Incident COVID-19 infections before Omicron in the U.S.

Rachel Lobay[a,1], Ajitesh Srivastava[b], Ryan J. Tibshirani[c], and Daniel J. McDonald[a]

[a]Department of Statistics, The University of British Columbia
[b]Department of Computer and Electrical Engineering, University of Southern California
[c]Department of Statistics, The University of California, Berkeley

Abstract

The timing and magnitude of COVID-19 infections are of interest to the public and to public health, but these are challenging to ascertain due to the volume of undetected asymptomatic cases and reporting delays. Accurate estimates of COVID-19 infections based on finalized data can improve understanding of the pandemic and provide more meaningful quantification of disease patterns and burden. Therefore, we retrospectively estimate daily incident infections for each U.S. state prior to Omicron. To this end, reported COVID-19 cases are deconvolved to their date of infection onset using delay distributions estimated from the CDC line list. Then, a novel serology-driven model is used to scale these deconvolved cases to account for the unreported infections. The resulting infections incorporate variant-specific incubation periods, reinfections, and waning antigenic immunity. They clearly demonstrate that the reported cases fail to reflect the full extent of disease burden in all states. Most notably, infections were severely underreported during the Delta wave, with an estimated reporting rate as low as 6.3% in New Jersey, 7.3% in Maryland, and 8.4% in Nevada. Moreover, in 44 states, fewer than 1/3 of infections appear as cases reports. Therefore, while reported cases offer a convenient proxy of disease burden, they fail to capture the full extent of infections, and can severely underestimate the true disease burden. This retrospective analysis also estimates other important quantities for every state, including variant-specific deconvolved cases, time-varying case ascertainment ratios, and infection-hospitalization ratios.
Keywords: COVID-19; SARS-CoV-2; Infections; Deconvolution; Time series; Seroprevalence; Antibody

1 Introduction

Reported COVID-19 cases are a staple in tracking the pandemic at varying geographic resolutions (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 have occurred, and likely much earlier. To see why, it is important to understand *whose* cases are being reported and what differentiates them from unreported cases as well as *when* these case reports happen. Figure 1 shows an idealized path of a symptomatic infection that is eventually reported to public health. This figure illustrates a number of sources of bias in the reporting pipeline. For instance, diagnostic testing mainly targets symptomatic individuals; thus, infected individuals exhibiting little to no symptoms are omitted (Centers for Disease Control and Prevention, 2022). In addition, testing practices, availability, and uptake vary temporally and spatially (European Centre for Disease Prevention and Control, 2020; Hitchings et al., 2021; Pitzer 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 eventual submission to public health (Pellis et al., 2021; Washington State Department of Health, 2020). For these reasons, reported cases are lagging indicators of the course of the pandemic. Furthermore, they do not represent the actual number of new infections that occur on any given day based on exposure to the pathogen. Since there was no large-scale surveillance effort in the United States that reliably tracked symptom onset, let alone infection onset, ascertaining the onset of all *infections* is challenging.

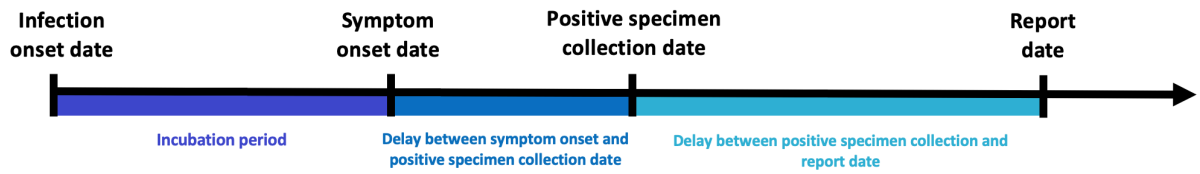

Figure 1: Idealized chain of events from infection onset to case report date for a symptomatic infection that is eventually reported to public health.

Contextualizing the course of the pandemic, understanding the effects of interventions, and drawing insights for future pandemics is challenging because the spatial and temporal behaviour of infections is unknown. While reported cases provide a convenient proxy of the disease burden in a population, it is incomplete, delayed, and misrepresents the true size and timing of the pandemic. Regardless of these difficulties, it is important to the public and to public health to perform a pandemic post-mortem. Estimates of daily incident infections are one such way to measure this and can guide understanding of the pandemic burden over space and time.

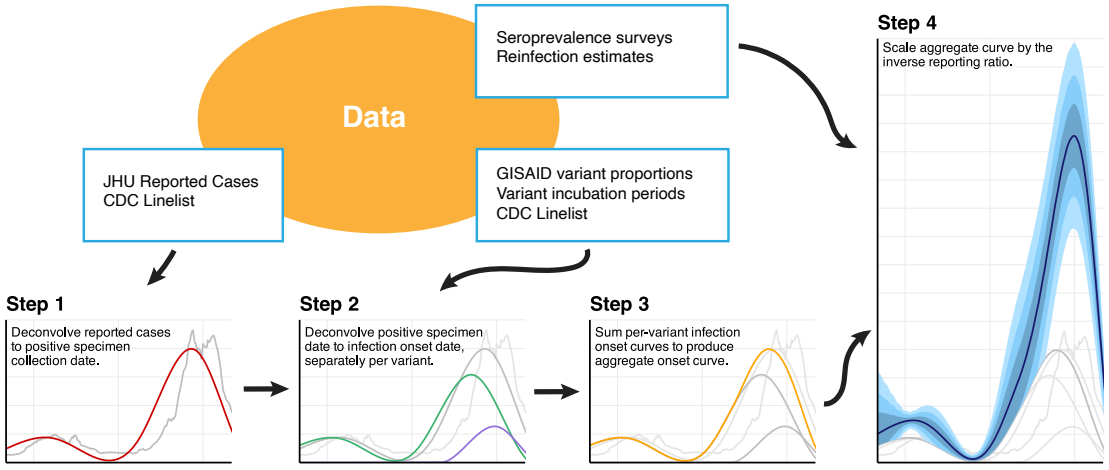

Figure 2: Flowchart of the data and major analysis steps required to get from reported cases to incident infection estimates. In Step 1, we use the CDC line list data to deconvolve reported cases (grey) backward to the date of positive specimen (red). Step 2 separately deconvolves these to the date of infection by variant (Epsilon in Purple, Ancestral in Green), before summing across all variants (orange) in Step 3. Finally, we use seroprevalence survey and time-varying reinfection data to account for the unreported infections.

In this work, we provide a data-driven reconstruction of daily incident infections for each U.S. state before the onset of Omicron. Using state-level line list data, we estimate state-date specific distributions for the delay from symptom onset to positive specimen date and positive specimen to case report date. We combine these with variant-specific incubation period distributions to deconvolve daily reported COVID-19 cases back to their infection onset, removing the effects of the delays. Finally, we adjust for unreported infections with seroprevalence and reinfection data, accounting for the waning of antigenic immunity over time. A graphical depiction of our procedure is shown in Figure 2. Our results examine features of our infection estimates and the implications of using them, rather than reported cases, to assess the impact of the pandemic. We also produce simple time-varying infection-hospitalization ratios (IHRs) for each state and compare these with case-hospitalization ratios (CHRs). While these analyses provide a glimpse into the utility of our infection estimates, we believe that there is much more to be explored, and we hope that our work serves as a benchmark for future retrospective analyses.

2 Methods

In what follows, we describe how we estimate the daily incident infections for each U.S. state from June 1, 2020 to November 29, 2021. Figure 2 summarizes the major analysis tasks. First, we estimate the delays from positive specimen to report date and use them to push back the reported cases to their sample collection dates. Next, we estimate the delay from symptom onset to sample collection, combine this with variant-specific infection-to-symptom delays, and use these to push back the cases to infection onset. The resulting case estimates are aggregated across variant categories and adjusted by the case ascertainment ratio, estimated with seroprevalence survey data and a model for antigenic immunity.

2.1 From reported cases to positive specimen collection

Deconvolution "pushes back" reported cases to the likely date of positive specimen collection. An important aspect of our methods is that deconvolution is not the same as a simple shift, rather it involves the distribution of delays (specific to each state and date). Simply shifting cases back in time would fail to reflect the fact that some cases take much longer to be reported than others (Appendix A).

We will start by describing how the model for deconvolution infers the likely dates of positive specimen collection from reported cases before describing how the CDC line list (Centers for Disease Control and Prevention, 2020a) was used to estimate the necessary delay distributions. Together, these are the ingredients for Step 1 in Figure 2. Define $y_{\ell,t}$ to be the number of new cases reported in location ℓ at time t, as reported by the John Hopkins Center for Systems Science and Engineering (JHU CSSE, Dong et al., 2020) and retrieved with the COVIDcast API (Reinhart et al., 2021). Let $\pi_{\ell,t}(k)$ be the probability that cases with positive specimen collection at time $t-k$ are reported at t. Then, we model $y_{\ell,t}$ as a Gaussian with mean

$$\mathbb{E}[y_{\ell,t} \mid x_{\ell,s}, s \leq t] = \sum_k \pi_{\ell,t-k}(k) x_{\ell,t-k}, \tag{1}$$

which is a probability weighted sum of the number of positive specimens collected k days earlier, $x_{\ell,t-k}$. We estimate $\mathbf{x}_\ell = \{x_{\ell,1}, \ldots, x_{\ell,T}\}$ by minimizing the negative log-likelihood with a penalty that encourages smoothness in time. Thus, our estimator is given by

$$\widehat{\mathbf{x}}_\ell = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_t \left(y_{\ell,t} - \sum_k \pi_{\ell,t-k}(k) x_{t-k} \right)^2 + \lambda \sum_t \left| x_t - 4x_{t-1} + 6x_{t-2} - 4x_{t-3} + x_{t-4} \right|. \tag{2}$$

The solution to this minimization problem is an adaptive piecewise cubic polynomial (Tibshirani, 2014, 2022) and can be accurately computed with ease (Jahja et al., 2022; Ramdas and Tibshirani, 2016). We select the tuning parameter λ with cross-validation to minimize the out-of-sample reconvolution error.

To estimate the $\pi_{\ell,t}(k)$ for all states ℓ, times t, and delays k, we use the CDC line list (Centers for Disease Control and Prevention, 2020a). The line list contains three key dates of interest for many cases that eventually appear in case reports: symptom onset, positive specimen collection, and report to the CDC. Handling missingness in these dates requires careful attention (Appendix B). Define $z_{\ell,t}$ to be a case report occurring at time t in location ℓ. We assume that positive samples are reported within 60 days and that no test is reported on the same date as it was collected. Under these assumptions, let $N_{\ell,t}$ be the total number of $z_{\ell,r}$ with positive specimen collection date r in a window $r \in [t-75+1, t+60]$ around t. Then, we compute the observed probability mass function (pmf)

$$\tilde{p}_{\ell,t}(k) = \frac{1}{N_{\ell,t}} \big(\#\ z_{\ell,r} \text{ with positive specimen at } r-k\big) \mathbf{1}(0 < k \leq 60), \tag{3}$$

where $\mathbf{1}(Z) = 1$ if Z is true and 0 otherwise. We also compute a similar national pmf, $\tilde{p}_t(k)\mathbf{1}(0 < k \leq 60)$, without restricting to location ℓ. Next, let $\alpha_{\ell,t}$ be the ratio of $N_{\ell,t}$ to the number of cases reported by JHU CSSE (Dong et al., 2020) in the window $[t-60+2, t+75]$. Then, compute $p_{\ell,t} = \alpha_{\ell,t}\tilde{p}_{\ell,t} + (1-\alpha_{\ell,t})\tilde{p}_t$. This construction was adopted to allow for more reliance on the state estimate when a larger fraction the JHU cases reports appear in the CDC line list. We calculate the mean $m_{\ell,t}$ and variance $v_{\ell,t}$ of the pmf $\{p_{\ell,t}(k)\}$ and estimate the best-fitting gamma distribution by solving the moment equations $m_{\ell,t} = \alpha_{\ell,t}\theta_{\ell,t}$ and $v_{\ell,t} = \alpha_{\ell,t}\theta_{\ell,t}^2$ for the shape $\alpha_{\ell,t}$ and scale $\theta_{\ell,t}$. Finally, we discretize the resulting gamma density to

the original support to produce an estimate $\widehat{\pi}_{\ell,t}(k)$ of the delay distribution $\pi_{\ell,t}(k)$. Additional details are deferred to Appendix C.

2.2 From positive specimen collection to infection onset

To continue, pushing positive specimen collection time back to infection onset (Step 2 in Figure 2), we use a procedure very similar to that described above and specified in Equations (1) and (2). However, because the delays involve the time from infection to symptom onset, these must be variant-specific. We use our estimates from Section 2.1, $\widehat{\mathbf{x}}_\ell$, but we weight them corresponding to the mix of variants in circulation. To estimate the daily proportions of the variants circulating in each state, we use GISAID genomic sequencing data from CoVariants.org (Elbe and Buckland-Merrett, 2017; Hodcroft, 2021), and estimate a multinomial logistic regression model. This procedure is now standard (Appendix D) (Annavajhala et al., 2021; Figgins and Bedford, 2021; Obermeyer et al., 2022). The resulting estimated probability of variant j is given by $\hat{v}_{j\ell,t}$.

To estimate variant-specific delays from infection to positive specimen collection, we convolve the location-time-specific symptom-to-test distributions (that are estimated from the CDC line list in the same way as in Section 2.1), with variant-specific incubation periods. The convolution of these yields a distribution $\widehat{\tau}_{j\ell,t}(k)$. Details on the convolution and its inputs are in Appendices F to H.

Analogous to Equations (1) and (2), for each variant j, we model the variant-specific, deconvolved cases as Gaussian with mean

$$\mathbb{E}\left[\widehat{v}_{j\ell,t}\widehat{x}_{\ell,t} \mid u_{j\ell,s}, s \le t\right] = \sum_k \widehat{\tau}_{j\ell,t-k}(k) u_{j\ell,t-k} \tag{4}$$

and estimate $\mathbf{u}_{j\ell}$ by minimizing the negative loglikelihood with a penalty to encourage smoothness:

$$\widetilde{\mathbf{u}}_{j\ell} = \underset{\mathbf{u}}{\operatorname{argmin}} \sum_t \left(\widehat{v}_{j\ell,t}\widehat{x}_{\ell,t} - \sum_k \widehat{\tau}_{j\ell,t-k}(k) u_{t-k} \right)^2 + \lambda \sum_t \left| u_t - 4u_{t-1} + 6u_{t-2} - 4u_{t-3} + u_{t-4} \right|. \tag{5}$$

We call the solution $\widetilde{\mathbf{u}}_{j\ell}$ the *variant-specific deconvolved cases* and emphasize that these are cases that will eventually be reported to public health. Because this deconvolution is performed separately for each location and variant, we sum over the variants at each time t, and denote the total deconvolved cases at location ℓ as $\widehat{\mathbf{u}}_\ell = \sum_j \widetilde{\mathbf{u}}_{j\ell}$ (Step 3 in Figure 2). Note that these deconvolved cases are now indexed by the time of infection onset rather than case report.

2.3 Inverse reporting ratio and the antibody prevalence model

To capture the unreported infections, it is necessary to adjust these deconvolved case estimates by the inverse reporting ratio, the ratio of the number of incident infections to incident reported infections (Step 4 in Figure 2). Seroprevalence of anti-nucleocapsid antibodies represents the percentage of people who have at least one resolving or past infection (Centers for Disease Control and Prevention, 2020b), so we develop a model that uses the change in subsequent seroprevalence measurements to estimate all new infections. We use two seroprevalence surveys to estimate the proportion of the population with evidence of previous infection in each state over time (Appendix I) (Centers for Disease Control and Prevention, 2021a,b).

To account for different surveys occurring on different dates with roughly weekly availability and measurement error, we treat actual seroprevalence $s_{\ell,m}$ as a latent variable available on Monday (using m rather than t to denote Mondays). Therefore, the observed seroprevalence survey measurements r^1_m and r^2_m are modelled as Gaussian,

$$r^1_{\ell,m} \mid s_{\ell,m},\ w^1_{\ell,m} \sim \mathrm{N}(s_{\ell,m},\ w^1_{\ell,m}\sigma^2_{\ell,r}), \tag{6}$$
$$r^2_{\ell,m} \mid s_{\ell,m},\ w^2_{\ell,m} \sim \mathrm{N}(s_{\ell,m},\ w^2_{\ell,m}\sigma^2_{\ell,r}), \tag{7}$$

with source-specific measurement errors, $w^1_{\ell,m}$ and $w^2_{\ell,m}$, that scale proportional to reported uncertainty.

To complete the model, we suppose that latent seroprevalence is modeled as Gaussian with mean given by a fraction of the previous seroprevalence measurement at m plus the reinfection-adjusted deconvolved cases

multiplied by the inverse reporting ratio at time m:

$$\mathbb{E}[s_{\ell,m+1} \mid s_{\ell,m}] = (1-\gamma)s_{\ell,m} + a_{\ell,m}(1-z_m) \sum_{t\in[m,m+1]} \widehat{u}_{\ell,t}, \tag{8}$$

where $\widehat{u}_{\ell,t}$ are deconvolved cases (from Section 2.2), z_m is the fraction of reinfections, and $a_{\ell,m}$ is the inverse reporting ratio. Note that γ is the fraction of people whose level of infection-induced antibodies falls below the detection threshold between time t and time $t+1$. The daily fraction of new infections z_t are based on surveillance work conducted by the Southern Nevada Health District (Ruff et al., 2022), and these estimates are broadly similar to those in other locations with available data (Hawaii Department of Health, 2022; New York State Department of Health, 2023; Ruff et al., 2022; Washington State Department of Health, 2022). Finally, we specify the time-varying evolution of the inverse reporting ratio as Gaussian with expectation,

$$\mathbb{E}[a_{\ell,m+1} \mid a_{\ell,m},\ a_{\ell,m-1},\ a_{\ell,m-2}] = 3a_{\ell,m} - 3a_{\ell,m-1} + a_{\ell,m-2}. \tag{9}$$

This construction for Equation (9) results in estimates that vary smoothly in time.

The antibody prevalence model specified by Equations (6) to (9) is a state space model with latent variables $\mathbf{s}_\ell$ and $\mathbf{a}_\ell$. In this way, the latent variables and all unknown parameters can be estimated using maximum likelihood, despite missing or irregularly-spaced survey measurements. Additionally, latent quantities can be extrapolated beyond the times of measured seroprevalence. Additional details of this methodology and the computation of the associated uncertainty measurements are in Appendix J.

2.4 Lagged correlation to hospitalizations and time-varying IHRs

From the COVIDcast API (Reinhart et al., 2021), we retrieve the daily number of confirmed COVID-19 hospital admissions for each state that are collected by the U.S. Department of Health and Human Services (HHS). We use our infection estimates $\widehat{\mathbf{u}}_\ell$ to compute the lagged correlation with hospitalizations. The goal of this analysis is to find the lag between infection and hospitalization rates that gives the highest average rank-based correlation across U.S. states. Thus, we consider a wide range of possible lag values ranging from 1 to 25 days. Then, to assess the impact of our modelling choices, particularly the contribution of the main steps to the lagged correlation analysis, we conduct an ablation study that is detailed in Appendix K.

For each considered lag, we calculate Spearman's correlation between the state infection and hospitalization rates for each observed between June 1, 2020 to November 29, 2021 with a center-aligned rolling window of 61 days. We then average these correlations across all states and times for each lag.

The lag that leads to the highest average correlation is used to estimate the time-varying IHRs for each state. The IHR is computed by dividing the number of individuals who are hospitalized due to COVID-19 by the estimated total number who were infected on the lagged number of days before. To stabilize these lagged IHR estimates, we average these hospitalizations and infections within a window of 31 days centered on the date of interest, rather than just using one pair of dates for each computation.

3 Results

3.1 Infection estimates and cases-to-infections ratios across the U.S. states

Prior to Omicron, the largest infection outbreaks were observed in the late summer and early fall of 2021 in Louisiana, Georgia, Idaho, and Montana (Figures 3 to 4). During this time, the state with the highest rate of infections on a single day is Louisiana, with 476 infections per 100K on July 20, 2021. For comparison, the state's 7-day average case rate peaks at 126 cases per 100K on August 13, 2021. Idaho follows with an infections peak of 457 per 100K on September 7, 2021, and a case peak of 76 per 100K occurring shortly thereafter on September 13, 2021. The period of lowest viral transmission is observed in the summer of 2020, when Vermont has fewer than 10 infections per 100K per week from June to August, the longest such lull observed for any state.

Nearly all states exhibit two major waves in infections—the Ancestral wave began in the fall of 2020 and extended into the winter season, while the Delta wave started in the late summer of 2021 and continued into mid-fall. In general, greater similarities in the strength and magnitude of outbreaks emerge in small clusters

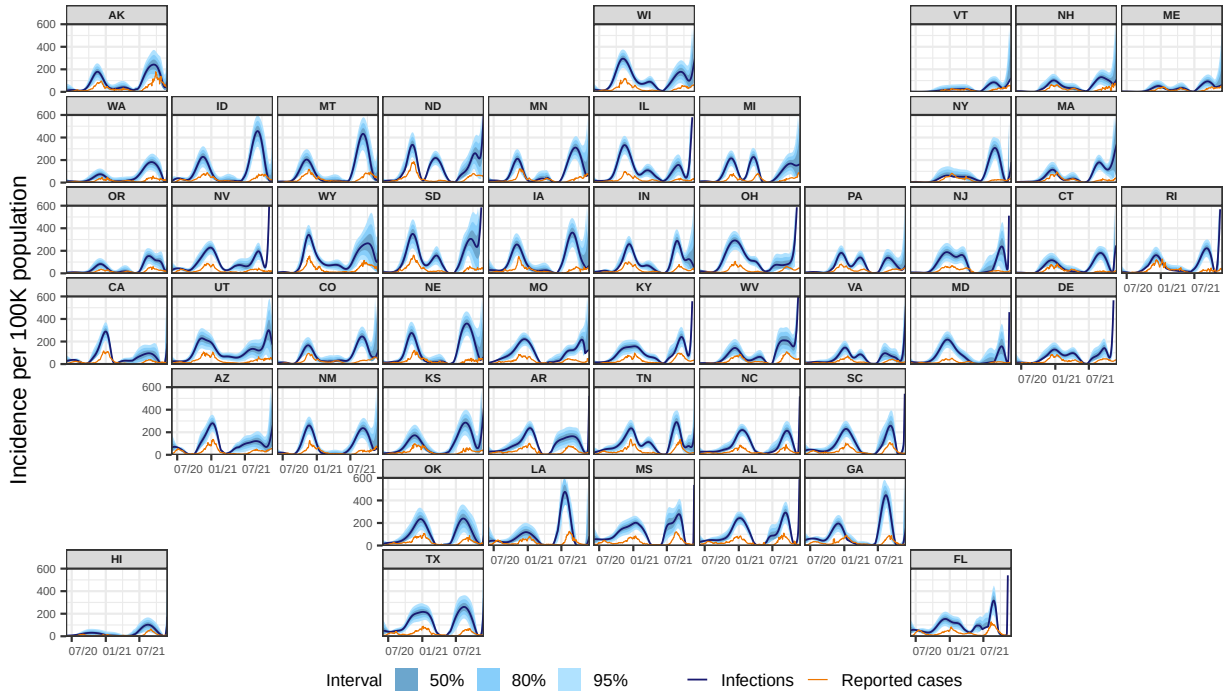

Figure 3: Estimates of the daily new infections per 100,000 population for each U.S. state from June 1, 2020 to November 29, 2021 (dark blue line). The blue shaded regions depict the 50, 80, and 95% intervals for the estimates, while the orange line represents the trailing 7-day average of reported cases per 100,000.

of states that border each other (Idaho and Montana; North and South Carolina) present waves of infections that mirror each other in amplitude and timing.

While the Ancestral, Alpha, and Delta waves are visible for most states, there are clear outbreaks in unreported infections that are not easily detectable from cases alone. For example, a wave of infections is evident in North and South Dakota over the spring of 2021 that is virtually undetectable from reported cases. Similarly, in late-summer 2021, the Delta wave is only faintly detectable from cases in a number of Northeastern states, while infections suggest that it has already begun in earnest.

Moreover, cases tend to severely underestimate infections during Delta for many states, more so than in earlier waves (Figure 3). The most extreme was New Jersey, where about 6.3% of estimated infections were eventually reported as cases. Similarly low are Maryland (7.3%), Nevada (8.4%), and South Dakota (10.0%). In 44 states, fewer than 1/3 of infections eventually appear in case reports. The cases-to-infections ratio was larger in earlier waves, and its effects were most apparent in different regions. During Alpha, Louisiana had the lowest ratio of infections to cases (11.9%) followed by California (13.6%). Such patterns are less apparent during the Ancestral wave, where Ohio and Maryland had the lowest ratio of reported cases to infections at 21.4% and 21.7%, respectively.

Figure 5 shows that using cases as a proxy for infections can lead to misunderstandings in the locations that are affected and the extent to which they are affected. For example, on October 20, 2020, while case rates are elevated in a handful of upper-Midwestern states (namely, North and South Dakota), infection rates are elevated to a similar extent in the surrounding states as well, indicating a wider impact than suggested by cases alone. On July 20, 2021, while the map of case rates shows low and geographically consistent impact, infection rates reveal that Texas, Louisiana, Georgia, and their neighbors are hotspots.

By focusing on states with elevated cases, infection outbreaks may be overlooked. For instance, on August 27, 2021, Montana and Idaho have some of the highest infection rates (Figure 5). In contrast, their case rates are unremarkable (the highest case rates tend to be in the Southeast). Infection outbreaks tend to precede case outbreaks, though the lead time can vary widely. During the Delta wave, infections in Montana peaked about 41 days before cases, while in Idaho, they peaked about 6 days before cases (Figure 3). During the

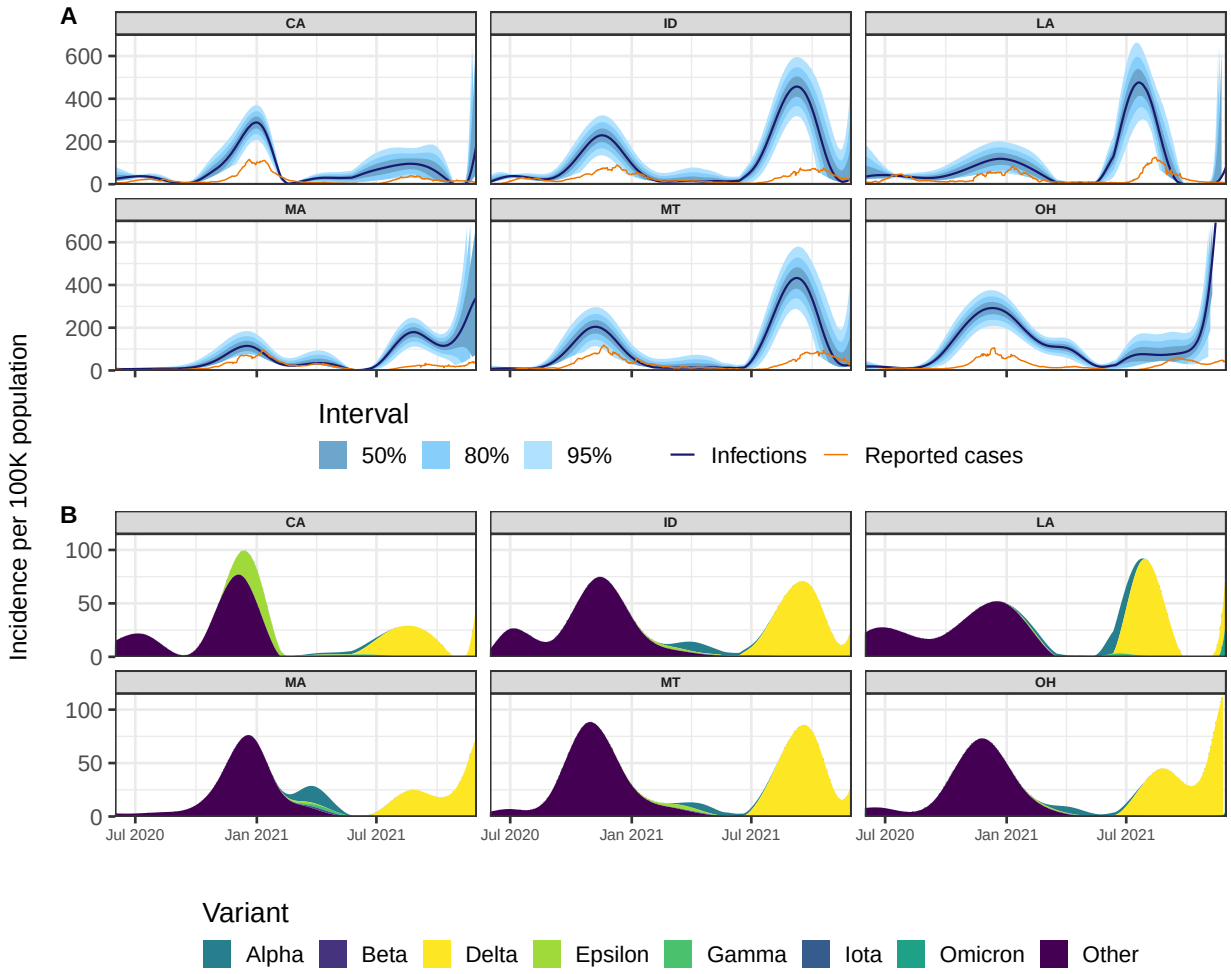

Figure 4: Panel A: Reported cases (orange) and estimates of daily new infections (dark blue) per 100K inhabitants. The blue shaded regions indicate 50, 80, and 95% confidence bands. Panel B: Deconvolved cases colored by variant per 100K inhabitants.

Ancestral wave, infections peaked about 12 days earlier than cases in Montana and 24 days earlier in Idaho, demonstrating a notable shift in lead times.

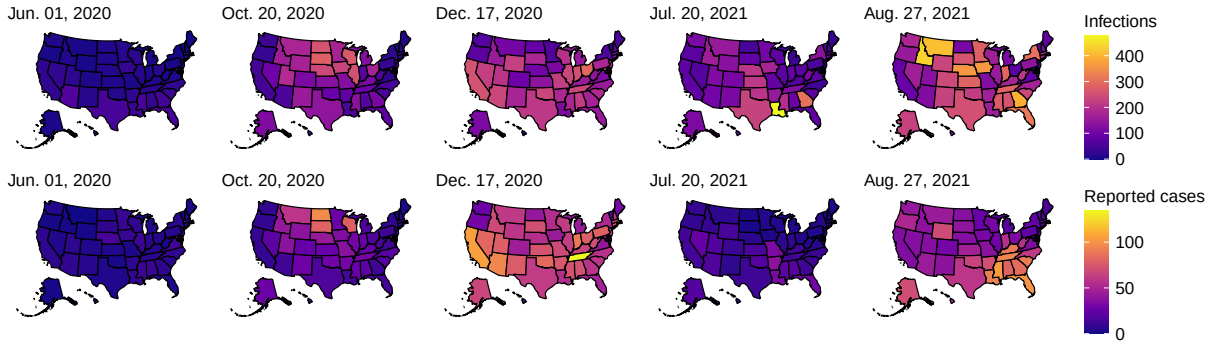

Figure 5: Choropleth maps of the state-level estimates of the daily new infections per 100K (top row) and the daily new cases per 100K (bottom row) for five select dates between June 1, 2020 and November 29, 2021. Note that the first date was chosen as a baseline, while the other dates were chosen because they present large counts of infections across all states. In particular, the third and fifth dates present the largest number of total infections across the 50 states within those calendar years.

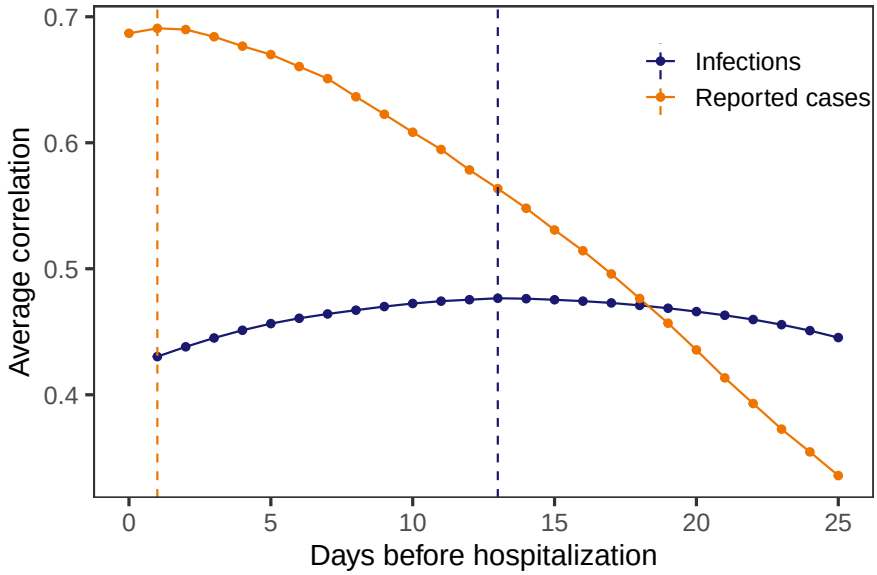

Figure 6: Spearman's rank correlation between each of infections and cases with hospitalizations per 100,000. A rolling window of 61 days is applied before averaging across all states and times for each lag. The vertical dashed lines indicate the lags for which the highest average correlation is attained.

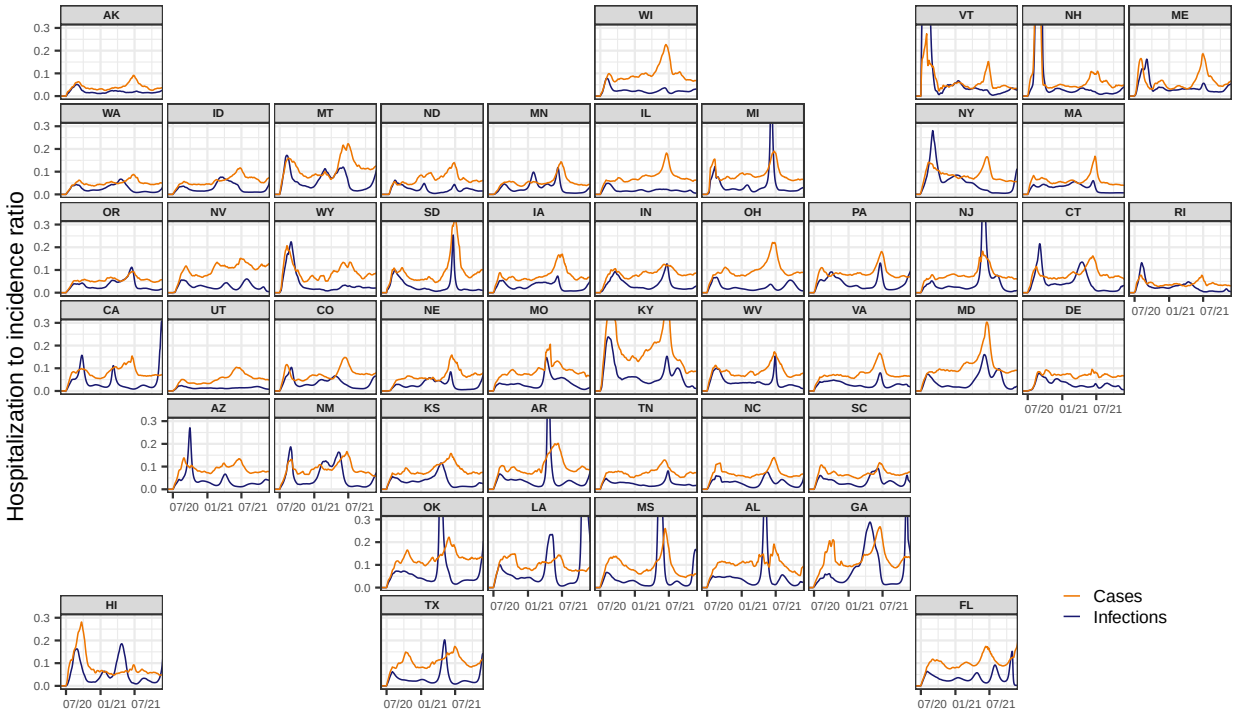

Figure 7: Time-varying IHR and CHR estimates for each state from June 1, 2020 to November 29, 2021, obtained using the respective correlation maximizing lag from Section 3.2. Note that the infection, case, and hospitalization counts are subject to a center-aligned 7-day average to remove spurious day of the week effects. Also note that the different starting points across states are due to the availability of the hospitalization data.

3.2 Insights from cross-correlations, IHRs and CHRs

The maximum Spearman's correlation between infections and hospitalizations is 0.48 and occurs at a lag of 13 days (Figure 6). In contrast, we find that the largest average Spearman correlation for cases is 0.69 and occurs at a lag of 1 day. That is, case reports are nearly contemporaneous to hospitalizations, while infection estimates clearly precede them.

We compute the time-varying infection-hospitalization ratios (IHRs) for each state using a 13-day lag and case-hospitalization ratios (CHRs) with a 1-day lag for comparison Figure 7). Overall, the relationship between infections and hospitalizations is complex. It is characterized by intermittent spikes that punctuate longer periods where the IHRs are relatively stable, remaining below 0.1 hospitalizations per infection.

Both IHRs and CHRs exhibit similar spatiotemporal trends as those noted for infections. Namely, states that are proximate (for example, North and South Carolina) show similar temporal patterns in IHRs and CHRs. In addition, similar spikes are evident across many states during waves of infections that are driven by variants of concern. For example, many states exhibit a striking increase in hospitalizations in mid-2021, which coincides with the rapid takeover of the Delta variant (Hodcroft, 2021).

4 Discussion

We retrospectively estimated daily incident infections for each U.S. state over the period June 1, 2020 to November 29, 2021. Our estimates support the intuition that the pandemic impacted states earlier and at a larger scale than is indicated by reported cases. They also emphasize that using cases as a proxy for infections can lead to erroneous conclusions about trends in infections. More importantly, we observe outbreaks in infections that are missed from inspecting cases alone such as the Delta wave in New Jersey, Connecticut, and Maryland. These sorts of omissions serve to emphasize that cases paint an incomplete picture of the pandemic, especially when outbreaks are largely driven by unreported infections. Furthermore, since case

reports generally follow symptom and infection onsets, cases have a built-in temporal bias. This 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 (Dunkel, 2020; Washington State Department of Health, 2020). Thus, while reported cases provide an indication of the trajectory of the pandemic, it is delayed and incomplete.

Our approach offers a number of advantages. By incorporating state-level case, line list, and variant circulation data, we are able to construct incubation and delay distributions that are spatiotemporally specific. Time-varying and state-specific seroprevalence data allows the reporting ratio estimates to similarly vary over space and time, a departure from existing work (Center for the Ecology of Infection Diseases, 2020; Unwin et al., 2020). Unlike previous approaches that use a single delay distribution to generate estimates for all states (Chitwood et al., 2022; Jahja et al., 2022; Miller et al., 2022), our work avoids this assumption of geographic invariance, an assumption that is far from realistic due to differences in the reporting pipelines, pandemic response, and variants in circulation, among other things. Similarly, prior methodology relies on only one incubation period distribution (Miller et al., 2022), whereas our method incorporates variant-specific incubation periods. This enhances our infection onset estimation by accounting for the differences across variants–specifically, that newer variants tend to have shorter incubation periods (Ogata et al., 2022; Tanaka et al., 2022; Wu et al., 2022).

Another limitation of previous approaches to estimate infections is that they often fail to account for reinfections. While reinfections constitute a small portion of the total infections until the arrival of high immune-escape variants (Omicron BA.1), disregarding them means that the infection-reporting ratio will tend to be underestimated with seroprevalence data alone. By accounting for reinfections as well as the waning of seropositivity, we more accurately estimate this ratio. However, future work could refine this analysis. Because the waning of immunity is likely to be variant-dependent (Pooley et al., 2023), our model's single waning parameter would be more accurately estimated as a mixture of variant-specific parameters with weights determined by the proportion of the variants circulating.

We chose to end our analysis on November 29, 2021, for two main reasons. The first is that Omicron and subsequent variants come with substantial increases in the risk of reinfection in comparison to previous variants, likely due to increased immune escape (Eythorsson et al., 2022; Pulliam et al., 2022; Wei et al., 2024). Access to reinfection data that is representative of each location under study is paramount for extending the analysis. While it would be ideal to use the reinfection rates over time for each U.S. state, many states do not publicly report reinfection data over the entire time period under examination, if at all. The second reason is that the case-ascertainment ratio after December 2021 can no longer be estimated with seroprevalence data alone. Specifically, while most state-level data suggests that reinfections still account for less than 20% of reported cases during Omicron (Hawaii Department of Health, 2022; New York State Department of Health, 2023; Ruff et al., 2022; Washington State Department of Health, 2022), seropositivity rapidly reaches nearly 100% of the population. Therefore, alternative data sources for estimating the case-ascertainment ratio must be considered. For example, wastewater surveillance data may be complementary to seroprevalence data, especially when testing is low, or serve as a substitute when it is unavailable (McManus et al., 2023). An alternative approach could integrate surveillance streams from surveys, helplines, or medical records if they offer a sufficiently strong signal of the disease intensity over time (European Centre for Disease Prevention and Control, 2020; Reinhart et al., 2021).

Our work develops a deconvolution-based approach to inferring infection onset, combining available line list data with variant circulation estimates and literature derived incubation periods. This approach is complemented with the development of a model that incorporates waning detectable antibody levels and major seroprevalence surveys. The resulting infection estimates as well as their geospatial and temporal trends are strongly grounded in both data and statistical models.

These well-informed, localized estimates of COVID-19 infections provide a clear and comprehensive understanding of the pandemic's progression over time. They contribute important information on the timing and magnitude of the disease burden for each location, and highlight trends that may not be visible from reported case data alone. Therefore, these infection estimates provide key information for the ongoing investigation on the true size and impact of the pandemic.

Data availability

The required materials and code for reproducing all figures and the numerical results are available at https://github.com/cmu-delphi/latent-infections/.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Acknowledgements

We would like to thank members of the Delphi research group for valuable feedback, and Change Healthcare and Optum/United Health Group for their invaluable data partnership and collaboration.

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.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and the Centers for Disease Control and Prevention.

DJM and RJT were supported by Centers for Disease Control and Prevention (CDC) Grant No. 75D30123C15907. DJM and RL received support from the National Sciences and Engineering Research Council of Canada and the University of British Columbia. AS was supported by the Centers for Disease Control and Prevention and the National Science Foundation under Award No. 2223933 and 2333494.

References

Annavajhala, M. K., Mohri, H., Wang, P., Nair, M., Zucker, J. E., Sheng, Z., Gomez-Simmonds, A., Kelley, A. L., Tagliavia, M., Huang, Y., et al., 2021. Emergence and expansion of SARS-CoV-2 B. 1.526 after identification in New York. Nature 597 (7878), 703–708, doi: 10.1038/s41586-021-03908-2.

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

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

Centers for Disease Control and Prevention, 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.

Centers for Disease Control and Prevention, 2021b. Nationwide commercial laboratory seroprevalence survey. https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv.

Centers for Disease Control and Prevention, 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 (8), e1010465, doi: 10.1371/journal.pcbi.1010465.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases 20 (5), 533–534, doi: 10.1016/S1473-3099(20)30120-1.

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

Elbe, S., Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges 1 (1), 33–46, doi: 10.1002/gch2.1018.

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., Palsson, R., 2022. Rate of SARS-CoV-2 reinfection during an Omicron wave in Iceland. JAMA Network Open 5 (8), e2225320–e2225320, doi: 10.1001/jamanetworkopen.2022.25320.

Figgins, M. D., Bedford, T., 2021. SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. MedRxiv, doi: 10.1101/2021.12.09.21267544.

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

Hitchings, M. D., Dean, N. E., García-Carreras, B., Hladish, T. J., Huang, A. T., Yang, B., 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 (7), 1396–1405, doi: 10.1093/aje/kwab023.

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

Jahja, M., Chin, A., Tibshirani, R. J., 2022. Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. Statistical Science 37 (2), 207–228, doi: 10.1214/22-STS856.

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 (8), 1589, doi: 10.3201/eid2908.221634.

Miller, A. C., Hannah, L. A., Futoma, J., Foti, N. J., Fox, E. B., D'Amour, A., Sandler, M., Saurous, R. A., Lewnard, J. A., 2022. Statistical deconvolution for inference of infection time series. Epidemiology 33 (4), 470–479, doi: 10.1097/EDE.0000000000001495.

New York State Department of Health, 2023. COVID-19 reinfection data. https://coronavirus.health.ny.gov/covid-19-reinfection-data.

Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., Bosso, M., Park, D. J., Babadi, M., MacInnis, B. L., et al., 2022. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science 376 (6599), 1327–1332, doi: 10.1126/science.abm1208.

Ogata, T., Tanaka, H., Irie, F., Hirayama, A., 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 (3), 1127, doi: 10.3390/ijerph19031127.

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 (1829), 20200264, doi: 10.1098/rstb.2020.0264.

Pitzer, V. E., Chitwood, M., Havumaki, J., Menzies, N. A., Perniciaro, S., Warren, J. L., Weinberger, D. M., 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 (9), 1908–1917, doi: 10.1093/aje/kwab089.

Pooley, N., Abdool Karim, S. S., Combadière, B., Ooi, E. E., Harris, R. C., El Guerche Seblain, C., Kisomi, M., Shaikh, N., 2023. Durability of vaccine-induced and natural immunity against COVID-19: A narrative review. Infectious Diseases and Therapy 12 (2), 367–387, doi: 10.1007/s40121-022-00753-2.

Pulliam, J. R., van Schalkwyk, C., Govender, N., von Gottberg, A., Cohen, C., Groome, M. J., Dushoff, J., Mlisana, K., Moultrie, H., 2022. Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. Science 376 (6593), eabn4947, doi: 10.1126/science.abn4947.

Ramdas, A., Tibshirani, R. J., 2016. Fast and flexible ADMM algorithms for trend filtering. Journal of Computational and Graphical Statistics 25 (3), 839–858, doi: 10.1080/10618600.2015.1054033.

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 (51), e2111452118, doi: 10.1073/pnas.2111452118.

Ruff, J., Zhang, Y., Kappel, M., Rathi, S., Watkins, K., Zhang, L., Lockett, C., 2022. Rapid increase in suspected SARS-CoV-2 reinfections, Clark County, Nevada, USA, December 2021. Emerging Infectious Diseases 28 (10), 1977, doi: 10.3201/eid2810.221045.

Tanaka, H., Ogata, T., Shibata, T., Nagai, H., Takahashi, Y., Kinoshita, M., Matsubayashi, K., Hattori, S., 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 (10), 6330, doi: 10.3390/ijerph19106330.

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 (1), 285–323, doi: 10.1214/13-AOS1189.

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

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 (1), 6189, doi: 10.1038/s41467-020-19652-6.

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

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

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 (1), 1008, doi: 10.1038/s41467-024-44973-1.

Wu, Y., Kang, L., Guo, Z., Liu, J., Liu, M., Liang, W., 2022. Incubation period of COVID-19 caused by unique SARS-CoV-2 strains: a systematic review and meta-analysis. JAMA network open 5 (8), e2228008–e2228008, doi: 10.1001/jamanetworkopen.2022.28008.