Retrospective estimation of latent COVID-19 infections over the pandemic in US states

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

Version: March 12, 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. Nevertheless, accurate estimates of latent COVID-19 infections can improve our understanding of the true size and scope of the pandemic and provide an indication of disease patterns and burden over time. Therefore, we estimate daily incident infections for each U.S. state. Our methods first deconvolve reported COVID-19 cases to their infection onset. We then use a serology-driven model to scale these deconvolved cases to unreported infections. Unlike existing approaches, our approach is state-specific, incorporates several variant-specific incubation periods, and accounts for reinfections. From its application to the gold standard case data, we find a disease burden that appears earlier and more extensively than indicated by cases alone. In addition, we observe similar epidemic patterns in surges and periods of waning observed in clusters of neighbouring states. Our findings help to better understand the impact of the pandemic in the U.S. at the level of the state.

1 Introduction

Reported COVID-19 cases are a staple in tracking the pandemic at varying geographic resolutions such as national, state and county levels (Dong et al., 2020; The New York Times, 2020; The Washington Post, 2020). Yet, for every case that is eventually reported to public health, several infections are likely to be missed. To see why, it is important to understand who's cases are being reported and what differentiates them from the unreported cases. Refer to Figure 1 for 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 likely to be missed (Centers for Disease Control and Prevention, 2022). In addition, testing practices, availability, and uptake vary across space and time (Pitzer et al., 2021; European Centre for Disease Prevention and Control, 2020; Hitchings et al., 2021). 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 submission to public health (Pellis et al., 2021; Washington State Department of Health, 2020). 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 is no large-scale surveillance effort in the United States that reliably tracked symptom onset, let alone infection onset, ascertaining the onset of 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 challenging because the true spatial and temporal behaviour is unknown. While reported cases provide some understanding 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 estimate the true extent of its effect—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 public and professional understanding of the pandemic burden over space and time.

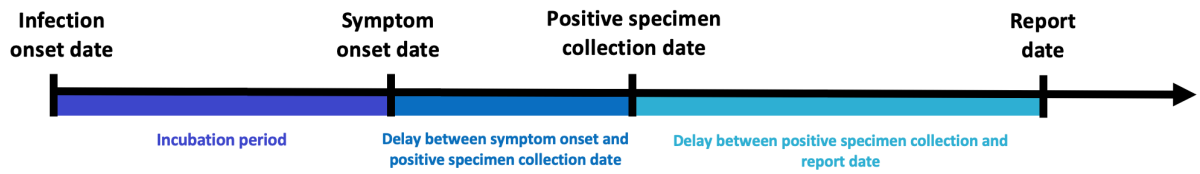

Figure 1: Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

In this work, we provide a statistically justified reconstruction of daily incident infections for each U.S. state from June 1, 2020 to November 29, 2021. We achieve this by breaking the task of estimating infection onset from report date into the more manageable parts of estimating the time from infection to symptom onset, symptom onset to positive specimen collection, and the time from positive specimen collection to report date (as depicted in Figure 1). 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 then use variant-specific incubation period distributions in conjunction with these delay distributions to deconvolve daily reported COVID-19 cases back to their infection onset. The resulting infection estimates are adjusted to account for the unreported infections by using seroprevalence data in a novel antibody prevalence model that is defined by its ability to account 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 apply our infection estimates to get 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 these infection estimates, we believe that there is much more to be explored, and we hope that our work will prove an important benchmark for others to undertake retrospective analyses.

2 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. We start with a brief introduction to each data source used and follow this with a description of each major analysis task in the order they are performed. Figure 2 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.

2.1 Data

We obtain literature estimates of the variant-specific incubation periods for the following eight variant categories: Ancestral, Alpha, Beta, Epsilon, Iota, Gamma, Delta, and Omicron. For the Ancestral variants, we use the literature estimates of the gamma distribution parameters (Tindale et al., 2020). 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. (2022); Grant et al. (2022); Ogata et al. (2022).[1] Since the literature

[1]To clarify, we use the estimates for Alpha and Omicron from Tanaka et al. (2022), those for Beta and Gamma from Grant et al. (2022), and those for Omicron from Ogata et al. (2022).

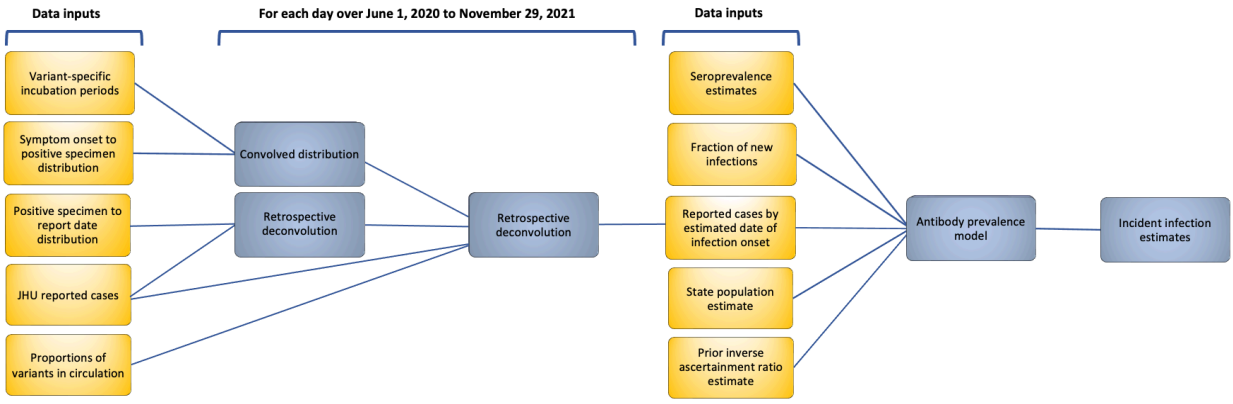

Figure 2: Flowchart of the inputted data and major analysis steps required to get from reported cases to incident infection estimates for each day over June 1, 2020 to November 29, 2021 for a state. Data sources are coloured in yellow, while data analysis steps are coloured in blue. The data sources that do not stem from an analysis step are literature estimates.

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 Hodcroft, 2021).

To estimate the daily proportions of the variants circulating in each state, we obtain the GISAID genomic sequencing data counts from CoVariants.org (Hodcroft, 2021; Elbe and Buckland-Merrett, 2017).[2] 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.

The COVIDcast API (Reinhart et al., 2021) 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, Dong et al., 2020). 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.

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 (Centers for Disease Control and Prevention, 2020b,c), the restricted dataset[3] is selected because it contains information on the state of residence which is essential for constructing state-specific delay distributions. 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.

In this dataset, the three key variables of interest are the dates of symptom onset, positive specimen date and report to the CDC. 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, we faced the fundamental issue that Jahja et al. (2022) described, in which cases with missing report dates may be filled with their symptom onset date. On top of this, we have the additional factor of positive specimen date to contend with. Previous correspondence with the CDC confirms that the imputation issue extends to this variable as well. So it is possible that all three variables may be imputed with the same date for a case. However,

[2]The complete list of EPI_SET Identifiers that were used to produce the CoVariants data are provided in the Acknowledgements section of their website (Hodcroft, 2021).

[3]The CDC does not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.

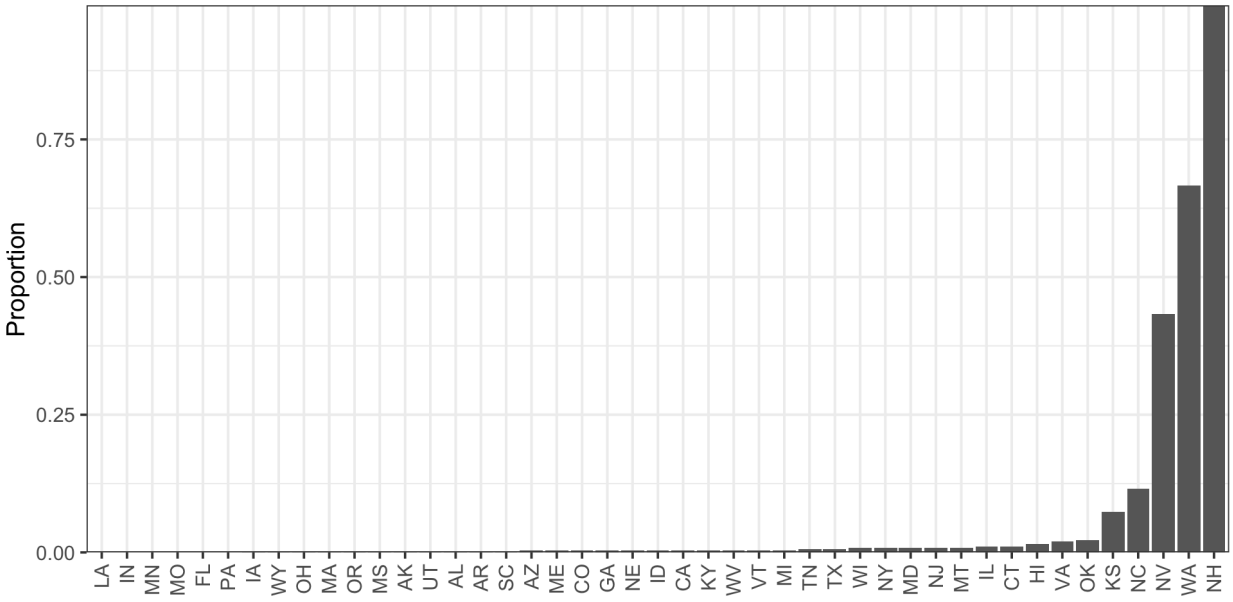

Figure 3: Proportions of complete cases with zero delay between positive specimen and report date in the restricted CDC line list dataset.

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 3 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. We explain how we incorporated these variations in the ordering of events into our analysis in Section 2.3.

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 4, from which it is clear the fraction of missing cases is substantial for many states, often surpassing 50% (Jahja et al., 2022). 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

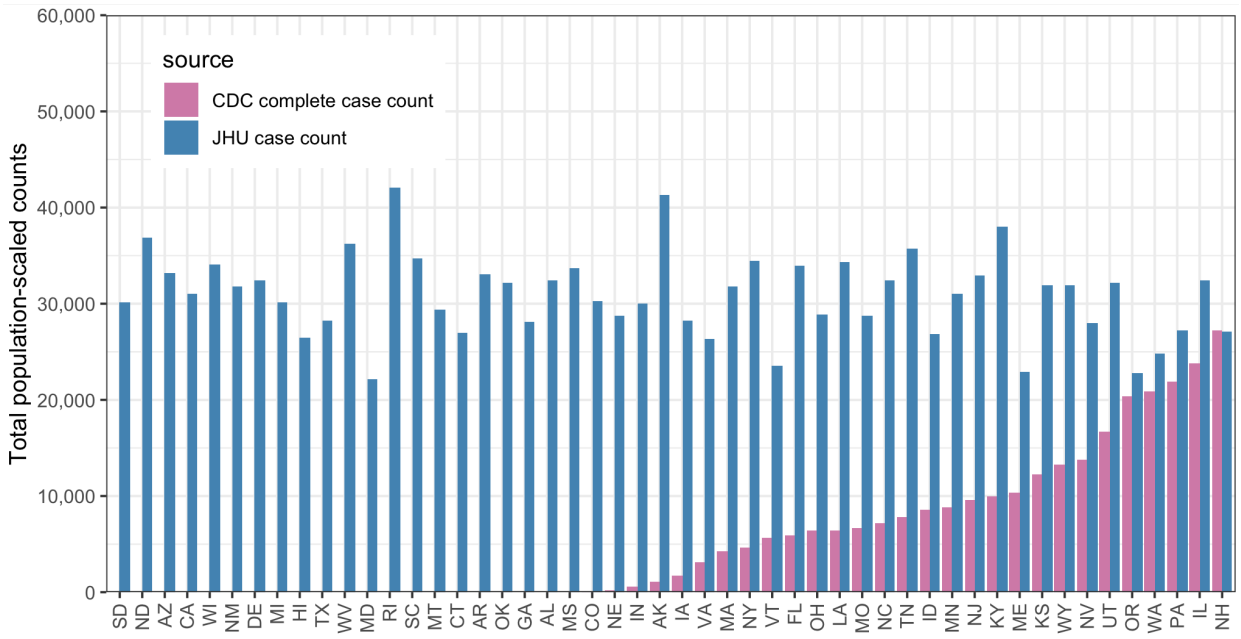

Figure 4: Complete case counts by state in the CDC line list versus the cumulative complete case counts from JHU CSSE as of June 6, 2022. All counts have been scaled by the 2022 state populations as of July 1, 2022 from U.S. Census Bureau, Population Division (2022).

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).

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 Nationwide Commercial Lab Seroprevalence Survey (Centers for Disease Control and Prevention, 2021a,b). In the former, the CDC collaborated with 17 blood collection organizations in the largest nationwide COVID-19 seroprevalence survey to date (Centers for Disease Control and Prevention, 2021a). The blood donation samples were used to construct monthly seroprevalence estimates for nearly all states from July 2020 to December 2021 (Jones et al., 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, Bajema et al., 2021). 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 (Bajema et al., 2021; Centers for Disease Control and Prevention, 2020d; Jones et al., 2021). 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 5). Such limitations indicate that reliance upon only one seroprevalence survey is inadvisable. For example, in the commercial dataset, the last estimate for North Dakota is in September 2020. In the blood donor dataset, Arkansas does not have estimates available until October 2020.

The date variables that come with the two seroprevalence datasets are different and so the date variables that we are able to construct from them are not the same. For the commercial dataset, we use the midpoint of the provided specimen collection date variable. A major difference in the structure of the two datasets is that the commercial dataset always has the seroprevalence estimates at the level of the state, while the blood

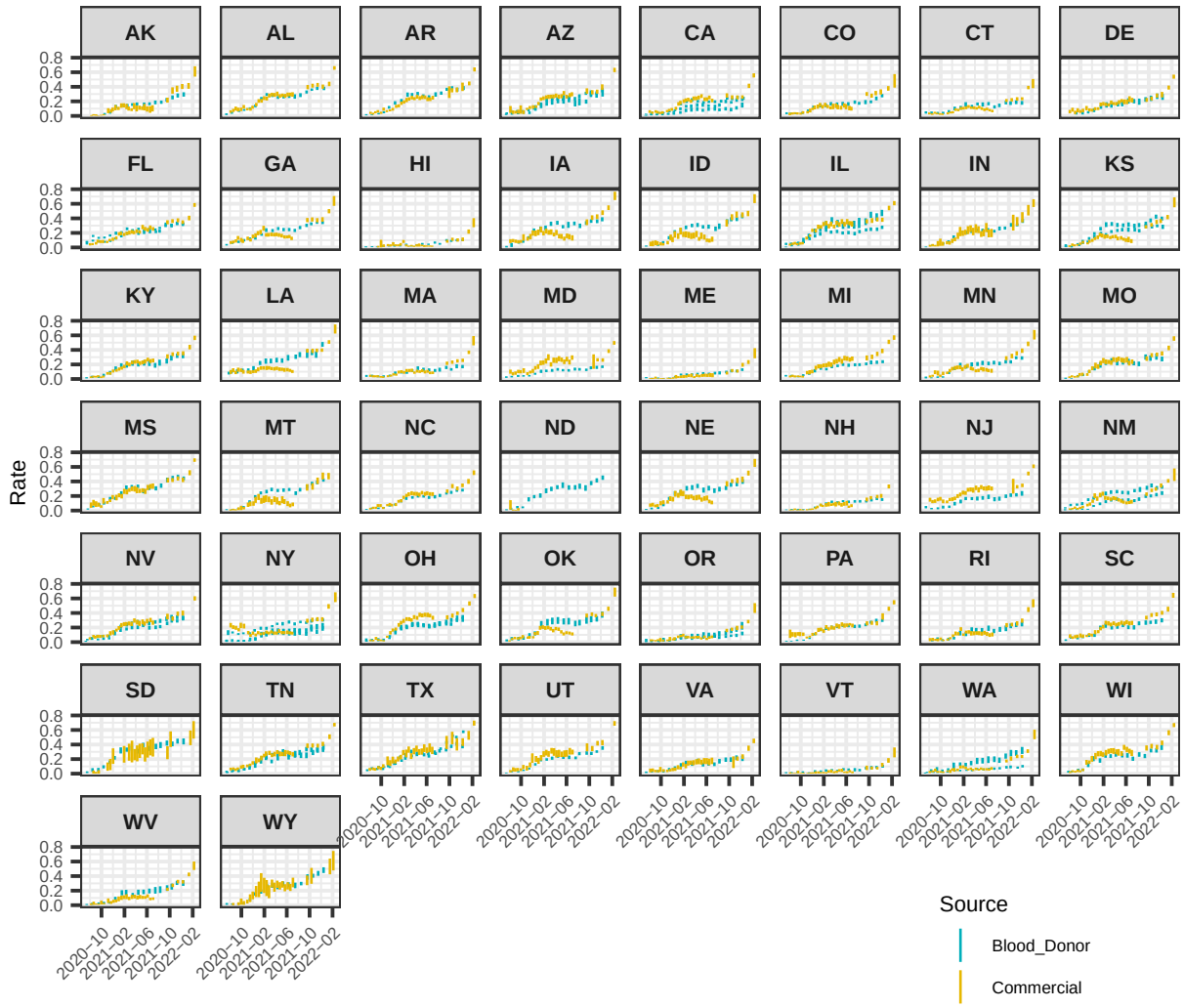

Figure 5: A comparison of the seroprevalence estimates from the Commercial Lab Seroprevalence Survey dataset (yellow) and the 2020–2021 Blood Donor Seroprevalence Survey dataset (blue). Note that the maximum and the minimum of the line ranges are the provided 95% confidence interval bounds to give a rough indication of uncertainty.

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.

For adjusting our infection counts, 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 (U.S. Census Bureau, Population Division, 2022). Unless otherwise specified, we use the July 1, 2020 estimates.

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. (2022). 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.

2.2 Estimating the delay from positive specimen to report date

We use the restricted CDC line list to estimate the distribution between positive specimen and report for each state at each time. To formalize this, 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 Jahja et al., 2022):

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

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 Jahja et al. (2022), 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 $k = 1, \ldots, 60$,

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

For each state, we estimate the positive specimen to report date distribution at each t by using the empirical distribution of all non-zero lags between the complete cases whose positive specimen dates fall in the interval around t designated by $[t - 75 + 1, t + 60]$).

Now, the task of estimating the positive specimen to report date distribution for each state at each time requires four distinct steps. First, we obtain the empirical distribution of all lags (excluding zero) from all cases with positive specimen dates falling in the roughly center-aligned interval. Next, we weight the state-specific empirical distribution by the proportion of CDC cases to JHU cases. That is, we compare the number of CDC cases used to create the empirical distribution to the number of cases reported by JHU in the time window of $[t - 60 + 2, t + 75]$ (to correspond appropriately to the interval for the CDC cases). This proportion is used as the weight for the state's empirical distribution, while the complement is used to weight the overall empirical distribution that is formed from the data for all states. This construction allows for more reliance on the state's distribution when there are more CDC cases relative to JHU (and vice versa). After implementing the shrinkage method, we fit a gamma density to the resulting empirical distribution by the method of moments. Finally, we discretize the resulting density to the support set of 1 to 60 days.

2.3 Estimating the delay from symptom onset to positive specimen date

The task of estimating the delay from symptom onset to positive specimen date follows the same procedure as for positive specimen to report date with a few key data-driven amendments. Firstly, we allow for negative

delays to account for the possibility that an individual case tests positive before symptom onset. Under this assumption, the lower and upper bounds for the support of -3 and 21 are chosen based on the largest delay values for the statewise 0.05 and 0.95 quantiles that are not outliers. Secondly, we allow for zero delay because the median statewise delay is very short at approximately 2 days. 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.

2.4 Estimating the incubation period distributions

One incubation period distribution is estimated for each variant under consideration. The variants we consider are Alpha, Beta, Gamma, and Delta, which are included because they are are designated to be variants of concern by WHO based on their potential to cause new waves, dethrone the dominant variant, and lead to changes in public health policy (World Health Organization, 2021). In addition, we include the Epsilon (California) and Iota (New York) variants because of their impact on those and the surrounding states (Yang et al., 2022; Duerr et al., 2021). 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 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 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. In other words, those are taken to be the lower and upper limits for the number of days that the virus could be incubating in someone. 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 Public Health Agency of Canada, 2021). The assumption underlying the upper bound is that 21 days is the maximum number of days that the virus could be incubating in someone (which is reasonable based on Zaki and Mohamed, 2021 and Cortés Martínez et al., 2022).

2.5 Convolutional estimates of the infection to positive specimen distributions

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.

2.6 Retrospective deconvolution

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 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 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

$$\underset{x}{\text{minimize}} \sum_{t \in \mathcal{T}} \left(y_t - \sum_{k=1}^{d} \hat{p}_t(k) x_{t-k} \right)^2 + \lambda \| D^{(4)} x \|_1.$$

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.

We solve this trend-filtering-regularized least squares deconvolution problem by employing the ADMM algorithm from Ramdas and Tibshirani (2016) that is described in Appendix A of Jahja et al. (2022). The solution to the problem is an adaptive piecewise cubic polynomial (Tibshirani, 2014, 2022).

We select the tuning parameter, λ, by using 3-fold cross validation similar to Jahja et al. (2022) in which every third infection count is reserved for testing. The tuning parameter that results in the smallest mean squared error is selected.

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.

2.7 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 (Centers for Disease Control and Prevention, 2020d), 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 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.

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

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

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^2_ϵ. Details of this methodology and the computation of the associated uncertainty measurements are deferred to the Supplementary Materials S1.2.

2.8 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.

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, the number of individuals who are hospitalized due to COVID-19 on a day are divided by the estimated total number who were infected on the lagged number of days before.

2.9 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).

3 Results

This work estimates incident infections for each U.S. state over June 1, 2020 to November 29, 2021 and to illustrate the disease burden and viral transmission dynamics at the level of the state across time. After converting the number of infections to rates (infections per 100,000 population), we perform a brief comparison between infection and case estimates within each state to see to what extent that surges in infections are evident in cases alone and point out instances where cases largely fail to capture surges in infections. To apply these infection estimates, we perform a lagged correlation analysis with hospitalizations and then compute the time-varying IHRs for each state. Finally, we look at infections within and across the states and point out patterns related to the major variants and geographical contiguity.

3.1 Infection estimates compared to reported cases

Naturally, outbreaks in infections precipitate those in cases and are reliably larger in magnitude (Figure 6). Hence, our infection estimates indicate that the pandemic had a differential impact across states earlier and at a larger scale than is suggested by cases. From the choropleth maps that compare the state-level rates of daily new infections and cases per $100,000$, we can observe that for the earliest time of June 1, 2020, there is little discrepancy between case and infection rates, while for the later times there are immense differences in the rates, such that case rates tend to underrepresent infections to a great extent (Figure 7).

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, take the wave in the spring of 2021 for North Dakota and South Dakota, where the relatively flat and steady case cadence fails to capture the major wave of infections. With respect to variants of concern, consider the late 2020 Ancestral wave for the midwestern states of Illinois, Indiana, and Ohio. 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 (Figure 6). 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 (Ontario Agency for Health Protection and Promotion, 2022; Garrett et al., 2022).

Finally, while the main Delta wave is somewhat evident from the case counts for all states (Figure 6), our estimates suggest that case counts tend to severely underestimate infections during this time for many states. The lowest of all states was in New Jersey, where about 4.6% (95% confidence interval: $[1.9, 67.7]$) of the estimated infections were reported. This was followed by Maryland with 7.4% ($[2.7, 83.8]$), Connecticut with 8.0% ($[3.1, 25.8]$), and Florida with 8.7% ($[4.8, 34.0]$). This underreporting issue extends to most states as in 39 states less than 30% of infections were reported during this time. Only 4 states of Alaska, Maine, Vermont and Virginia reported at least 40% infections. No states were found to surpass 50% for reported infections for this time.

Similar patterns were observed during the earlier period of Alpha domination, where Louisiana had the lowest reported infections at 11.7% (95% confidence interval: $[6.7, 31.5]$) and was followed by California at 14.4% (95% confidence interval: $[7.7, 68.2]$). There were 23 states that reported at least 40% and 22 states that reported at least 50% of their infections.

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% (95% confidence interval: $[16.2, 34.0]$) and 22.3% (95% confidence interval: $[14.8, 40.5]$), respectively. During this time, 28 states that reported at least 40% and 14 states that reported at least 50% of their infections.

Figure 8 zooms in on the infection estimates for the states that exhibit the most underreporting for each of the three variant-specific time periods to emphasize the marked differences between case and infection rates. To supplement this, Figure 9 shows the division of infections in states by the proportions of variants in circulation specific to the state. From these plots, it is clear that few variant categories tends to dominate and drive infections at a time. The general progression in terms of variant starts with the Ancestral category from 2020 up to early 2021, to the Alpha variant in mid 2021, which eventually gets eclipsed by the Delta variant in mid to late 2021. This supports our division of our results by the three main variant-driven time periods.

Figure 6: 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.

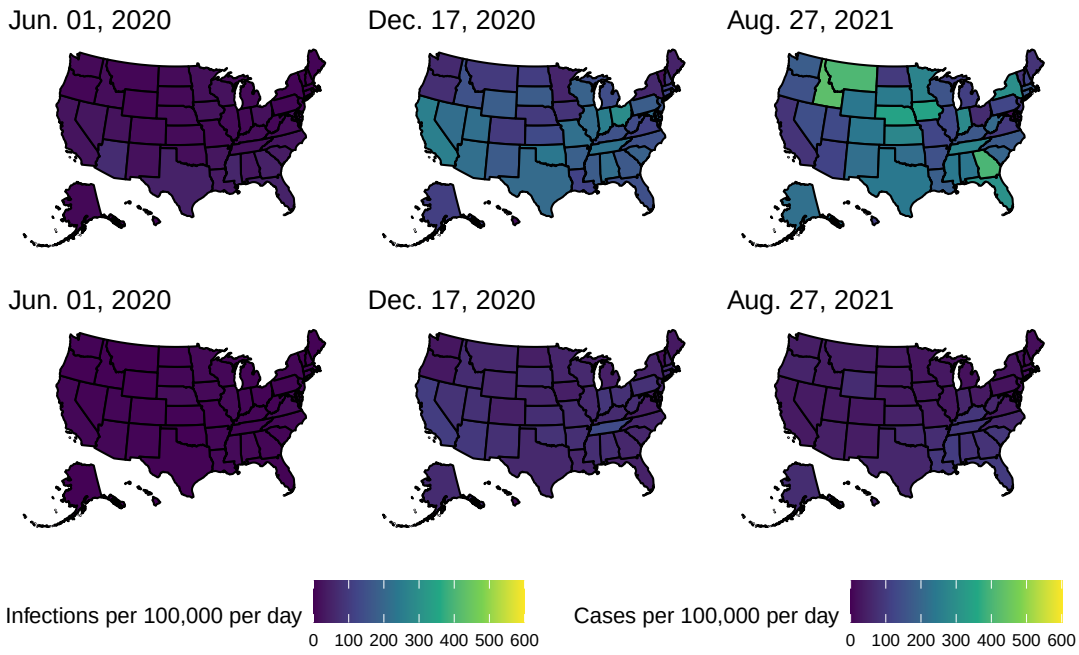

Figure 7: Choropleth maps of the state-level estimates of the number of daily new infections per 100,000 population (top row) and the daily new cases per 100,000 population (bottom row) for three times over the June 1, 2020 to November 29, 2021 period. The first date was chosen simply as a baseline, while the second and third dates were chosen based on the day that had the largest number of infections across the 50 states from each year. These maps are generated from the `usmap` package in R (Di Lorenzo, 2023).

We perform a lagged correlation analysis where we systematically investigate the rank-based (i.e., Spearman's) correlation between our infection and confirmed hospitalization rates per 100,000 population over a broad range of lag values (Figure 10). Examining the correlation between infections and hospitalizations shows that the maximum average correlation across states of 0.513 is first observed at a lag of 13 days. In contrast, we find that the greatest average rank-based correlation for cases of 0.691 is achieved 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.

We undertake an ablation study for the lagged correlation of infections, the results of which are shown in Figure 11. 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.

In terms of the average correlation produced, the deconvolved case estimates by infection onset and the deconvolved case estimates by positive specimen date reach almost the same maximum average correlation. While that is not a clear differentiator by itself, there is a clear time-based benefit of opting for the infection estimates by the date of infection onset over symptom onset because they provide similar information on hospitalizations about 6 days before the latter tends to occur.

Unsurprisingly, the deconvolved case and infection estimates achieve their maximum correlation at the same lag. And yet, the average correlation to hospitalizations tends to be greater for the deconvolved case estimates than for the infection estimates (and the reported infections by symptom onset). If the goal is

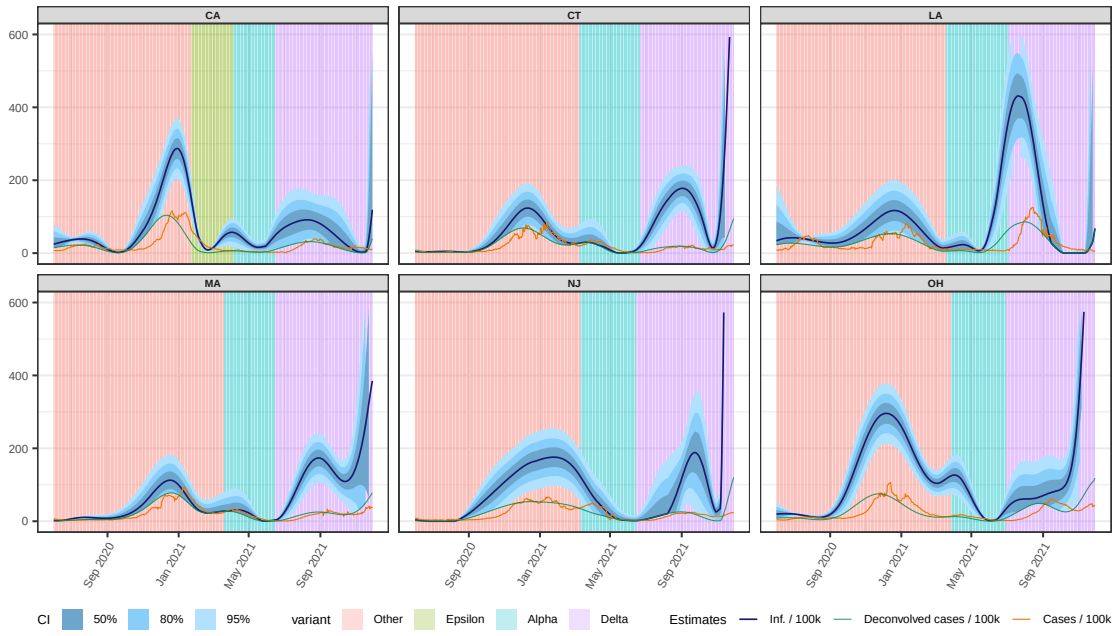

Figure 8: Estimates of the number of daily new infections per 100,000 for six U.S. states from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% confidence intervals. The background is shaded to indicate the top variant in circulation at the time.

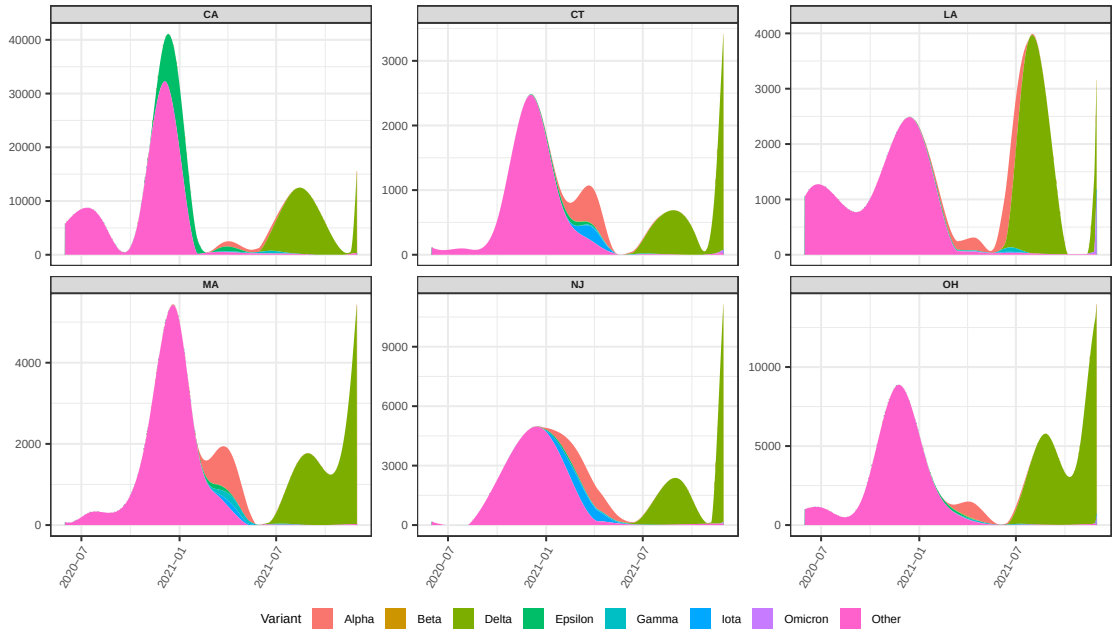

Figure 9: Estimates of the deconvolved case counts by variant for six U.S. states from June 1, 2020 to November 29, 2021. The area under the curve is coloured to indicate the corresponding variant.

to find the part of this process that is most informative to hospitalizations, then it is clear that producing the deconvolved case estimates is the more informative step and these estimates are potentially the most meaningful signal for future hospitalizations. The reason for this finding may stem from a difference in disease

severity between the reported and unreported infections: The unreported infections tend to be less severe and less likely to lead to hospitalization than those that are reported.

As a counterpart to our lagged correlation analysis, we compute the time-varying IHRs for each state using the optimal lag for infection and hospitalization rates. We also included the CHRs that are computed using the optimal lag for cases and hospitalizations for comparison (Figure 12).

For each state, the CHRs tend to show an amplified version of the course of the IHRs over time. 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 (Hodcroft, 2021). This finding aligns with previous studies that found an increased risk in hospitalizations with Delta in comparison to other variants (Twohig et al., 2022; Nyberg et al., 2022). 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 or Wyoming).

There does not tend to be a strict upwards or downwards trajectory or even a mild waning pattern in the IHRSs (as one might expect if we were to tread further into the pandemic over which the virus mutates to variants that are generally more infectious, but that pose less of a risk to hospitalization (Lorenzo-Redondo et al., 2022; Blauer, 2022a)). Overall, we observe intermittent spikes that punctuate longer periods where the IHRs tend to be less than 0.2 hospitalizations per infection. These spikes tend to align with the emergence of new variants.

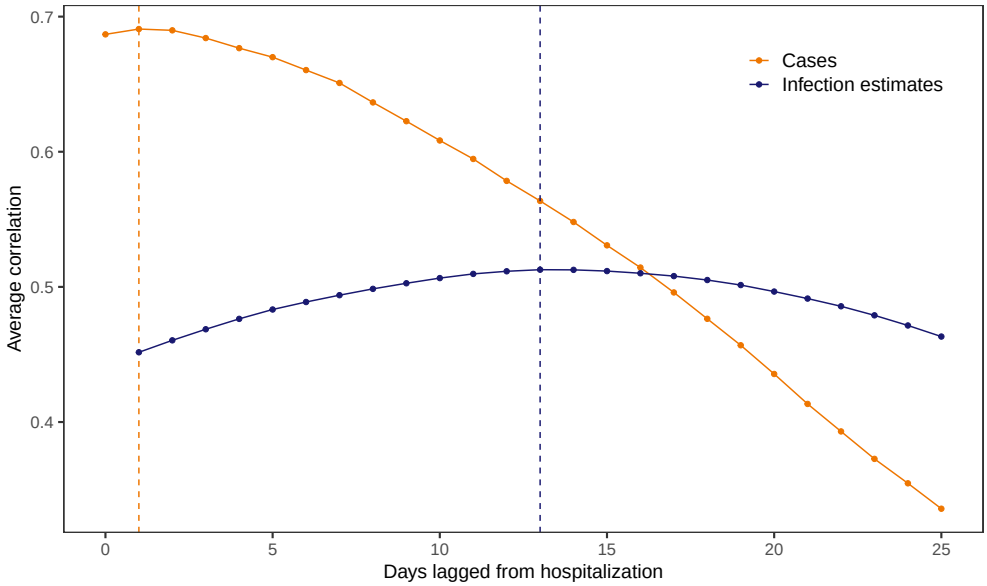

Figure 10: Lagged Spearman's correlation between infection and hospitalization rates per 100,000 as well as between case and hospitalization rates per 100,000. The averages shown are for each lag, across U.S. states and days over June 1, 2020 to November 29, 2021, and taken over a rolling window of 61 days. Note that the infections, cases, and hospitalization counts are 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.

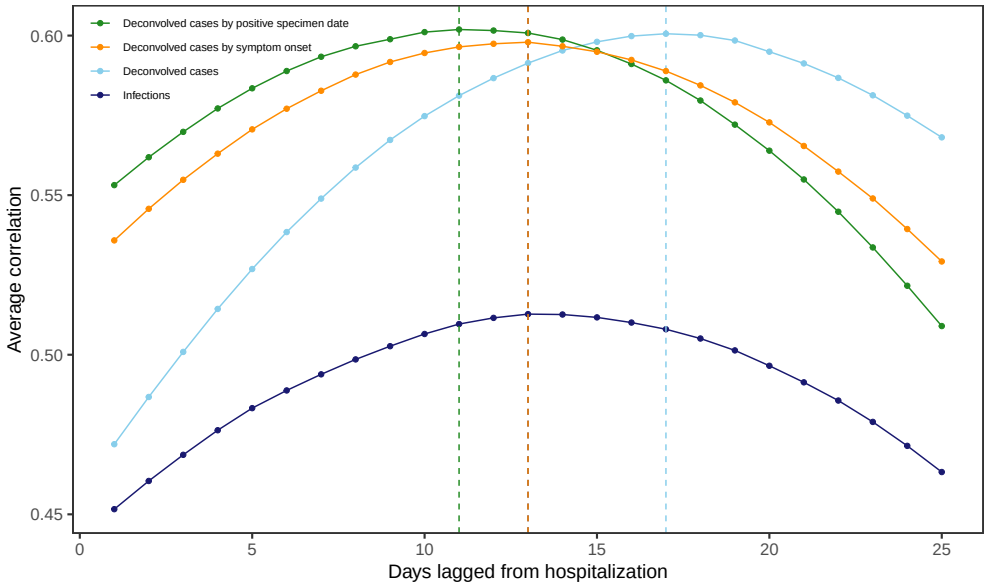

Figure 11: Lagged Spearman's correlation between the infection and hospitalization rates per 100,000 averaged for each lag across U.S. states and days over June 1, 2020 to November 29, 2021, and taken over a rolling window of 61 days. The infection rates are based on the counts for the deconvolved case 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.

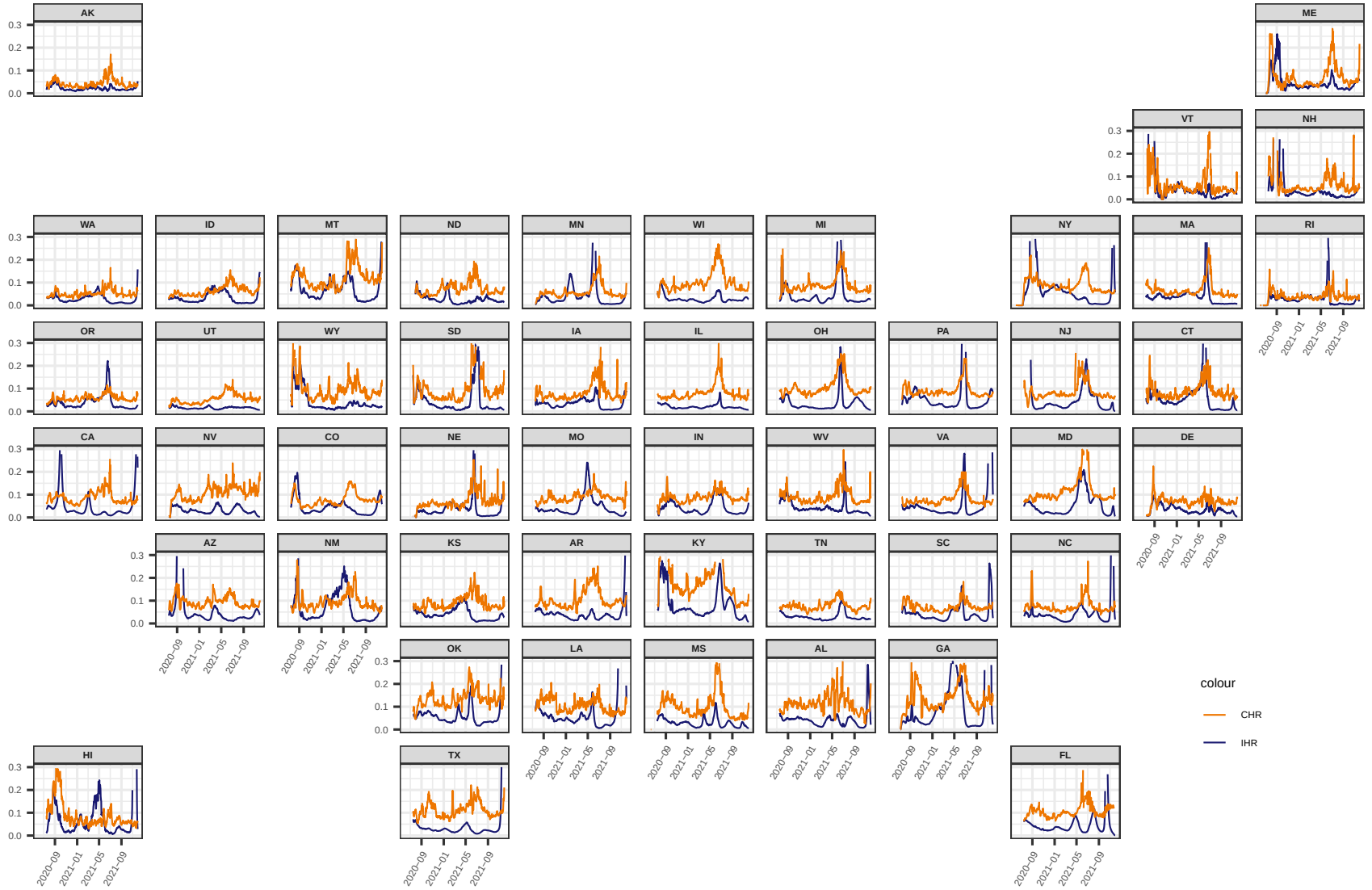

Figure 12: Time-varying IHR and CHR estimates for each state from June 1, 2020 to November 29, 2021, obtained using the corresponding optimal lag from the systematic lag analysis. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious day of the week effects. Also note that the different starting points across states are due to the availability of the hospitalization data.

3.2 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 (Figure 6, Figure 8). 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]). These are closely followed by Montana with 432 on September 8, 2021 (95% confidence interval: [282, 581]), Louisiana with 431 on July 20, 2021 (95% confidence interval: [252, 610]), and Wyoming with 350 on November 13, 2020 (95% confidence interval: [256, 444]).

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]), South Dakota with 342 infections per 100,000 on July 3, 2021 (95% confidence interval:[177, 539]), and Illinois with 340 infections per 100,000 on July 3, 2021 (95% confidence interval: [177, 539]). During this time, 74% of the top rates for each state were observed in the late fall or winter of 2020.

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.

4 Discussion

We obtained retrospective estimates of daily incident infections for each U.S. state for June 1, 2020 to November 29, 2021. Our infection estimates suggest that the pandemic has an impact in states earlier and at a larger scale than is indicated by cases. Since case reporting is not consistent across time and states, case counts underestimate the true number of infections and, hence, the impact of the pandemic (Centers for Disease Control and Prevention, 2022; Simon, 2021). For example, some states report the number of individuals tested rather than the numbers of tests performed (Schechtman, 2020; Chitwood et al., 2022). Additionally, while the definition of a confirmed COVID-19 case tends to be fairly uniform across the United States due to general adherence to the CSTE case definitions, state reporting standards have been known to vary (Centers for Disease Control and Prevention, 2020a; CMU Delphi Research Group, 2020). For instance, there may be inconsistencies across locations if some cases are labelled as confirmed based on positive antigen tests instead of PCR tests (The COVID Tracking Project, 2021).

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 report dates generally follow symptom and infection onset, cases are a fundamentally flawed indicator of disease burden because they 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 (Washington State Department of Health, 2020; Dunkel, 2020). Furthermore, no indication of uncertainty is provided for even the gold standard case estimates (CMU Delphi Research Group, 2020). Thus, while reported cases provide an indication of the trajectory of the pandemic, it is a delayed and incomplete version. Estimating the new number of infections by symptom or infection onset date would more closely align with the definition of incidence as we know it (Jahja et al., 2022).

From the correlation analysis between daily infection estimates and hospitalizations, a lag of 13 days gives the maximum average correlation across states. This is in agreement with the 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 (more precisely, 8.0 days (95% interval: [2.7, 18.5]) for the first wave to 9.7 days (95% interval: [4.1, 19.6]) for the second wave, (Ward and Johnsen, 2021)). However, we should note the first study is based on a small sample size for outbreak cases reported well before our study start date. As well, both sets of estimates depend upon the healthcare system and the population structure, amongst other things (Ward and Johnsen, 2021). Nevertheless, their relative agreement with our estimate of 13 days for the U.S. states lends some credence to of our results.

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 (Russell et al., 2023). 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 (similar to Fox et al., 2023 though with the additional element of being time-varying). We strongly believe this would be a worthwhile direction to pursue in future work should the necessary information be available.

The remainder of our discussion consists of an in-depth look into the advantages and limitations of our approach and of other comparable approaches, followed by a high level summary of our work and its major contributions.

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 case, line list, and variant circulation data, we are able to construct incubation and delay distributions that are unique for each state. By using time-varying and state-specific seroprevalence data, we are able to allow the reporting ratio to vary over both time and state, which is an advantage over such ratios that are non-time varying but state-specific and those that are time-varying but the same for all states (Unwin et al., 2020; Center for the Ecology of Infection Diseases, 2020). Existing approaches that use the delay distribution to generate infection estimates often only construct one delay distribution that is used for all states (Chitwood et al., 2022; Jahja et al., 2022). That is, they operate under the assumption of geographic invariance, where it is assumed that all states have the same patterns of delay from onset to case report, which is unlikely to be true due to differences in reporting pipelines, pandemic response, and variants in circulation, amongst other things.

Another major limitation of these existing approaches to derive infections is that they do not to account for reinfections. Now, it may be contended that reinfections do not account for a substantial fraction of the infections until later in the pandemic, so they are not absolutely necessary to include in the earlier stages of the pandemic. Still, at no stage did infection confer lifelong immunity. Rather antibody levels and immunity are known to wane over time. And we believe it is important to account for such defining characteristics of the virus when tracking infections over time. Therefore, we account for reinfections and the waning of detectable antibody levels in our custom antibody prevalence model. 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 (Pooley et al., 2023), 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 of how newer variants may escape detection (National Institutes of Health, 2022; U.S. Food and Drug Administration, 2023). 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 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 (Wei et al., 2024; Pulliam et al., 2022; Eythorsson et al., 2022). So having quality reinfection data that is representative of each location under study is of the utmost importance for the Omicron era.

While it would be ideal to use confirmed rates over time for each U.S. state, most states do not publicly report reinfection data over the entire time period we considered. So we have turned to suspected reinfection data over time for Clark County, USA, as that surveillance is among the most detailed and reputable that we have found for the United States. Nevertheless, using such localized data raises questions of representativeness and the applicability of such estimates to Nevada and all other states. Furthermore, this data has no information available beyond suspected third infections, which imposes an irremediable bias. However, based on the third infection data available there, we expect that the probability of being reinfected more than three times is likely very low for time frame considered and so the omission of these would impact our infection estimates to a minimal extent.

The vast majority of issues we encountered when trying to reconstruct the infection time series for each state are due to an absence or a lack of data. Such is the primary issue we had with the restricted line list. In comparison to the number of JHU cases (which we are treating as a gold standard) for the same release date, we noted there are about 10 million cases that are unaccounted for in the CDC line list. Moreover, the missingness does not appear to be random and uniformly distributed across states. Rather it is unequally distributed, suggesting that the dataset is likely biased. However, more information on the cases that are missing versus present would be required to determine the extent the missing cases led to a nonrepresentative, and therefore, biased sample, and could be a topic of further study.

Seroprevalence data also runs the risk of being nonrepresentative of the intended population (Bajema et al., 2021). For example, in the blood donor dataset some states have region specific-estimates, which clearly do not stand for the entire state. Another source of systematic variation is in the characteristics of the individuals who opt for blood tests versus those who do not. For instance, there may be a healthy user bias, in which a number of those who opt for blood tests are generally more inclined to partake in proactive healthy behaviors (such as checking on basic health markers by taking an annual blood test) than those who do not (Parsley Health, 2018). Alternatively, a number of individuals may be recommended for blood tests by their doctors due to signs of ill-health (ex. mineral deficiencies or underlying medical conditions). The extent that each such bias persists depends on the purpose of the blood test and whether it was used as a proactive or reactive medical tool. Since such information is unavailable to us, all we can conclude is that participant-driven sources of bias impact the seroprevalence samples to an undetermined extent. There are additional concerns about the performance of antibody testing for individuals with mild or asymptomatic disease as well as about the loss of immunity over time (Kaku et al., 2021; Seow et al., 2020; Ibarrondo et al., 2020).

In this work, we do not attempt to directly address infection underascertainment due to the increase in asymptomatic infections across variants (Public Health Ontario, 2023). We simply note that this would likely pose a greater problem later in the pandemic, particularly after the Delta era (Fan et al., 2022). We hope that such infections would be largely represented by the seroprevalence and reinfection estimates, but there is undoubtedly increasing reliance on such estimates to be able to do this over time (owing to the simultaneous decline in the reporting cadence and the apparent rise in asymptomatic infections over time) (Ontario Agency for Health Protection and Promotion, 2022; Garrett et al., 2022; Blauer, 2022b; Ren et al., 2021). Consequently, there is an increasing uncertainty over time that is not captured by the model or the estimates.

Due to such concerns with the seroprevalence data, one further area of research is on investigating the utility of various sources to estimate the incidence of infections. Intuitively, one might expect that leveraging data from multiple sources would likely lead to more accurate and stable estimates than those from using one source. Wastewater surveillance data is one promising source that may be complementary to seroprevalence data, especially when testing is low (McManus et al., 2023). However, there has been limited success in

predicting incidence using such data. The extent that wastewater concentration data is a useful in estimating COVID-19 incidence is unclear owing to problems with viral occurrence and detectability in wastewater that render detection inconsistent across locations (ex. due to temperature, per-capita water use, and in-sewer travel time) (McManus et al., 2023; Hart and Halden, 2020; Li et al., 2023). 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 (European Centre for Disease Prevention and Control, 2020).

Overall, 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.

Our approach is predicated upon having case, line list, viral circulation, and seroprevalence data for each state, all of which are readily available (or available upon request in the case of restricted line list data). As a result of this, we are able to demonstrate the feasibility of estimating COVID-19 infections at the state level by using standard sources of data.

Our framework is quite versatile as it lends itself to more localized, county or community level estimates, or globalized, country-specific estimates. Fundamentally, to produce estimates of infections for different geographic regions, one would simply need to input the required data and re-run the code pipeline. In this way, one could readily adapt our approach to generate estimates for the provinces in Canada or the regions in England.

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.

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 (Elbe and Buckland-Merrett, 2017), on which this research is based.

References

Bajema, K. L., Wiegand, R. E., Cuffe, K., Patel, S. V., Iachan, R., Lim, T., Lee, A., Moyse, D., Havers, F. P., Harding, L. et al. (2021) Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. *JAMA Internal Medicine*, **181**, 450–460.

Blauer, B. (2022a) Comparing cases, deaths, and hospitalizations indicates Omicron is less deadly. https://coronavirus.jhu.edu/pandemic-data-initiative/data-outlook/comparing-cases-deaths-and-hospitalizations-indicates-omicron-less-deadly.

— (2022b) Reduce data reporting cadence for an endemic disease? Not quite yet. https://coronavirus.jhu.edu/pandemic-data-initiative/data-outlook/reduce-data-reporting-cadence-for-an-endemic-disease-not-quite-yet.

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

Centers for Disease Control and Prevention (2020a) August 5, 2020 CSTE case definition. https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/08/05/.

— (2020b) COVID-19 case surveillance public use data. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf.

— (2020c) COVID-19 case surveillance restricted access detailed data. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t.

— (2020d) COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/#national-lab.

— (2021a) 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.

— (2021b) Nationwide commercial laboratory seroprevalence survey. https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv.

— (2022) Estimated COVID-19 burden. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html.

Chitwood, M. H., Russi, M., Gunasekera, K., Havumaki, J., Klaassen, F., Pitzer, V. E., Salomon, J. A., Swartwood, N. A., Warren, J. L., Weinberger, D. M. et al. (2022) 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.

CMU Delphi Research Group (2020) Delphi epidata API. https://cmu-delphi.github.io/delphi-epidata.

Cortés Martínez, J., Pak, D., Abelenda-Alonso, G., Langohr, K., Ning, J., Rombauts, A., Colom, M., Shen, Y. and Gómez Melis, G. (2022) 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.

Di Lorenzo, P. (2023) *usmap*. URL: https://usmap.dev. R package version 0.6.2.

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

Duerr, R., Dimartino, D., Marier, C., Zappile, P., Wang, G., Lighter, J., Elbel, B., Troxel, A. B., Heguy, A. et al. (2021) Dominance of Alpha and Iota variants in SARS-CoV-2 vaccine breakthrough infections in New York City. *The Journal of Clinical Investigation*, **131**, e152702.

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

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

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

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

Eythorsson, E., Runolfsdottir, H. L., Ingvarsson, R. F., Sigurdsson, M. I. and Palsson, R. (2022) Rate of sars-cov-2 reinfection during an omicron wave in iceland. *JAMA Network Open*, **5**, e2225320–e2225320.

Fan, Y., Li, X., Zhang, L., Wan, S., Zhang, L. and Zhou, F. (2022) SARS-CoV-2 Omicron variant: Recent progress and future perspectives. *Signal Transduction and Targeted Therapy*, **7**, 141.

Fox, S. J., Javan, E., Pasco, R., Gibson, G. C., Betke, B., Herrera-Diestra, J. L., Woody, S., Pierce, K., Johnson, K. E., Johnson-León, M. et al. (2023) Disproportionate impacts of COVID-19 in a large US city. *PLOS Computational Biology*, **19**, e1011149.

Garrett, N., Tapley, A., Andriesen, J., Seocharan, I., Fisher, L. H., Bunts, L., Espy, N., Wallis, C. L., Randhawa, A. K., Ketter, N. et al. (2022) High rate of asymptomatic carriage associated with variant strain Omicron. *MedRxiv*.

Grant, R., Charmet, T., Schaeffer, L., Galmiche, S., Madec, Y., Von Platen, C., Chény, O., Omar, F., David, C., Rogoff, A. et al. (2022) 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.

Hart, O. E. and Halden, R. U. (2020) 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.

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

Hitchings, M. D., Dean, N. E., García-Carreras, B., Hladish, T. J., Huang, A. T., Yang, B. and Cummings, D. A. (2021) 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.

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

Ibarrondo, F. J., Fulcher, J. A., Goodman-Meza, D., Elliott, J., Hofmann, C., Hausner, M. A., Ferbas, K. G., Tobin, N. H., Aldrovandi, G. M. and Yang, O. O. (2020) Rapid decay of anti–SARS-CoV-2 antibodies in persons with mild COVID-19. *New England Journal of Medicine*, **383**, 1085–1087.

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

Jones, J. M., Stone, M., Sulaeman, H., Fink, R. V., Dave, H., Levy, M. E., Di Germanio, C., Green, V., Notari, E., Saa, P. et al. (2021) Estimated US infection-and vaccine-induced SARS-CoV-2 seroprevalence based on blood donations, July 2020-May 2021. *JAMA*, **326**, 1400–1409.

Kaku, N., Nishimura, F., Shigeishi, Y., Tachiki, R., Sakai, H., Sasaki, D., Ota, K., Sakamoto, K., Kosai, K., Hasegawa, H. et al. (2021) Performance of anti-SARS-CoV-2 antibody testing in asymptomatic or mild COVID-19 patients: A retrospective study in outbreak on a cruise ship. *PLoS One*, **16**, e0257452.

Li, X., Zhang, S., Sherchan, S., Orive, G., Lertxundi, U., Haramoto, E., Honda, R., Kumar, M., Arora, S., Kitajima, M. et al. (2023) 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.

Lorenzo-Redondo, R., Ozer, E. A. and Hultquist, J. F. (2022) COVID-19: Is Omicron less lethal than Delta? *British Medical Journal*, **378**.

McManus, O., Christiansen, L. E., Nauta, M., Krogsgaard, L. W., Bahrenscheer, N. S., von Kappelgaard, L., Christiansen, T., Hansen, M., Hansen, N. C., Kähler, J. et al. (2023) Predicting COVID-19 incidence using wastewater surveillance data, Denmark, October 2021–June 2022. *Emerging Infectious Diseases*, **29**, 1589.

National Institutes of Health (2022) 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.

Nyberg, T., Ferguson, N. M., Nash, S. G., Webster, H. H., Flaxman, S., Andrews, N., Hinsley, W., Bernal, J. L., Kall, M., Bhatt, S. et al. (2022) 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.

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

Ontario Agency for Health Protection and Promotion (2022) 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.

Parsley Health (2018) 5 essential blood tests you need every year. https://www.parsleyhealth.com/blog/5-essential-blood-tests-need-every-year/.

Pellis, L., Scarabel, F., Stage, H. B., Overton, C. E., Chappell, L. H., Fearon, E., Bennett, E., Lythgoe, K. A., House, T. A., Hall, I. et al. (2021) 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.

Pitzer, V. E., Chitwood, M., Havumaki, J., Menzies, N. A., Perniciaro, S., Warren, J. L., Weinberger, D. M. and Cohen, T. (2021) 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.

Pooley, N., Abdool Karim, S. S., Combadière, B., Ooi, E. E., Harris, R. C., El Guerche Seblain, C., Kisomi, M. and Shaikh, N. (2023) Durability of vaccine-induced and natural immunity against COVID-19: A narrative review. *Infectious Diseases and Therapy*, **12**, 367–387.

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

Public Health Ontario (2023) COVID-19 Omicron variant of concern and communicability – What we know so far. https://www.publichealthontario.ca/-/media/documents/ncov/covid-wwksf/2022/01/wwksf-omicron-communicability.pdf.

Pulliam, J. R., van Schalkwyk, C., Govender, N., von Gottberg, A., Cohen, C., Groome, M. J., Dushoff, J., Mlisana, K. and Moultrie, H. (2022) Increased risk of sars-cov-2 reinfection associated with emergence of omicron in south africa. *Science*, **376**, eabn4947.

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

Reinhart, A., Brooks, L., Jahja, M., Rumack, A., Tang, J., Agrawal, S., Al Saeed, W., Arnold, T., Basu, A., Bien, J. et al. (2021) An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, **118**, e2111452118.

Ren, R., Zhang, Y., Li, Q., McGoogan, J. M., Feng, Z., Gao, G. F. and Wu, Z. (2021) Asymptomatic SARS-CoV-2 infections among persons entering China from april 16 to october 12, 2020. *Jama*, **325**, 489–492.

Ruff, J., Zhang, Y., Kappel, M., Rathi, S., Watkins, K., Zhang, L. and Lockett, C. (2022) Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. *Emerging Infectious Diseases*, **28**, 1977.

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

Schechtman, K. (2020) Counting COVID-19 tests: How states do it, how we do it, and what's changing. https://covidtracking.com/analysis-updates/counting-covid-19-tests.

Seow, J., Graham, C., Merrick, B., Acors, S., Pickering, S., Steel, K. J., Hemmings, O., O'Byrne, A., Kouphou, N., Galao, R. P. et al. (2020) Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nature Microbiology*, **5**, 1598–1607.

Simon, S. (2021) Inconsistent reporting practices hampered our ability to analyze COVID-19 data. Here are three common problems we identified. https://covidtracking.com/analysis-updates/three-covid-19-data-problems.

Tanaka, H., Ogata, T., Shibata, T., Nagai, H., Takahashi, Y., Kinoshita, M., Matsubayashi, K., Hattori, S. and Taniguchi, C. (2022) Shorter incubation period among COVID-19 cases with the BA. 1 Omicron variant. *International Journal of Environmental Research and Public Health*, **19**, 6330.

The COVID Tracking Project (2021) The COVID tracking project: The data. https://covidtracking.com/data.

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

The Washington Post (2020) 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.

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

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

Tindale, L. C., Stockdale, J. E., Coombe, M., Garlock, E. S., Lau, W. Y. V., Saraswat, M., Zhang, L., Chen, D., Wallinga, J. and Colijn, C. (2020) Evidence for transmission of COVID-19 prior to symptom onset. *eLife*, **9**, e57149.

Twohig, K. A., Nyberg, T., Zaidi, A., Thelwall, S., Sinnathamby, M. A., Aliabadi, S., Seaman, S. R., Harris, R. J., Hope, R., Lopez-Bernal, J. et al. (2022) 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.

Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J., Vollmer, M. A., Whittaker, C., Filippi, S. L. et al. (2020) State-level tracking of COVID-19 in the United States. *Nature Communications*, **11**, 6189.

U.S. Census Bureau, Population Division (2022) 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.

U.S. Food and Drug Administration (2023) 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.

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

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

Wei, J., Stoesser, N., Matthews, P. C., Khera, T., Gethings, O., Diamond, I., Studley, R., Taylor, N., Peto, T. E., Walker, A. S. et al. (2024) Risk of sars-cov-2 reinfection during multiple omicron variant waves in the uk general population. *Nature Communications*, **15**, 1008.

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

Yang, S., Hemarajata, P., Hilt, E. E., Price, T. K., Garner, O. B. and Green, N. M. (2022) 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.

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

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 1 is conceptualized as a Gaussian state space model (as in Durbin and Koopman, 2012; Helske, 2017).

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 (Helske, 2017; Durbin and Koopman, 2012). 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 (Durbin and Koopman, 2012).

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 (Durbin and Koopman, 2012). 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

$$\hat{a}_t = \mathbb{E}[\alpha_t \mid y_n, \ldots, y_1] \tag{4}$$
$$V_t = \text{Var}(\alpha_t \mid y_n, \ldots, y_1). \tag{5}$$

The filtering and smoothing steps are based on recursions that are described in Appendix A of Helske (2017) 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. (2020) 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 (Durbin and Koopman, 2012), 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

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