Supplementary Information: Retrospective estimation of latent COVID-19 infections before Omicron in the US

Rachel Lobay

2024-06-01

Contents

List of Figures

List of Tables

S1 Supplementary Methods

S1.1 A general description and depiction of convolution

In general, the goal of convolution is to propagate the input signal forward in time using a probability distribution. In the 1D and discrete context, it is simply a rolling, weighted average of the past. So for an input sequence $\{x_t\}_{t=1}^n$ and time-constant weights $\{z(k)\}_{k=-\infty}^0$, the output sequence $\{y_t\}_{t=1}^n$ is given by

$$y_t = \sum_{s=0}^{t} z(s) x_{t-s}. \tag{1}$$

Figure S1 presents a depiction of the convolution procedure for an example signal x_t (smoothed cases, orange line). Essentially, to push the cases forward in time, we take the appropriately aligned (forward-in-time) delay distribution $z(s)$ (blue shaded region) and convolve it with the smoothed cases signal counts by it to get the convolved estimates (blue line). This process is repeated as we march forward in time, as shown through the stop-motion panels, such that it eventually covers the entire line of cases. An important takeaway from this is that convolution is not the same as a simple shift of the data. A shift is a special case: when $z(d) = 1$ and $z(s) = 0$, $s \neq d$, we shift x forward by d. Rather, convolution generally weights the entire past by non-zero probabilities. Deconvolution proceeds in the same fashion, but in the opposite direction, going backward in time and undoing the effect of a convolution.

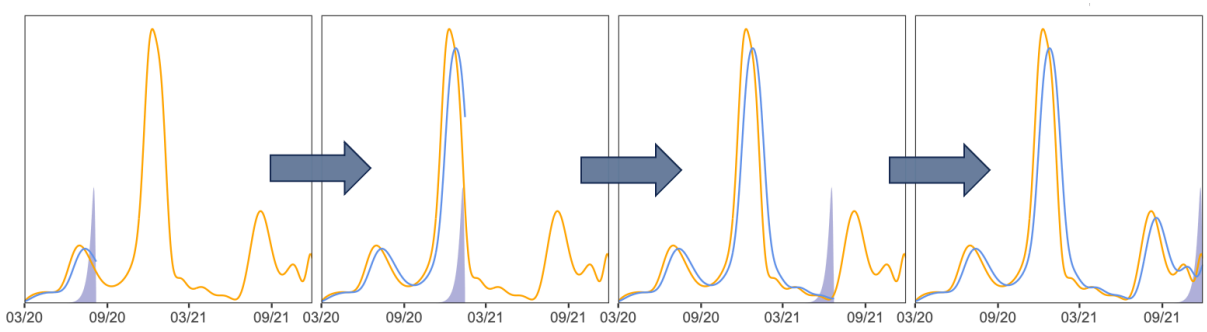

Figure S1: A general depiction of convolving smoothed cases (orange line) with the corresponding delay probabilities (shaded blue area) to get the convolved estimates (blue line) over four different times.

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

Because the CDC line list is updated monthly and cases may undergo revision, we use a single version of it that was released on June 6, 2022. We consider this version to be finalized in that it is well-beyond our study end date such that the dataset is unlikely to be subject to further significant revisions.

Table S1 presents the percent of pairwise occurrences for the different possible permutations of events in the line list. Essentially, most cases follow the idealized ordering shown by Figure 1 and so we adhere to this construction as much as possible. Unfortunately, the line list has significant missing data, notably with respect to our variables of interest. Approximately 62.3% of cases are missing the symptom onset date, 55.4% are missing positive specimen date, and 8.96% of cases are missing the report date. Furthermore, cases with missing report or positive specimen dates may be filled with their symptom onset date[1]. So it is possible that all three variables may have the same date for a case. However, we only actually deal with pairs of these events; we do not use all three at once in our construction of the delay distributions. Therefore, we restrict our investigation of coincident missingness to the possible pairs.

Due to the contamination in the zero delay cases (those whose symptom onset was used to fill missing positive specimen or report date, the true extent of which is unknown to us), we omit all cases where the positive specimen and report dates have zero delay from our analysis. We choose to allow for zero and negative delay for symptom onset to report because correspondence with the CDC confirms the possibility that a person could test positive before symptom onset and it is a reasonable ordering to expect if, for example, the individual is aware that they have been exposed to an infected individual.

Order of events	**Percent pairwise occurrence**	**Handling**
IO $\rightarrow$ SO $\rightarrow$ PS $\rightarrow$ RE	PS $\geq$ 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 $\rightarrow$ PS and PS $\rightarrow$ 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 $\rightarrow$ PS $\rightarrow$ SO $\rightarrow$ RE	PS <SO: 2.91 SO $\leq$ 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 $\rightarrow$ PS $\rightarrow$ RE $\rightarrow$ 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).

The CDC restricted line list contains 74,849,225 cases (rows) in total compared to 84,714,805 cases reported by the JHU CSSE; that is, line list is missing about 10 million cases. The extent that this issue impacts each state is shown in Figure S3, from which it is clear the fraction of missing cases is substantial for many states, often surpassing 50%[1]. 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.

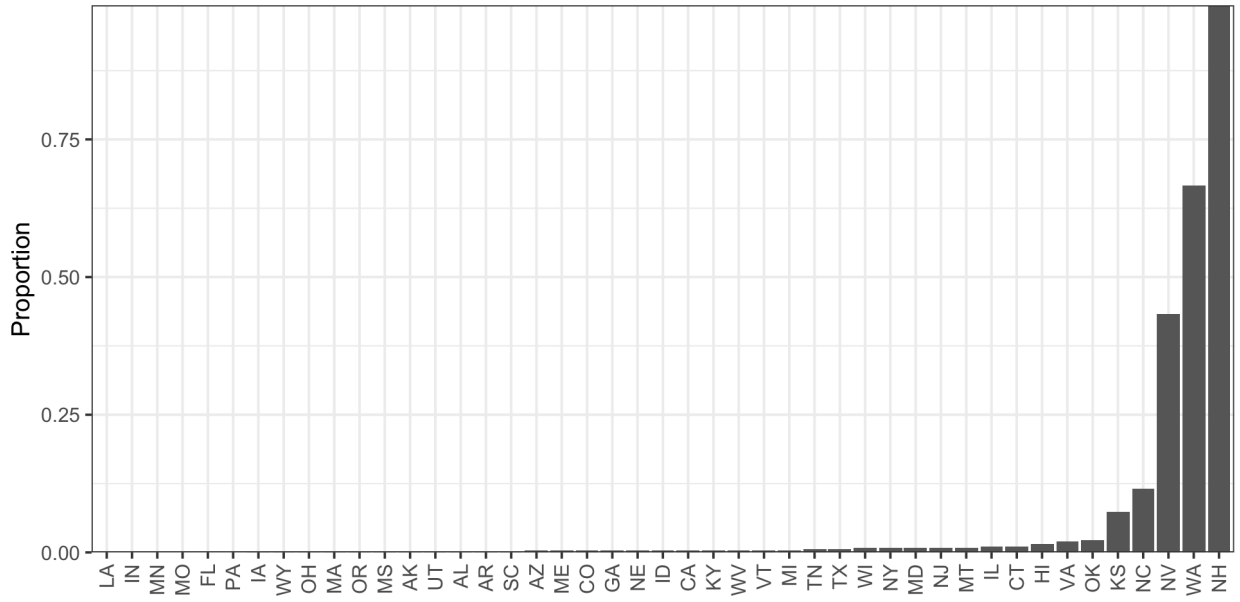

Figure S2: Proportion of complete cases with zero delay between positive specimen and report date in the restricted CDC line list dataset.

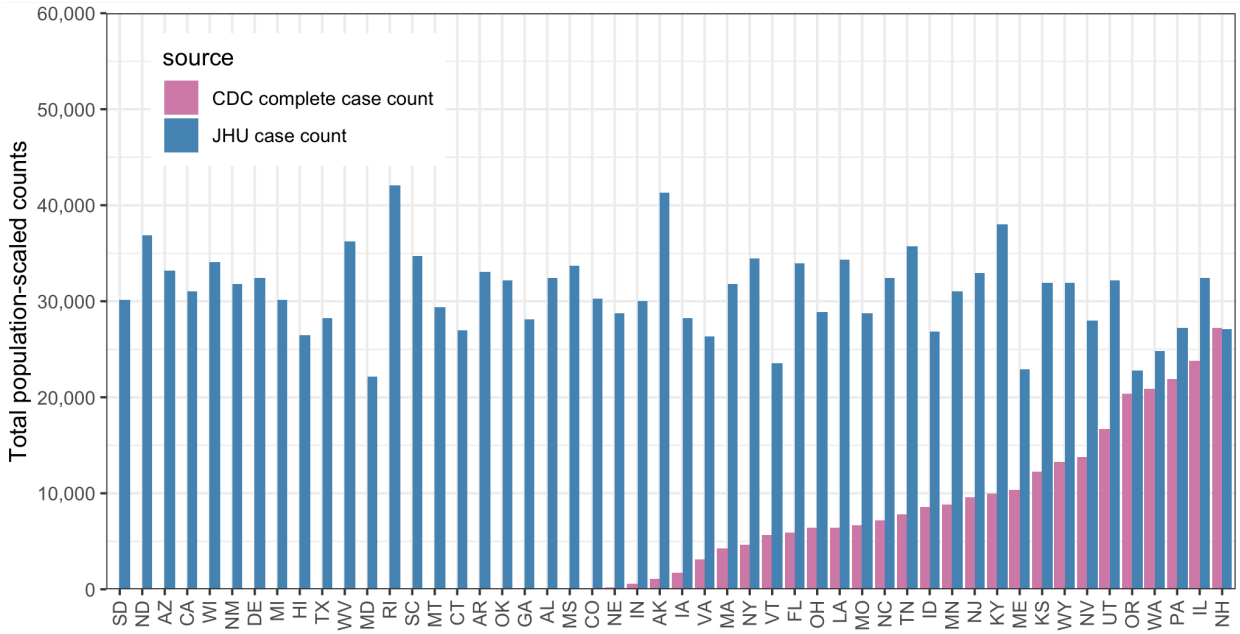

Figure S3: Complete case counts by state in the CDC line list versus the cumulative complete case counts from JHU CSSE as of June 6, 2022. All counts have been scaled by the 2022 state populations as of July 1, 2022 from[2].

S1.3 Additional details on delay distribution calculations

In the line list, we observe unusual spikes in reporting in 2020 in comparison to majority of 2021. When stratified by report date, a few states are seen to contribute unusually large case counts on isolated dates very late in the reporting process (well beyond 100 days following specimen collection). These large accumulations of cases over time are likely due breakdowns of the reporting pipeline. Such anomalies are not likely to be reliable indicators of the delay from positive specimen to case report. Therefore, we prune these reporting backlogs systematically. The heuristic is to find any large batches of cases that were all simultaneously reported on the same date with a lengthy delay (as would happen if a state "found" a tranche of previously unreported positive tests).

For this, we operate only on part of the line list intended for the positive specimen to case report delay estimation (as this is where we use case reports from the line list).

Then, we divide this line list into three roughly evenly-spaced and non-overlapping intervals over 2020 and early 2021 (the interval end points on July 16, 2020, October 16, 2020, and January 15, 2021). For each interval, we separate the cases with reporting delays that are at least 50 days long into equally-sized bins of 50 days.

Then, for each interval and bin, we perform the following pruning procedure:

1. Obtain the total count of these delays for each state.
2. Check whether each such count on the log scale is at least the median (for the bin) plus 1.5 times the interquartile range. Retain only those that exceed this criterion as candidates for pruning.
3. For the candidate states, compute the case counts by report date.
4. If there is a report date for a candidate state with a case count greater than or equal to a pre-specified threshold, then remove/prune those cases from the line list. We set the threshold to be 2000 cases for the first two bins, and then lower it to 500 for the remaining bins based on inspection.

Finally, in New Hampshire, for a small handful of report dates, all cases reported by JHU appear in the CDC line list, and all are recorded as having positive specimen collection date equal to the report date. The resulting estimate of the delay distribution (see Section 4.1) $\widetilde{p}_{\mathrm{NH},t}(k)$ would be a point mass at $k = 0$ and the weight $\alpha_{\mathrm{NH},t} = 1$ resulting $\widehat{\pi}_{\mathrm{NH},t}(k)$ also being a point mass at $k = 0$. In this specific case, we force $\alpha_{\mathrm{NH},t} = \min_{\ell \neq \mathrm{NH}} \alpha_{\ell,t}$.

S1.4 Variant circulation proportions

To estimate the daily proportions of the variants circulating in each state, we obtain the GISAID genomic sequencing data from CoVariants.org[3,4]. These counts represent the total number of cases belonging to a particular variant using a sample of positive tests over a biweekly period. To estimate the population proportion of each variant, we apply multinomial

logistic regression for the eight variant categories separately for each state. Multinomial logistic regression is a standard technique to model the frequency of SARS-CoV-2 variants[5–7].

We let $V_{j\ell,t}$ to be the probability of a new cases at time t in location ℓ corresponding to variant j. Let $v_{j\ell,t}$ be the analogous observed proportion. Then nonparametric multinomial logistic regression models the log odds as the system

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

This is estimated along with a constraint to ensure that the estimated proportions will sum to 1 across all J variants. The specification of the log odds as a third-order polynomial in time produces smoothness of the estimated proportions. Figure S4 shows the proportions by variant for California before (left) and after (right) the smoothing procedure.

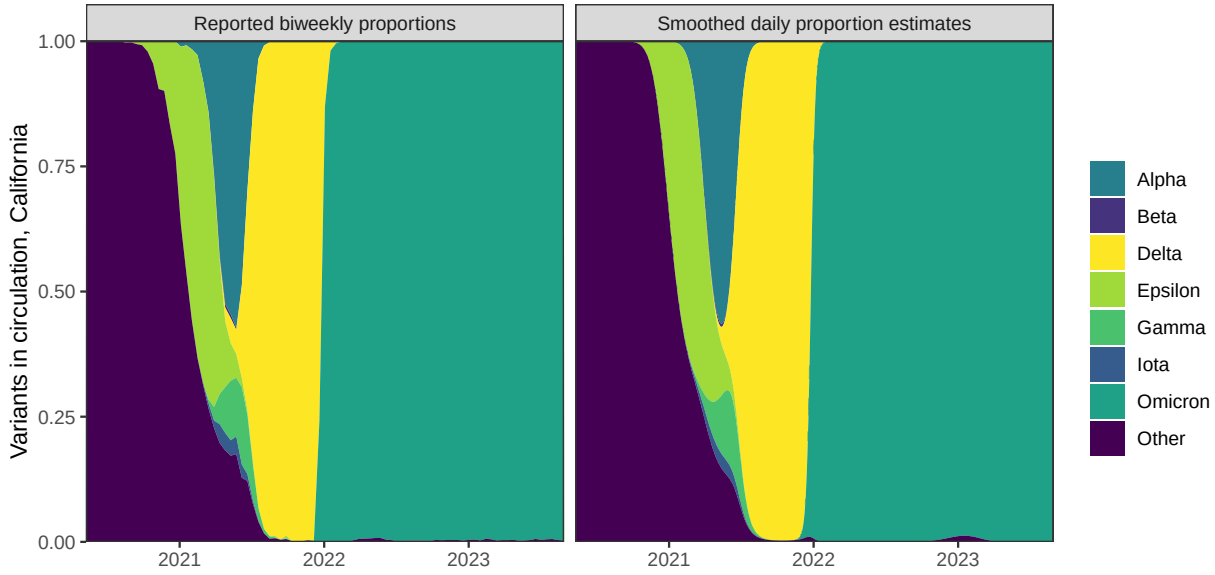

Figure S4: Left: Original biweekly proportions of the variants in circulation for California. Right: Daily proportions of the variants in circulation for California.

S1.5 Constructing the delay from infection to test

The result of Step 1 (Section 4.1) is $\widehat{x}_{\ell,t}$, case estimates by positive specimen date for each state. To continue, pushing this back to infection estimates, we need the variant-specific delays from infection to positive specimen collection. As shown in Figure 1}, this delay can be broken into two separate pieces: (1) the delay from infection to symptom onset, and (2) the delay from symptom onset to positive specimen collection. The first requires different methods and is specific to the variant causing the infection, while the second is estimable from the CDC line list.

S1.5.1 Estimating the incubation period distributions

To account for the incubation period, the time between infection and symptom onset, we use estimates from the existing literature, modified slightly for coherence with each other: we model each incubation as a gamma distribution with different parameters. We focus on the following eight variants (shown in Figure S4, which saw significant circulation in one of the US states during our study period: Ancestral/Other, Alpha, Beta, Epsilon, Iota, Gamma, Delta, and Omicron. Alpha, Beta, Delta, Gamma, and Omicron are all variants of concern[8], while we include the Epsilon (California) and Iota (New York) variants because of large impact on those and neighbouring states[9,10].

The incubation period of the Ancestral variant has been modelled as a gamma distribution[11], so we simply use the reported shape and scale parameters. For the Alpha, Beta, Gamma, Delta and Omicron variants, the mean and standard deviation are reported[12–14]. Therefore, we use method of moments to match the mean and variance to estimate the gamma parameters, using the moment equations given in Section 4.1. Then, we discretize each resulting density shown in Figure S5 to the support set, which is taken to be from 1 and 21 days. This range assumes that symptoms require at least 1 day to develop[15] and that an asymptomatic infection will resolve within 21 days[16,17].

We were unable to locate incubation period estimates for the geo-specific Epsilon and Iota variants, so we use the incubation period for Beta because Epsilon, Iota, and Beta are all children from the same parent in the phylogenetic tree of the Nextstrain Clades[3]. All other circulating variants are grouped together with the Ancestral variant. There was little available sequencing data prior to Alpha-emergence, but unfortunately, later in the pandemic, it is impossible to separate Ancestral from other rare variants, though these also saw minimal circulation after the middle of 2021.

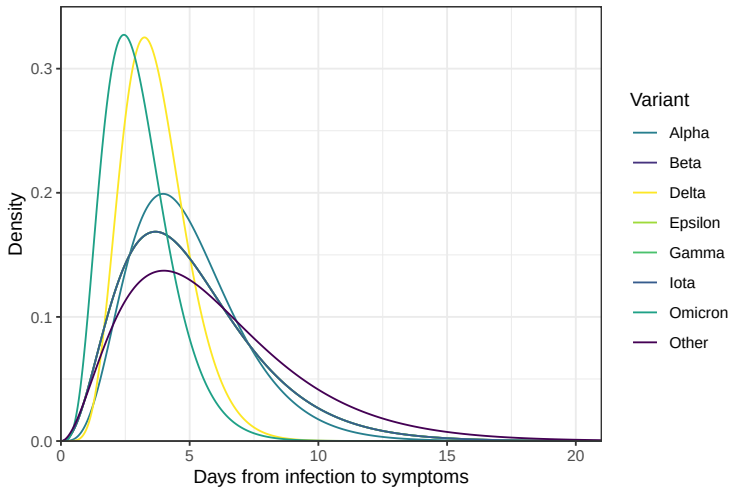

Figure S5: Gamma density for the incubation period of each of the eight variant categories. Note that the Ancestral variant uses reported shape and scale parameters[11], while the remaining variants convert reported estimates for the mean and variance using the method of moments to produce the gamma parameters[12–14].

S1.5.2 Estimating the delay distributions for symptom onset to positive specimen

Estimating the delay from symptom onset to positive specimen date follows a similar procedure as described in Section 4.1 with a minor adjustment. Here, we allow k to range from -3 to 21 (rather than 1 to 60). These upper and lower bounds are based on the largest delay values for the state-wide 0.05 and 0.95 quantiles. The median delay is very short at approximately 2 days, and an asymptomatic individual may test positive following a known exposure, before the onset of symptoms. We show both types of delays for a sample of states over several dates in Figure S6. Unlike the delay from positive specimen collection to report, the delay from symptoms to positive specimen can conceivably be negative. The most obvious reason for this would be if a person knew they had been exposed to an infectious individual and so got tested prior to the development of symptoms. Required regular testing for jobs in health care settings, construction, or the film industry could also produce negative delays.

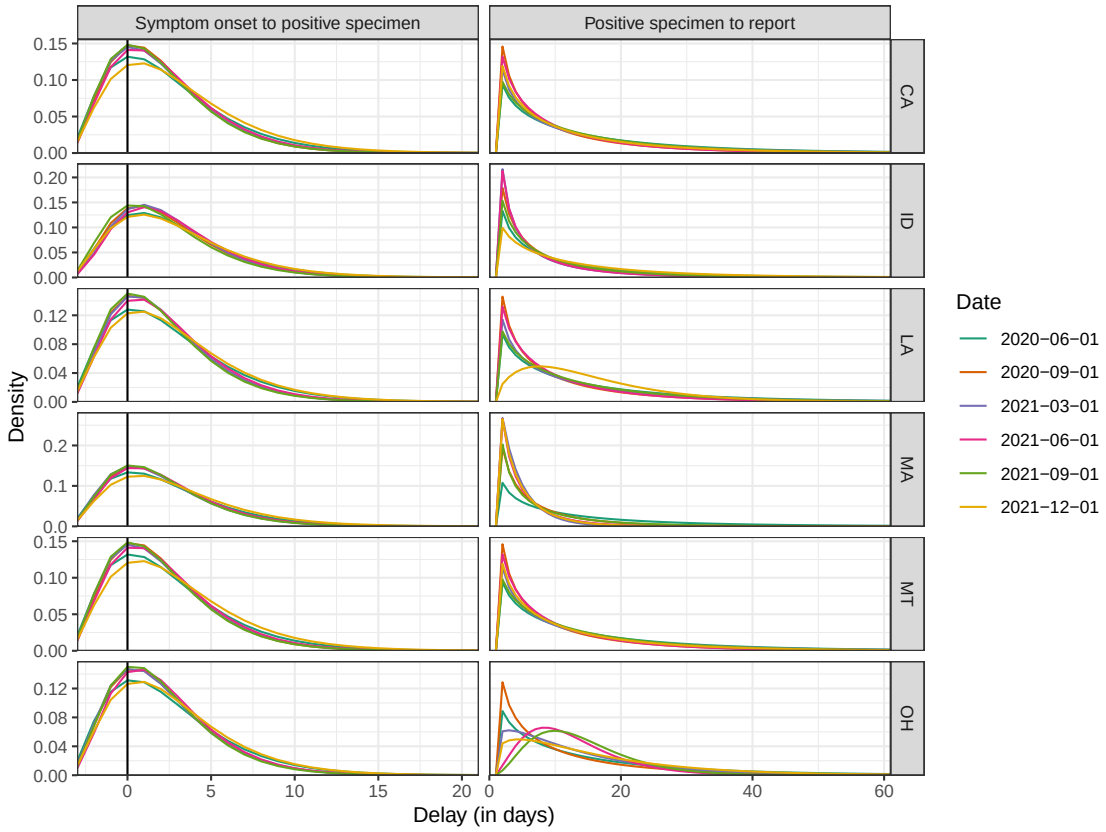

Figure S6: Depictions of the estimated delay from symptom onset to positive specimen date (left) and from positive specimen date to report date (right) for a sample of six states over several dates.

S1.5.3 Details on constructing the infection-to-test distributions

Finally, to produce the delay from infection to positive specimen collection we convolve the variant-specific incubation periods from Section S1.5.1, denoted by $\{i_j(k) : 0 < k \leq 21\}$ with the location-time-specific symptom-to-positive-test distributions from Section S1.5.2, denoted by $\{q_{\ell,t}(k) : -3 \leq k \leq 21\}$. The convolution of these yields a distribution $\mathbf{q}_{\ell,t} * \mathbf{i}_j = \{\tau_{j\ell,t}(k) : -3 \leq k \leq 42\}$. Figure S6 shows the delays used for a sample of 6 states: symptom to positive specimen (left column) and positive specimen to report (right column). The convolution distribution $\tau_{j\ell,t}$ requires convolving the distribution in the left column with the variant-specific incubation periods shown in Figure S5.

S1.6 Details about seroprevalence data

We use two major contemporaneous surveys to estimate the proportion of the population with evidence of previous infection in each state over time: the 2020–2021 Blood Donor Seroprevalence Survey and the Nationwide Commercial Lab Seroprevalence Survey[18,19]. In the former, the CDC collaborated with 17 blood collection organizations in the largest nationwide COVID-19 seroprevalence survey to date[18]. The blood donation samples were used to construct monthly seroprevalence estimates for nearly all states from July 2020 to December 2021[20]. In the latter survey, the CDC collaborated with two private commercial laboratories to test blood samples from people that were in for routine or clinical management (presumably unrelated to COVID-19[21]) for the antibodies to the virus. 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 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[20–22]. 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 S7). For example, in the commercial dataset, the last estimate for North Dakota is in September 2020. In the blood donor dataset, Arkansas does not have estimates available until October 2020.

A major difference in the structure of the two datasets is that the commercial dataset always has the seroprevalence estimates at the level of the state, while the blood donor dataset can either have estimates for the state or for multiple separate regions within the state. For the commercial dataset, we use the midpoint of the provided specimen collection date variable. For the blood donor dataset, we use the median donation date if the seroprevalence estimates are designated to be for entire state. If they are instead for regions in the state, since there is reliably one measurement per region per month, we aggregate the measurements into one per month per state by using a weighted average (to account for the given sample sizes of the regions). The median of the median dates is taken to be the date for the weighted average. If there are multiple measurements in a week from a seroprevalence source, then the average is used.

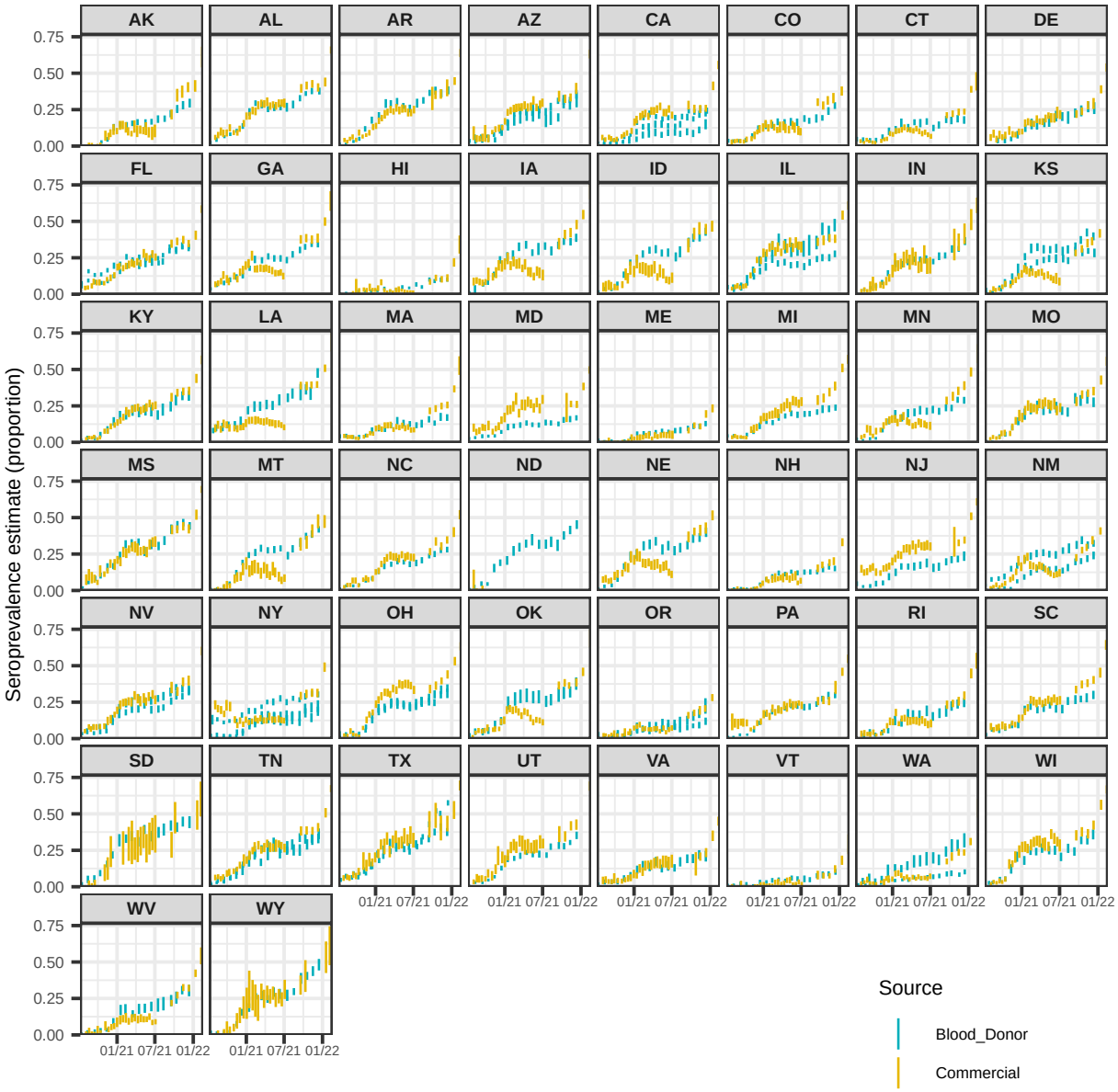

Figure S7: A comparison of the seroprevalence estimates from the Commercial Lab Seroprevalence Survey dataset (yellow) and the 2020–2021 Blood Donor Seroprevalence Survey dataset (blue). Note that the maximum and the minimum of the line ranges are the provided 95% confidence interval bounds to give a rough indication of uncertainty.

S1.7 State space representation of the antibody prevalence model

The antibody prevalence model can be expressed as a linear Gaussian state space model[23,24]. For $m = 1, \ldots, M$, let α_m be a vector of latent state processes at time m and y_m be a vector

of observations at time m. The form of the (general) linear Gaussian state space model is

$$y_m = Z\alpha_m + \sigma_r^2 \epsilon_m, \qquad \epsilon_m \sim N(0, H_m) \tag{3}$$

$$\alpha_{m+1} = T_m \alpha_m + R\eta_m, \quad \eta_m \sim N(0, Q) \tag{4}$$

where $\alpha_1 \sim N(a_1, P_1)$ and ϵ_m and η_m are mutually and serially independent.

To express the antibody prevalence model in state space form, we define the components in Equations (3) and (4) as follows (omitting the location subscript for simplicity):

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad Z = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad H_m = \begin{bmatrix} w_m^1 & 0 \\ 0 & w_m^2 \end{bmatrix}$$

$$\alpha_m = \begin{bmatrix} s_m \\ a_m \\ a_{m-1} \\ a_{m-2} \end{bmatrix} \qquad T_m = \begin{bmatrix} (1-\gamma) & \widehat{u}_m^{\Sigma}(1-z_m) & 0 & 0 \\ 0 & 3 & -3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad Q = \begin{bmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{bmatrix}$$

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

where σ_r^2 is the variance of observations, σ_ϵ^2 is the variance of the seroprevalence estimates, σ_η^2 is the trend variance, and $\widehat{u}_m^{\Sigma}$ denotes the new deconvolved cases between m and $m+1$. Because the inverse reporting ratios should be more variable than the seroprevalence estimates, we enforce that the estimate of σ_η^2 is a multiple of σ_ϵ^2.

Finally, w_m^1 and w_m^2 are the time-varying inverse variance weights computed from the commercial and blood donor datasets, respectively. For each source, we compute the weights for the observed seroprevalence estimates using the formula for the standard error of a proportion. These weights are then re-scaled to sum to the number of observed seroprevalence measurements for the source. Finally, the ratio of the average observed weights from the two sources is used to scale all weights relative the commercial source. This transformation is done purely for computational purposes to make estimation of the unknown parameters easier.

The prior distribution for α_1 is estimated using both data-driven constraints and externally sourced information. The initial value of the seroprevalence component, $\tilde{s}_1$ is the average of the initial seroprevalence measurements from each source rounded down to two decimal places, The corresponding initial variance estimate, $\sigma_{\tilde{s}_1}^2$, is taken to be the mean of the standard errors of the two seroprevalence estimates. For the initial values of the trend components, we use the inverse of the ascertainment ratio estimate for each as of June 1, 2020 from Table 1 in[25]: the mean for $\tilde{a}_1$ and half the width of the confidence interval converted to the inverse for $\sigma_{\tilde{a}_1}^2$.

The initial σ_r^2 is taken to be the average of the estimated variances from the observed seroprevalence measurements regressed linearly on time. The initial value of the multiplier is set to be 100 for all states. The σ_ϵ^2 and γ values are estimated separately for each state, then fixed to their averages on the log-scale.

Following maximum likelihood estimation of remaining parameters we use the Kalman filter to obtain the smoothed point estimates and variances of the weekly inverse reporting ratios. We use forward and backward extrapolation to extend these estimates outside of the observed seroprevalence range[23], followed by linear interpolation to produce daily values. We then multiply these by the corresponding deconvolved case estimates before converting to per-capita values. Annual estimates of the state populations as of July 1 of 2020 are taken from the Dec. 2022 press release from the US Census Bureau[2].

Supplementary References

1. Jahja, M., Chin, A. & Tibshirani, R. J. Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. *Statistical Science* **37**, 207–228 (2022).
2. U.S. Census Bureau, Population Division. Annual estimates of the resident population for the United States, regions, states, District of Columbia, and Puerto Rico: April 1, 2020 to july 1, 2022. (2022).
3. Hodcroft, E. CoVariants: SARS-CoV-2 mutations and variants of interest. (2021).
4. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).
5. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).
6. Annavajhala, M. K. *et al.* Emergence and expansion of SARS-CoV-2 B. 1.526 after identification in New York. *Nature* **597**, 703–708 (2021).
7. Figgins, M. D. & Bedford, T. SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. *MedRxiv* 2021–12 (2021).
8. World Health Organization. Tracking SARS-CoV-2 variants. (2021).
9. Yang, S. *et al.* Investigation of SARS-CoV-2 Epsilon variant and hospitalization status by genomic surveillance in a single large health system during the 2020-2021 winter surge in Southern California. *American Journal of Clinical Pathology* **157**, 649–652 (2022).
10. Duerr, R. *et al.* Dominance of Alpha and Iota variants in SARS-CoV-2 vaccine breakthrough infections in New York City. *The Journal of Clinical Investigation* **131**, e152702 (2021).
11. Tindale, L. C. *et al.* Evidence for transmission of COVID-19 prior to symptom onset. *eLife* **9**, e57149 (2020).
12. Tanaka, H. *et al.* Shorter incubation period among COVID-19 cases with the BA. 1 Omicron variant. *International Journal of Environmental Research and Public Health* **19**, 6330 (2022).
13. Grant, R. *et al.* Impact of SARS-CoV-2 Delta variant on incubation, transmission settings and vaccine effectiveness: Results from a nationwide case-control study in France. *The Lancet Regional Health–Europe* **13**, 100278 (2022).
14. Ogata, T., Tanaka, H., Irie, F., Hirayama, A. & Takahashi, Y. Shorter incubation period among unvaccinated delta variant coronavirus disease 2019 patients in Japan. *International Journal of Environmental Research and Public Health* **19**, 1127 (2022).
15. Public Health Agency of Canada. COVID-19 for health professionals: transmission. (2021).

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

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

18. Centers for Disease Control and Prevention. 2020-2021 nationwide blood donor seroprevalence survey infection-induced seroprevalence estimates. (2021).

19. Centers for Disease Control and Prevention. Nationwide commercial laboratory seroprevalence survey. (2021).

20. Jones, J. M. *et al.* Estimated US infection-and vaccine-induced SARS-CoV-2 seroprevalence based on blood donations, July 2020-May 2021. *JAMA* **326**, 1400–1409 (2021).

21. Bajema, K. L. *et al.* Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. *JAMA Internal Medicine* **181**, 450–460 (2021).

22. Centers for Disease Control and Prevention. COVID Data Tracker. (2020).

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

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

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