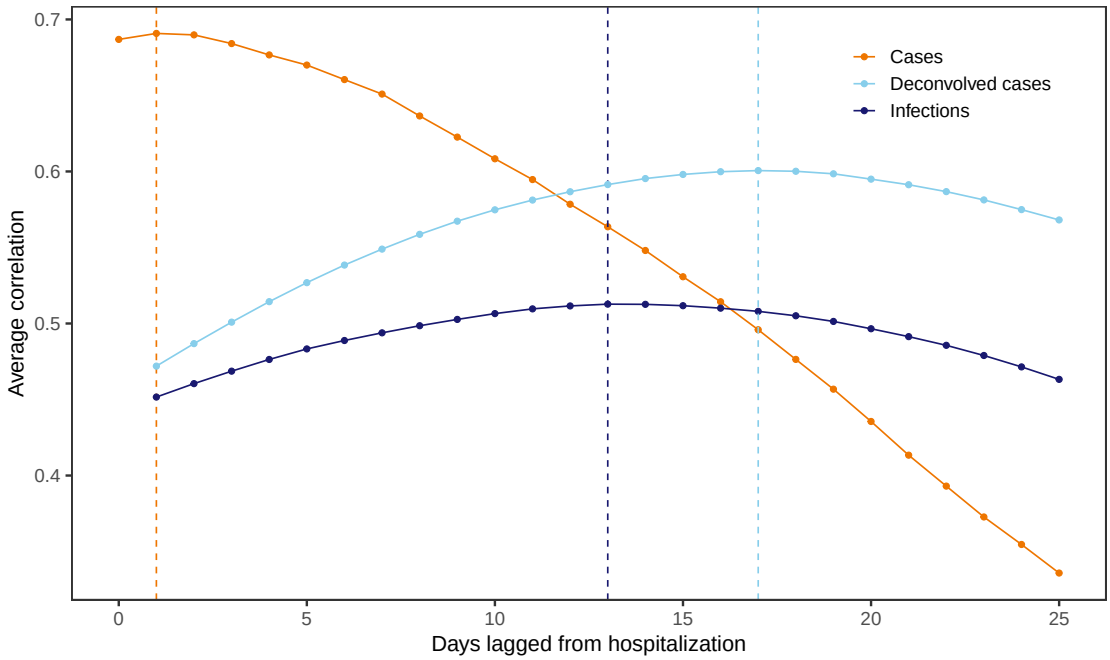

Figure 5: Spearman's correlation between each of cases, deconvolved cases, and infections with hospitalizations per 100,000. These are calculated for each lag, state and rolling window of 61 days before averaging. The vertical dashed lines indicate the lags for which the highest average correlation is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

both likely 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].

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

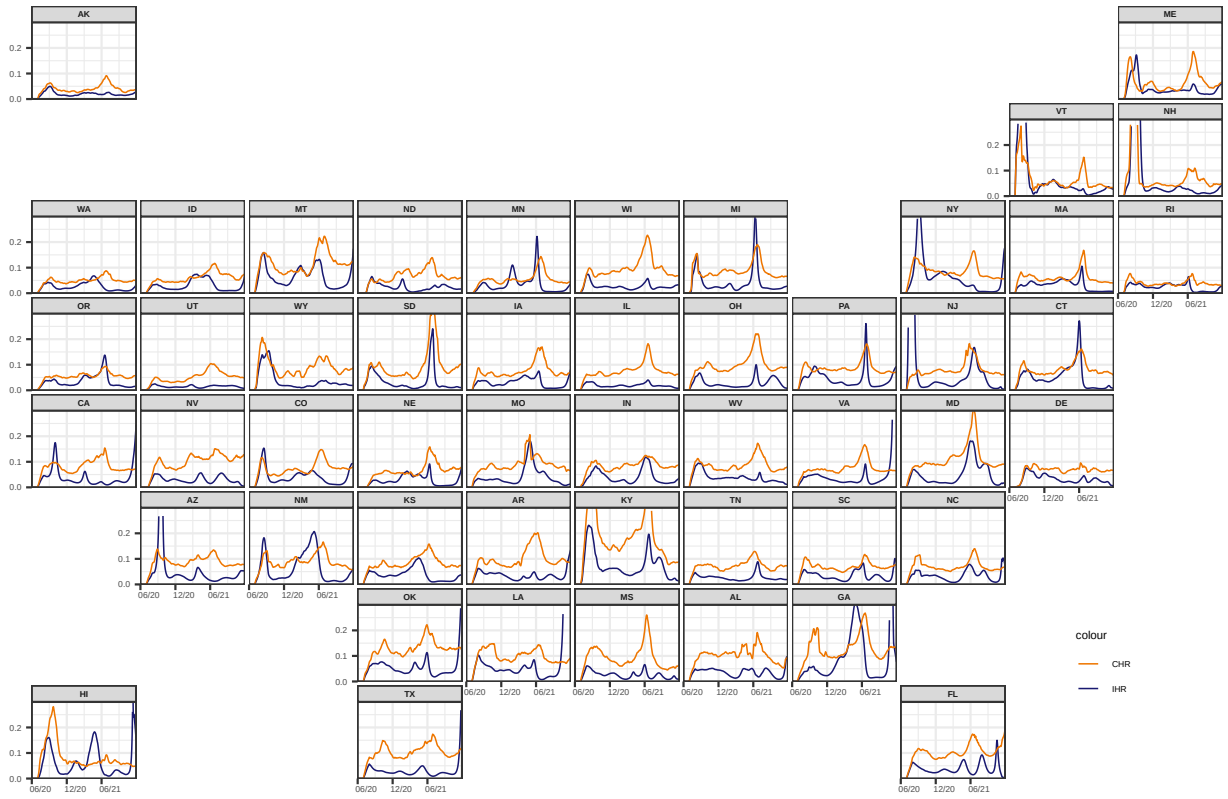

Figure 6: Time-varying IHR and CHR estimates for each state from June 1, 2020 to November 29, 2021, obtained using the corresponding optimal lag from the systematic lag analysis. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious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

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 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 end date of November 29, 2021, and ultimately decided to not continue into the Omicron wave is because the Omicron variants come with substantial increases in the risk of reinfection in comparison to previous variants, likely due to increased immune escape [27–29]. So having reinfection data that is representative of each location under study is crucial for extending the analysis. 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 additional issues that inhibit us from pushing the period of analysis past December 2021. While most state-level data suggests that reinfections still 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 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 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.

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 Section S1.6.

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 $s \in [t - 75 + 1, t + 60] = \mathcal{S}_t$

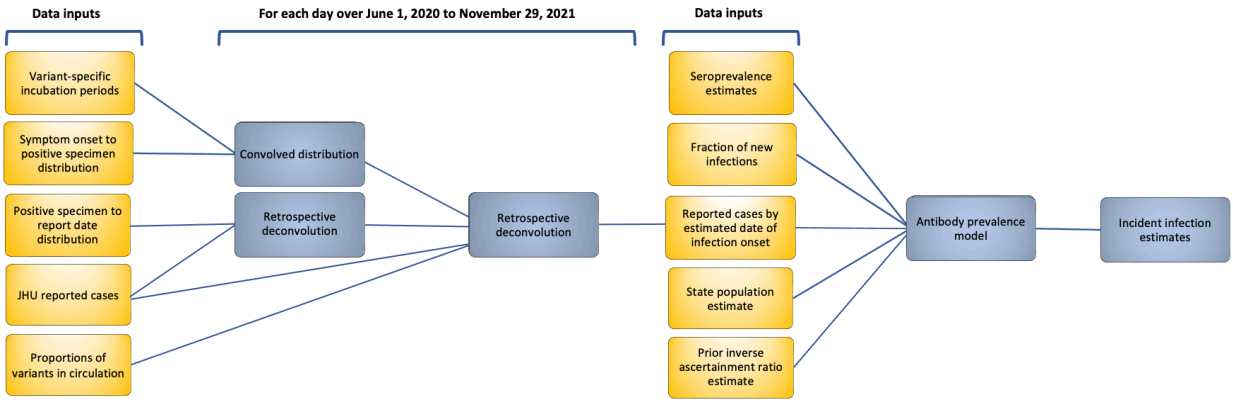

Figure 7: Flowchart of the inputted data and major analysis steps required to get from reported cases to incident infection estimates for each day over June 1, 2020 to November 29, 2021 for a state. Data sources are coloured in yellow, while data analysis steps are coloured in blue. The data sources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

and positive specimen date greater than $s - 60$. Then, we first compute

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

Next we compute a similar national quantity $\tilde{p}_t(k) = \frac{1}{N_t} \sum_{s \in \mathcal{S}_t} \big(\#\ z_s$ 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 Section S1.7.

4.2 Estimating the incubation period distributions

To account for the incubation period, the time between infection and symptom onset, we use estimates from the existing literature, modified slightly for coherence with each other: we model each incubation as a gamma distribution with different parameters. We focus on the following eight variants, which dominated at various points during our study period: Ancestral, Alpha, Beta, Epsilon, Iota, Gamma, Delta, and Omicron. Alpha, Beta, Delta, Gamma, and Omicron are all variants of concern[40], while we include the Epsilon (California) and Iota (New York) variants because of large impact on those and neighbouring states[41,42].

The Ancestral variant has been modelled as a gamma distribution[43], so we simply use those reported 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[44–46]. To match these moments to the gamma distribution, we solve the same moment equations described in Section 4.1. Then, we discretize each resulting density to the support set, which is taken to be from 1 and 21 days. This range assumes that symptoms require at least 1 day to develop[47] and that an asymptomatic infection will resolve within 21 days[48,49].

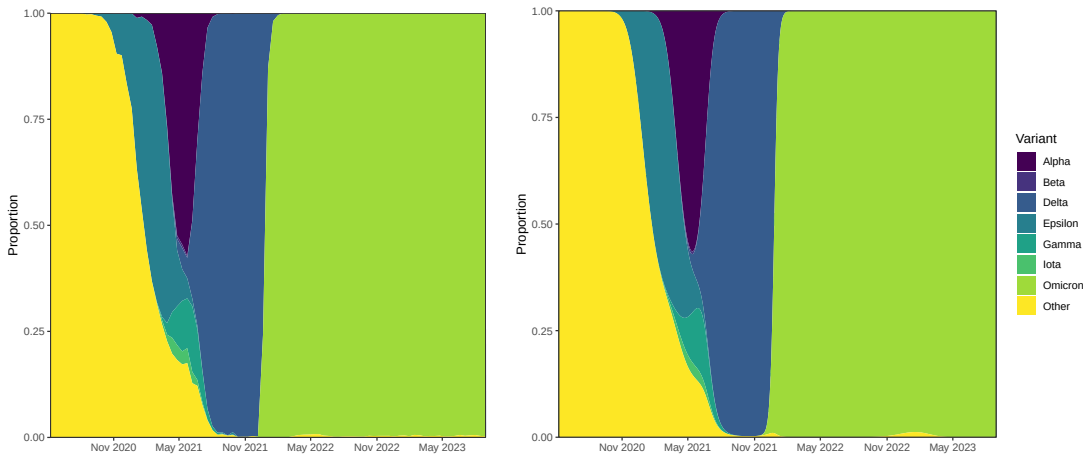

Figure 8: Left: Original biweekly proportions of the variants in circulation for California. Right: Daily proportions of the variants in circulation for California.

We were unable to locate incubation period estimates for the geo-specific Epsilon and Iota variants, so we use the incubation period for Beta because Epsilon, Iota, and Beta are all children from the same parent in the phylogenetic tree of the Nextstrain Clades[14]. 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.

ATTN: Perhaps plot the incubation periods? Maybe sharing a panel with the circulation proportions below.

4.3 Variant circulation proportions

To estimate the daily proportions of the variants circulating in each state, we obtain the GISAID genomic sequencing data from CoVariants.org[14,50]. These counts represent the total number of cases belonging to a particular variant using a sample of positive tests over a biweekly period. To estimate the population proportion of each variant, we apply multinomial logistic regression for the eight variant categories separately for each state.

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

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

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 8 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 Section 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.

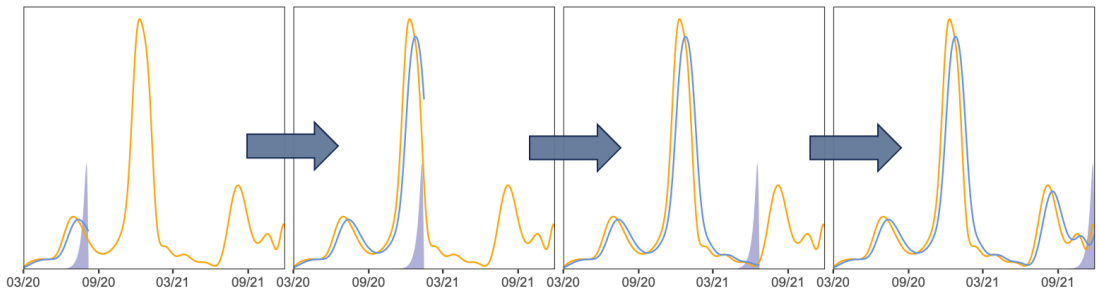

Figure 9: A general depiction of convolving smoothed cases (orange line) with the corresponding delay probabilities (shaded blue area) to get the convolved estimates (blue line) over four different times.

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).

We will start by describing the first type of deconvolution performed from report to positive specimen date in detail. For this problem, let $t = 1, \ldots, T'$ index the extended deconvolution period from March 1, 2020 to March 1, 2023, extended to minimize the effects of boundary issues. 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 The COVIDcast API[37]. Recall that $\widehat{\pi}_{\ell,t}(k)$ is the associated probability that these reported cases were collected k days earler.

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

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

ATTN: These sums aren't quite right: the outer t makes x_{t-k} run negative? Actually this wouldn't be a problem, except that we don't deconvolve before $t = 1$, we renormalize the delays. Perhaps too much detail, but the math annoys me. Ideas? The two parts of this optimization problem trade data fidelity (the sum of squared errors) with smoothness in the resulting estimates (the absolute error of the 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 easily[53]. We select λ with 3-fold cross validation[23] 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 estimate, 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 Section 4.1, denoted by $\{q_{\ell,t}(k) : -3 \leq k \leq 21\}$, with the variant-specific incubation periods from Section 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 Section 4.3. The analogous optimization problem is therefore:

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

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 will 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 infection[54], 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:

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

where $\widehat{u}_{\ell,t}$ is deconvolved cases, $z_{\ell,t}$ is the percentage of reinfections, $a_{\ell,t}$ is the inverse reporting ratio, and $\epsilon_{\ell,t}$ represents noise. Note that γ_ℓ is the percentage 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 measured using various surveys.

For the proportion of the population in each state with evidence of previous infection across time, we use two major seroprevalence surveys: the 2020–2021 Blood Donor Seroprevalence Survey and the Nationwide Commercial Lab Seroprevalence Survey[55,56]. See Section S1.8 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 District[30]. These results are broadly similar to those in other locations with available data[30–33].

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,

$$r^1_{\ell,m} = s_{\ell,m} + \tau_{\ell,m}, \qquad \tau_{\ell,m} \sim \mathrm{N}(0, w^1_{\ell,m}\sigma^2_{\ell,r}), \tag{6}$$

$$r^2_{\ell,m} = s_{\ell,m} + \varphi_{\ell,m}, \qquad \varphi_{\ell,m} \sim \mathrm{N}(0, w^2_{\ell,m}\sigma^2_{\ell,r}), \tag{7}$$

$$s_{\ell,m+1} = (1-\gamma_\ell)s_{\ell,m} + a_{\ell,m}z_{\ell,m}\widehat{u}^{\Sigma}_{\ell,m} + \epsilon_{\ell,m}, \qquad \epsilon_{\ell,m} \sim \mathrm{N}(0, \sigma^2_{\ell,\epsilon}), \tag{8}$$

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

$$a_{\ell,m+1} = 3a_{\ell,m} - 3a_{\ell,m-1} + a_{\ell,m-2} + \eta_{\ell,m}, \qquad \eta_{\ell,m} \sim \mathrm{N}(0, \sigma^2_\eta). \tag{9}$$

This antibody prevalence model is a state-space model with latent variables $\mathbf{s}_\ell$ and $\mathbf{a}_\ell$ and unknown parameters γ_ℓ, σ^2_r, σ^2_ϵ, and σ^2_η. 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 Section S1.2.

4.6 Lagged correlation to hospitalizations and time-varying IHRs

From The COVIDcast API[37], 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 $\widehat{\mathbf{u}}_\ell$ 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

Code availability

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] 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).

[13] 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).

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

[15] 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).

[16] 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).

[17] Russell, C. D., Lone, N. I. & Baillie, J. K. Comorbidities, multimorbidity and COVID-19. *Nature Medicine* **29**, 334–343 (2023).

[18] Fox, S. J. *et al.* Disproportionate impacts of COVID-19 in a large US city. *PLOS Computational Biology* **19**, e1011149 (2023).

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

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

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

[22] 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).

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

[24] 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).

[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] 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).

[28] Pulliam, J. R. *et al.* Increased risk of sars-cov-2 reinfection associated with emergence of omicron in south africa. *Science* **376**, eabn4947 (2022).

[29] 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).

[30] Ruff, J. *et al.* Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. *Emerging Infectious Diseases* **28**, 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] McManus, O. *et al.* Predicting COVID-19 incidence using wastewater surveillance data, Denmark, October 2021–June 2022. *Emerging Infectious Diseases* **29**, 1589 (2023).

[35] Hart, O. E. & Halden, R. U. 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] Li, X. *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] Reinhart, A. *et al.* An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences* **118**, 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] 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).

[42] 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).

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

[44] 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).

[45] 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).

[46] 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).

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

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

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

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

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

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

[59] 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).

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 dataset used or estimation methodology

S1.1 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).

S1.2 State space representation of the antibody prevalence model

The antibody prevalence model from Equation 6 is conceptualized as a Gaussian state space model (as in[57,58]).

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

$$y_t = Z\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, H_t) \tag{10}$$

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

where $\alpha_1 \sim N(a_1, P_1)$ and there is independence amongst α_1, ϵ_t and η_t[57,58]. 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[57].

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[57]. 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,

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

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

$$\hat{a}_t = \mathbb{E}[\alpha_t \mid y_n, \ldots, y_1] \tag{12}$$
$$V_t = \text{Var}(\alpha_t \mid y_n, \ldots, y_1). \tag{13}$$

The filtering and smoothing steps are based on recursions that are described in Appendix A of **(author?)** [58] 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:

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

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

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

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 **(author?)** [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 $\sigma_s^2 = 3 \times 10^{-6}$,$\gamma = 0.99$, 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[57], 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 S1.3 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.

S1.3 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

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

S1.4 Ablation analysis of infection-hospitalization correlations

We undertake an ablation study for the lagged correlation of infections, the results of which are shown in Figure 5. 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.

S1.5 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[59]. Unless otherwise specified, we use the July 1, 2020 estimates.

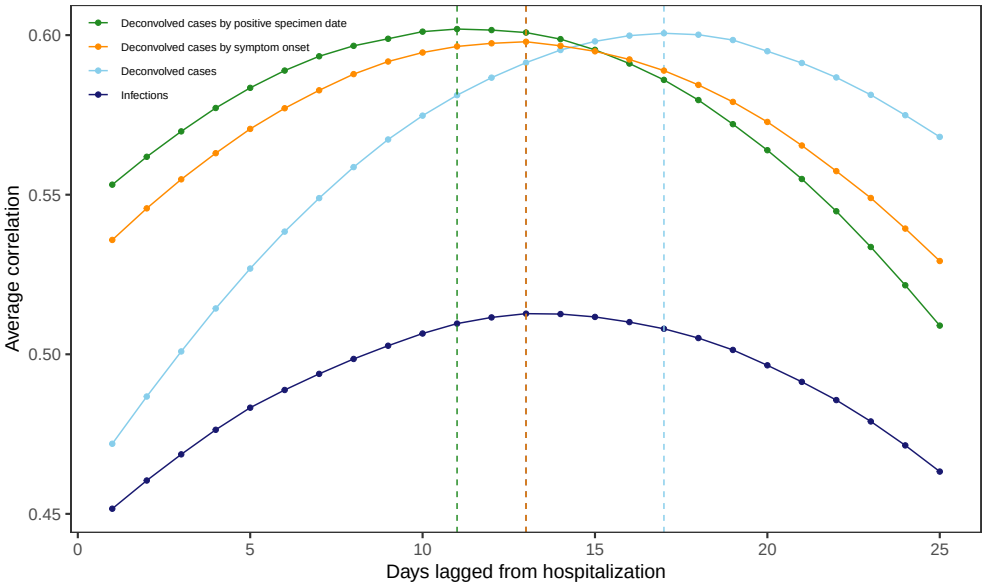

Figure S1: Lagged Spearman's correlation between the infection and hospitalization rates per 100,000 averaged for each lag across U.S. states and days over June 1, 2020 to November 29, 2021, and taken over a rolling window of 61 days. The infection rates are based on the counts for the deconvolved case 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.

S1.6 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 **(author?)** [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 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 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% [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.

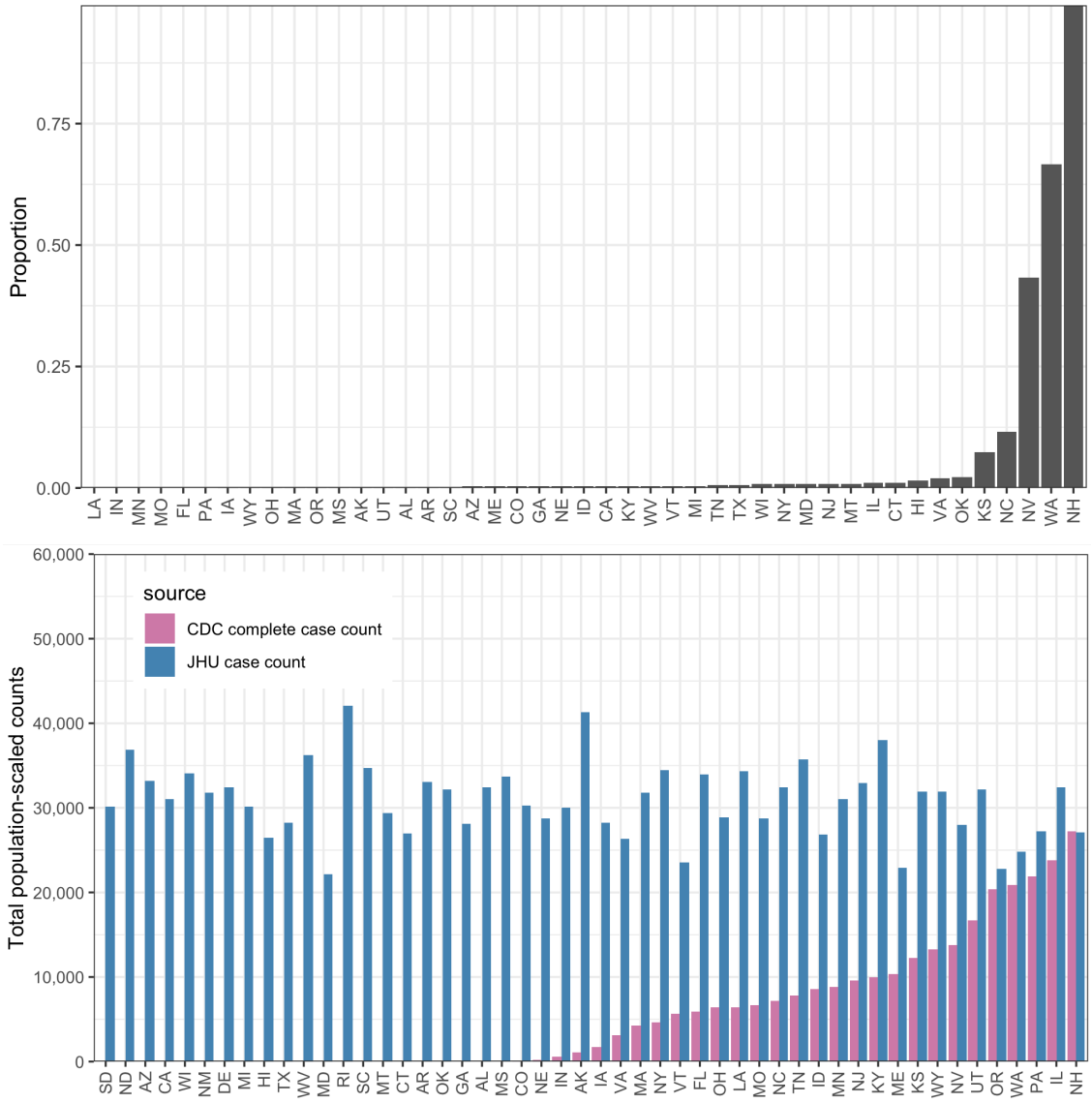

Figure S2: Top panel: Proportion of complete cases with zero delay between positive specimen and report date in the restricted CDC line list dataset. Bottom panel: Complete case counts by state in the CDC line list versus the cumulative complete case counts from JHU CSSE as of June 6, 2022. All counts have been scaled by the 2022 state populations as of July 1, 2022 from **(author?)** [59].

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).

S1.7 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]):

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

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 **(author?)**[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.

S1.8 Details about seroprevalence data

In the former, the CDC collaborated with 17 blood collection organizations in the largest nationwide COVID-19 seroprevalence survey to date[55]. The blood donation samples were used to construct monthly seroprevalence estimates for nearly all states from July 2020 to December 2021[?] . 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-19,[?]). 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? ?] . 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 S3). 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. In addition, this blood donor dataset lacks measurements for any states in 2022 (as the corresponding survey ended in December 2021). Finally, as can be seen from Figure S3, the final commercial seroprevalence measurement from 2022 shows a large increase relative to the immediately preceding measurement for each state. Since such an increase may signal unreliability or instability of the final estimates, we decided to remove them from our analysis. Note that North Dakota is the only state to which this exclusion does not apply as there are no commercial seroprevalence measurements beyond 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.

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.

S1.9 Ablation study for the lagged correlation analysis

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).

S1.10 Possible investigation of reinfections

Possible change to the paper based on Ajitesh's feedback - Main contribution is the model, shows without reinfection & here's an extension that shows how to include reinfection data.

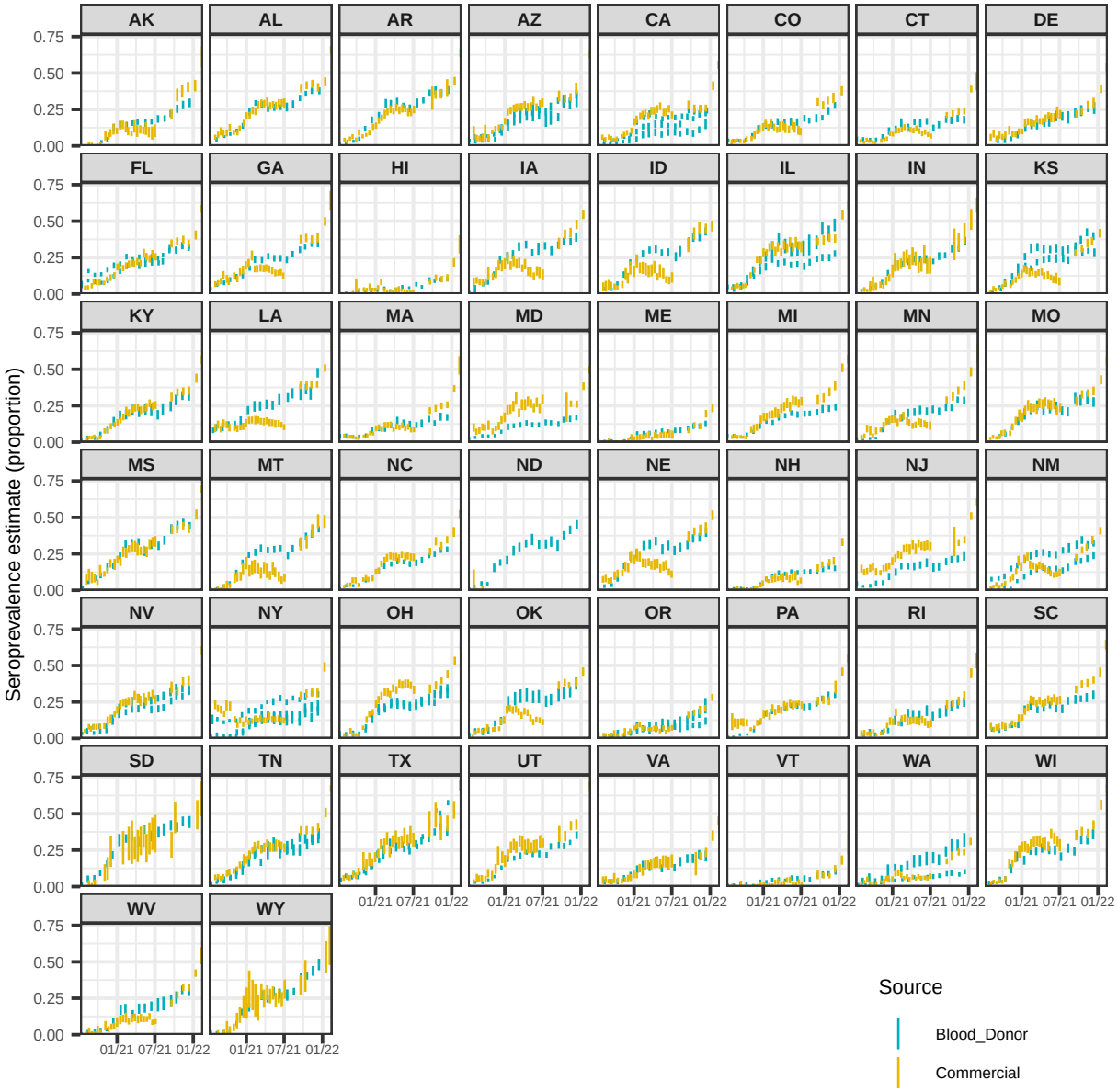

Figure S3: A comparison of the seroprevalence estimates from the Commercial Lab Seroprevalence Survey dataset (yellow) and the 2020–2021 Blood Donor Seroprevalence Survey dataset (blue). Note that the maximum and the minimum of the line ranges are the provided 95% confidence interval bounds to give a rough indication of uncertainty. ATTN: This figure doesn't use space very well. Let's remove the gap between panels and make the facet labels (the state names) normal sized. The x-axis could just show 2021, 2022 (rather than so many ticks). Alternatively, another map layout? And I don't think "Rate" is correct for the y-axis.