Challenges in Real-Time Estimation of Changing Epidemic Severity Rates

Jeremy Goldwasser

October 2024

Abstract

Severity rates like Case-Fatality Rate and Infection-Fatality Rate are ubiquitous metrics in public health. To guide decision-making in response to changes like new variants or vaccines, it is imperative to understand how these rates shift in real time. We demonstrate that standard ratio estimators for time-varying severity rates may exhibit high statistical bias. These ratios may fail to detect increases in fatality risk, or falsely signal nonexistent surges. We justify our theoretical analyses with experimental results on real and simulated data from COVID-19. Finally, we highlight strategies to mitigate this bias, drawing connections with R_t estimation.[1]

1 Introduction

A number of public health metrics express the probability that a second, more serious outcome will follow a primary event. For example, the Case-Fatality Rate (CFR) is commonly used as a proxy for the underlying Infection-Fatality Rate (IFR) to assess the intensity of an epidemic. Other examples of such "severity rates" include the Hospitalization-Fatality Rate and Case-Hospitalization Rate.

In an ideal setting, severity rates can be obtained directly from line-list data of individual patient outcomes Bellan et al. (2020); Challen et al. (2021); Roth et al. (2021); Xie et al. (2024). In fast-moving epidemics like COVID-19, however, large-scale tracking is infeasible, especially in real-time Overton et al. (2022). Instead, these rates are estimated from aggregate count data. While many works assume they are constant over time Baud et al. (2020); Ghani et al. (2005); Jewell et al. (2007); Reich et al. (2012), in reality they are constantly changing in response to factors such as new therapeutics, vaccines, and variants McNeil (2020). Time-varying severity rates are typically estimated with a ratio of the two aggregate data streams such as cases and deaths. These ratios have been widely used to report COVID CFRs, both in academic literature Horita and Fukumoto (2022); Liu et al. (2023); Luo et al. (2021); Wjst and Wendtner

[1] Code is available at `https://github.com/jeremy-goldwasser/Severity-Bias`.

(2023); Yuan et al. (2020) and major news publications like the Atlantic Madrigal and Moser (2020) and Wall Street Journal Kamp and Krouse (2020). While other methods exist, ratio estimators are so common that IFR, for example, is often referred to as the Infection-Fatality *Ratio* COVID-19 Forecasting Team (2022); Luo et al. (2021).

In this work, we demonstrate these ratio estimators exhibit fundamental statistical bias, and identify three factors that drive it. Bias arises as a consequence of changing severity rates, precisely when time-varying estimates should be most useful. This bias may be influenced by changes in primary incidence levels, as well as long time delays between events. During COVID-19, the ratio estimators would have failed to quickly identify the rise in hospitalization-fatality rate (HFR) during the onset of the Delta wave. After the initial Omicron surge, the ratios spiked as the true HFRs fell. We study the sources of this bias, and suggest alternative methodology which overcomes it.

2 Methods

2.1 Severity Rate Estimators

The time-varying severity rate is defined as

$$p_t = \mathbb{P}(\text{secondary event will occur} \mid \text{primary event at time } t). \tag{1}$$

Let $\{X_t\}$, $\{Y_t\}$ denote the time series of interest. In the case of CFR, for example, X_t and Y_t are the total number of new cases and deaths, respectively, at day t.

The canonical estimator for time-varying severity rates is a ratio between X_t and Y_t events, offset by a lag ℓ. This lagged approach is formally introduced in Thomas and Marks, but had been used in numerous prior works, e.g. Horita and Fukumoto (2022); Kamp and Krouse (2020); Liu et al. (2023); Luo et al. (2021); Madrigal and Moser (2020); Wjst and Wendtner (2023); Yuan et al. (2020). The real-time estimator only uses data until the present timestep t:

$$\hat{p}_t^{\text{Lagged}} = \frac{Y_t}{X_{t-L}} \tag{2}$$

Alternative methods use the delay distribution that relates the two time series. Let $\pi_k^{(t)}$ denote the probability that the secondary event occurs k days after the primary event, given it occurs at all. A number of tools exist to estimate delay distributions from aggregate or line-list data Charniga et al. (2024). It is necessary to truncate the delay distribution at d days, in essence assuming all secondary events occur within this period.

The expected number of secondary events at any given day can be expressed by convolving the delay distribution against hospitalizations and severity rates Nishiura et al. (2009); Qu et al. (2022).[2]

[2]Throughout this work, we assume primary incidence is known, and condition on $X_{s\leq t}$ implicitly. We also assume the delay distribution π is the same over all time.

$$
\begin{aligned}
E[Y_t] &= \sum_{k=0}^{\infty} X_{t-k} \mathbb{P}(\text{secondary at } t \mid \text{primary at } t-k) \\
&= \sum_{k=0}^{\infty} X_{t-k} \mathbb{P}(\text{secondary after } k \mid \text{secondary occurs, primary at } t-k) \\
&\qquad \times \mathbb{P}(\text{secondary occurs} \mid \text{primary at } t-k) \\
&= \sum_{k=0}^{\infty} X_{t-k} \pi_k p_{t-k}. \qquad (3)
\end{aligned}
$$

If the severity rates are a constant p, Eq. 3 simplifies to $E[Y_t] = p \sum_{k=0}^{\infty} X_{t-k} \pi_k$. Nishiura et al. rearranged this expression to estimate the stationary rate, using a plug-in estimate of the delay distribution and smoothing with cumulative counts.

$$
\hat{p}_t = \frac{\sum_{s=t_0}^{t} Y_s}{\sum_{s=t_0}^{t} \sum_{k=0}^{d} X_{s-k} \hat{\pi}_k}. \qquad (4)
$$

This estimator is widely used in practice Garske et al. (2009); Russell et al. (2020a,b). Assuming the true rate is indeed stationary and the delay distribution is correctly specified, it is unbiased. Overton et al. adapted Eq. 4 for the time-varying setting, using daily rather than cumulative counts:

$$
\hat{p}_t^{\text{Conv}} = \frac{Y_t}{\sum_{k=0}^{d} X_{t-k} \hat{\pi}_k}. \qquad (5)
$$

This convolutional ratio can be understood as a generalization of Equation 2. It reduces to the same ratio if $\hat{\pi}$ is a point mass distribution where all secondary events occur after ℓ days. They also are equivalent if primary events are constant. Otherwise, it expresses a more accurate relation the two time series. In practice, however, we have not come across work that applies it for time-varying estimation. Rather, the lagged ratio is standard. [AH: I'm just a little bit confused here... Overton introduces (5) for the time-varying setting but but there is no work that applies it for "time-varying estimation"? What do Overton do with it then? Just a little bit more context here would be helpful for me...]

To stabilize estimates, smoothed counts are often used in practice Liu et al. (2023); Luo et al. (2021); Wjst and Wendtner (2023). For the sake of simplicity of presentation, we generally focus on the versions described above. However, we formalize the smoothed versions in Equations 11 and 12, and analyze them experimentally.

2.2 Well-Specified Analysis

In this section, we demonstrate that these time-varying severity ratios are biased when the true rates are changing, even when the delay distribution is correctly

specified. Assume the true delay distribution is a constant π over all time with maximum length d. [AH: Can we remove this assumption that π has maximum length d? Replacing d with ∞ in the subsequent analysis does not change its validity, and would make the presentation cleaner (the previous exposition does not assume bounded support for π, and you also don't need to assume later on that d has larger support than the estimator $\hat{\pi}$ for π has smaller support than π in the misspecified analysis.] We first analyze the convolutional ratio (Eq. 5), assuming oracle knowledge of the true delay distribution π.

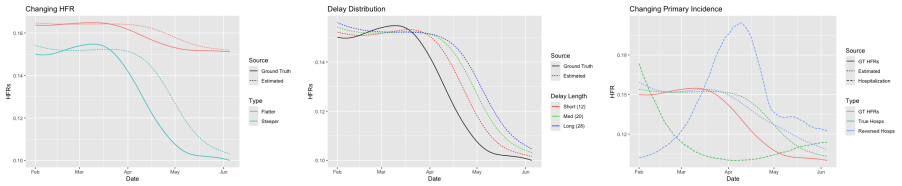
$$\begin{aligned}
\text{Bias}(\hat{p}_t^{\text{OracleConv}}) = E[\hat{p}_t^{\text{OracleConv}}] - p_t &= \frac{E[Y_t]}{\sum_{k=0}^{d} X_{t-k}\pi_k} - p_t \\
&= \frac{\sum_{k=0}^{d} X_{t-k}\pi_k p_{t-k}}{\sum_{k=0}^{d} X_{t-k}\pi_k} - \frac{p_t \sum_{k=0}^{d} X_{t-k}\pi_k}{\sum_{k=0}^{d} X_{t-k}\pi_k} \\
&= \sum_{k=0}^{d} \frac{X_{t-k}\pi_k}{\sum_{j=0}^{d} X_{t-j}\pi_j}(p_{t-k} - p_t). \qquad (6)
\end{aligned}$$

The degree of bias in Eq. 6 depends on three factors. [AH: I think this decomposition of the sources of bias is fantastic... it really makes it clear that change in severity rate is the underlying source of bias and its sign (pointwise over k), and the delay distribution / primary incidence curve work to take a weighted average over those pointwise $p_{t-k} - p_t$ bias terms.
]

1. **Changes in severity rate**. The central component of this bias expression is the $p_{t-k} - p_t$ term. When severity rates are constant in the d preceding days, this estimator is unbiased. This is in line with the unbiasedness of estimator using cumulative counts assuming a globally stationary rate Nishiura et al. (2009). But when severity rates change before t, the numerator will likely not equal 0, in which case the estimator will be biased. Figures 1a and 4a illustrate this: The estimated severity rates are most inaccurate at periods where the true rate is changing sharply.

   The bias is in the opposite direction of the trend we want to detect. For example, if the severity rate is falling, then $p_{t-k} > p_t$ for many $k \in \{1, \ldots, d\}$. As a result, the bias is positive, meaning the ratio estimates do not decline at the true rate. In fact, the estimated severity may even rise, not fall. Conversely, when true severity rates are rising, the ratio estimates will be too low.

2. **The delay distribution**. How much the changing severity rates impact the bias depends on the shape of the delay distribution. In general, the bias is greatest when the delay distribution is long-tailed enough to upweight significant differences in severity rate. While this distinction may appear subtle, the Results section highlights its surprisingly large effects. The simple example in Figures 1b and 4a shows significant differences in bias between shorter and longer delay distributions.



(a) (b) (c)

Figure 1: Simple examples of severity rate bias. NCHS HFRs and HHS hospitalizations from early 2022. Using fitted delay distribution (2.4).

3. **The primary incidence curve.** Changing primary incidences will also affect the bias, presuming the severity rate changes roughly monotonically in the recent past. Intuitively, this up- or down-weights the terms $X_{t-k}\pi_k(p_{t-k} - p_t)$ for dates further from the present, which are likely to contribute the most bias. Falling primary incidences will amplify the bias, whereas rising events will minimize it. Figures 1c and 4b visualize this trend on the convolutional ratio.

As noted previously, the convolutional ratio is equivalent to the lagged ratio if $\hat{\pi}$ is a point mass distribution at ℓ. [AH: In the above, do you mean to use π rather that $\hat{\pi}$? And in the below, should it say "OracleLagged" and "OracleConv"?] In this oracle setting, this indicates all secondary events occur after exactly ℓ days, a highly unrealistic situation. Nevertheless, if this is the case, then

$$\text{Bias}(\hat{p}_t^{\text{Lagged}}) = \text{Bias}(\hat{p}_t^{\text{Conv}}) = p_{t-\ell} - p_t.$$

This toy setting and others are discussed in Appendix A.

2.3 Misspecified Analysis

The above section considered the bias of the convolutional ratio where the true delay distribution is known. We now consider the more general case, in which it is instead replaced with a plug-in estimate. Nevertheless, we still assume this distribution is constant over time, and its length d exceeds that of the plug-in distribution. Note the bias of the lagged estimator is a special case, where the

plug-in distribution has all its mass at the lag time ℓ.

$$\begin{aligned}
\text{Bias}(\hat{p}_t) &= \frac{E[Y_t]}{\sum_{k=0}^{d} X_{t-k}\hat{\pi}_k} - p_t \\
&= \frac{\sum_{k=0}^{d} X_{t-k}\pi_k p_{t-k}}{\sum_{k=0}^{d} X_{t-k}\hat{\pi}_k} - \frac{\sum_{k=0}^{d} X_{t-k}\hat{\pi}_k p_t}{\sum_{k=0}^{d} X_{t-k}\hat{\pi}_k} \\
&= \sum_{k=0}^{d} \frac{X_{t-k}}{\sum_{j=0}^{d} X_{t-j}\hat{\pi}_j}(\pi_k p_{t-k} - \hat{\pi}_k p_t) \\
&= \sum_{k=0}^{d} \frac{X_{t-k}}{\sum_{j=0}^{d} X_{t-j}\hat{\pi}_k}(\pi_k p_{t-k} - (\pi_k + \hat{\pi}_k - \pi_k)p_t) \\
&= \sum_{k=0}^{d} \frac{X_{t-k}\pi_k}{\sum_{j=0}^{d} X_{t-j}\hat{\pi}_j}(p_{t-k} - p_t) + \\
&\qquad p_t \sum_{k=0}^{d} \frac{X_{t-k}}{\sum_{j=0}^{d} X_{t-j}\hat{\pi}_j}(\hat{\pi}_k - \pi_k) \qquad (7)
\end{aligned}$$

[AH: For the above, I would proceed from the third line, and write the bias as:

$$\text{Bias}(\hat{p}_t) = \sum_{k=0}^{d} \frac{X_{t-k}\hat{\pi}_k}{\sum_{j=0}^{d} X_{t-j}\hat{\pi}_j}\left((\pi_k/\hat{\pi}_k)p_{t-k} - p_t\right)$$

Now you can discuss the bias of the non-oracle case in terms of "multiplicative" terms, in line with your discussion of the oracle case bias breakdown of change in severity rate, delay distribution, and primary incidence curve:

1. **Changes in severity rate.** The $p_{t-k} - p_t$ term is replaced by a $(\pi_k/\hat{\pi}_k)p_{t-k} - p_t$ term. Interestingly, this means that when if $p_{t_k} < p_t$ and $\hat{\pi}_k < \pi_k$, then the kth pointwise term here can have *smaller* bias than in the oracle case (similar for $>$). But since $\hat{\pi}_k$ has the constraint that it is a distribution (sums to 1), it's not clear that you can "win" overall.

2. **The delay distribution.** Same discussion here, except it depends on the delay distribution estimate rather than the underlying delay distribution. Considered by itself, this would suggest that we should favor $\hat{\pi}_k$ with shorter tails, to avoid upweighting significant differences in severity rate. However, this needs to be balanced against the $\pi_k/\hat{\pi}_k$ term above, and taking a delay distribution estimate too different from the underlying delay distribution can hurt us overall.

3. **The primary incidence curve.** The discussion here is exactly the same as in the well-specified case.

Taking all of this together, we can summarize the bias comparison between the oracle case and non-oracle case into: (a) the "pointwise severity rate bias" terms

are affected by the delay distribution estimate ratio $\pi_k/\hat{\pi}_k$, and (b) the weight terms taking a weighted average of the "pointwise severity rate bias" terms to the overall bias of $\hat{p}_t$, where π is entirely replaced by $\hat{\pi}$.

I know this isn't an additive discussion into the oracle bias and a misspecification term... I still would love to know if it is a fact that $\hat{\pi} = \pi$ minimizes the bias.]

For the lagged estimator,

$$\begin{aligned}\text{Bias}(\hat{p}_t^{\text{Lagged}}) &= \sum_{k=0}^{d} \frac{X_{t-k}}{X_{t-\ell}}(\pi_k p_{t-k} - \{k = \ell\}p_t) \\ &= \sum_{k=0}^{d} \pi_k \Big(\frac{X_{t-k}}{X_{t-\ell}} p_{t-k} - p_t\Big). \end{aligned} \tag{8}$$

The only difference between the first term in the final equality of Eq. 7 and Eq 6 is the true versus plug-in delay distribution in the denominator of the weighting term.

It would be ideal if we could say something about the misspecified bias in terms of the *actual* oracle bias plus a misspecification term. This is clearly not possible from the previous equation, since the $\hat{\pi}_j$ in the denominator cannot be factored out by any means.

It is possible to express this as a bound, after making a strange assumption about the delay distribution. However, what the bounds tell us is not interesting. Depending on the direction of the bias & the assumption, it either provides

1. An unsatisfying lower bound - unsatisfying in that the lower bound may be considerably *better* than the oracle estimator's bias.

2. A worst-case bound. This isn't particularly interesting, since we're more interested in showing that the bias with misspecified delay distribution is demonstrably bad.

I'm happy to write up these bounds, but I don't think they're interesting enough to make their way into the main paper.

Lastly, I looked into the mean squared error of this misspecified estimator. Can we break the MSE into oracle and plug-in error terms? Unfortunately, the answer again appears to be no. The expression also has the plug-in distribution in a denominator that cannot be nicely analyzed.

In light of all of this, I think it is most sensible to just write about the bias expressions as presented. There still are relatively interesting things to say:

1. It decomposes into *roughly* an oracle and misspecification term, though the oracle term is deeply reliant on the plug-in delay distribution;

2. The lagged bias has a ratio of primary incidence, comparing k days ago to the lag time ℓ. This reliance yields interesting results. e.g. In Figure 4b, as hospitalizations start to decline, lagged bias falls while oracle convolutional bias rises.

Per Ryan's comment on line 379, it is important to do this well. The lagged ratio generally does much worse than the convolutional estimator, indicating substantial error as a consequence of delay misspecification. Equation 7 describes this drop in performance, but it is unfortunately challenging to analyze.

For the sake of legibility, I've hidden Addison and Ryan's Overleaf commentary encouraging this investigation into misspecification error. It's entirely possible there's stuff I've missed; feel free to uncomment parts out if need be.

2.4 Experimental Setup

2.4.1 HFR Estimation

Our experiments focus on the Hospitalization-Fatality Rate (HFR) during COVID-19. Hospitalization reporting was much more complete than case reporting throughout the pandemic. Hospitals were mandated to report new daily admissions to the Department of Health and Human Services (HHS) or face penalties Department of Health and Human Services (2023). The time-to-death delay distribution is indeed supported on integers starting at $k = 0$, since hospitalizations are aligned by admission date.

To estimate real-time HFRs, we pulled daily hospitalizations and deaths from the `epidatr` API, developed by the Delphi Group. Like HHS for hospitalizations Department of Health and Human Services (2023), John Hopkins University (JHU) provided the definitive resource for real-time death counts Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (2023). These counts reflect the times at which deaths were reported to health authorities, not necessarily when they actually happened. Therefore raw JHU death counts are highly volatile due to reporting idiosyncrasies like day-of-week effects and data dumps. As a result, we used a 7-day trailing average of counts.

The daily aggregates from JHU and HHS were updated over the course of the pandemic. We pulled both the counts available in real time as well as their finalized versions. Often, the most recent date with available counts lagged several days behind the present. To account for this, we estimated HFRs each week, pulling real-time data 6 days after each date. In the rare chance that requested counts still were unavailable, we imputed with the most recently observed date.

The two ratio estimators (Eq. 2 and 5) require choices of lag and delay distribution. Appendix C evaluates the robustness of findings against different hyperparameter values. The experiments in Section 3.1 use a lag of 20 days, which maximizes the cross-correlation between hospitalizations and deaths over all time. We let the delay distribution be a discrete gamma, a common choice. We set its mean to this oracle lag, as lags are often chosen to be the mean of the delay distribution. This mean of 20 matches nicely with a UK study (CITE) that finds a median hospitalization-to-death time of 11 days, and a CDC report that 63% of COVID deaths are reported within 10 days. We set the standard

deviation to 18, because the delay distributions fit by the UK study had standard deviations that were roughly 90% of their means.

2.4.2 Validation Data

While the true HFRs are unknown, there are sound ways to approximate them. One such approach is to use line-list HFRs from the National Hospital Care Survey National Center for Health Statistics (NCHS) (2023). The NHCS recorded weekly HFRs from inpatient deaths in a representative subset of 601 hospitals across the US.

HFRs from aggregate hospitalization and death counts are significantly higher than those from NHCS because not all deaths occur in hospitals. A CDC analysis reported the percentage of inpatient deaths every month from 2020 through 2022; roughly 60% of COVID deaths occurred in hospitals in 2022, down from nearly 70% in 2021 and 2022. To account for non-inpatient deaths, we divided the NHCS curve by these percentages. Finally, we smoothed the resulting HFRs with a spline. To do so, we used the `smooth.spline` function in `R`, which chooses the smoothness hyperparameter with generalized cross validation.

We considered two other sources for ground truth HFRs, discussed in Appendix B.2. Unlike NHCS, these HFRs are obtained from aggregate counts, not line-list data. Fortunately, they are fairly consistent with the rescaled NHCS data, bolstering our trust in it. Of course, the NHCS curve is merely an approximation for the ground truth; its values, especially after mid-2022, may be incorrect. Nevertheless, it is a useful aide with which to judge the fidelity of our HFR estimates.

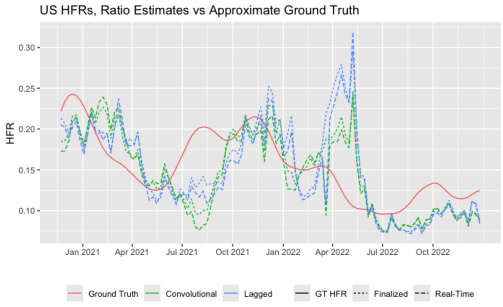
3 Results

3.1 National COVID Data

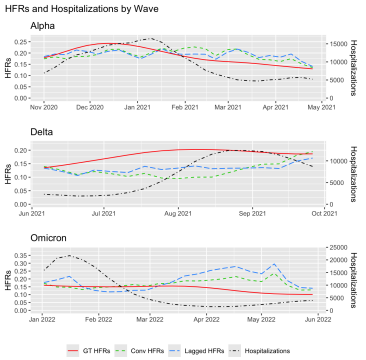
Figure 2 highlights the bias of these ratio estimators. Both the lagged and convolutional ratios respond very slowly to changes in the HFR. As the HFR declines throughout the Alpha wave in early 2021, both ratios stay around 0.2 for several months. More troublingly, they are very slow to detect the rising HFR in the early Delta period (summer 2021).

The most significant bias comes in the middle of the Omicron wave in spring 2022. The true HFRs sharply decline in this period, from a high of roughly 17% in March to a low of 9% only two months later. At the same time, the HFR estimates *rise*, peaking over 20% as the true HFR reaches its nadir. This dramatic surge signals a serious false alarm.

The analysis in Section 2.2 explains for these three failure cases. Recall the bias moves in the opposite direction of the true severity rate. Rising HFRs engender negative bias, hence the delayed rise in the Delta wave. Falling HFRs correspond to positive bias, as observed in early 2021 and 2022. In addition, the enormity of the bias during Omicron can be attributed to the precipitous



(a) Comparing convolutional and lagged ratios against approximate ground truth. Finalized and real-time counts, Nov. 2020 - Dec 2022.



(b) HFRs and hospitalizations in three periods with major bias. Finalized counts.

Figure 2: Convolutional ratio estimates are biased regardless of which delay distribution is selected.

decline in hospitalizations, as falling primary incidence has been shown to exacerbate the bias. Average daily hospitalizations declined from over 20,000 in mid-January to only 1,500 by April 1. Finally, the delay distribution is relatively long with JHU deaths due to its alignment by report date. This is shown to have a substantial impact on the bias, as analyzed in Appendix B.1.

We performed several robustness checks to assess the stability of these findings. Figure 2a compares the convolutional and lagged ratio estimators, finding bias in both. The convolutional estimator is slightly better, but still very problematic. Appendix C explores the effect of different hyperparameters and locations. By and large, the ratio estimators are biased regardless of these considerations.

3.2 Simulated Data

We further evaluated these methods on simulated deaths whose true HFRs is known. Throughout these experiments, we used observed, finalized HHS hospitalization reports X_t, and various models for the HFR and delay distribution. Given a series of time-varying HFRs p_t and delay distribution π, deaths are defined without noise as according to 3

$$Y_t := \sum_{k=0}^{d} X_{t-k} \mathbb{P}(\text{die at } t \mid \text{hosp at } t-k) = \sum_{k=0}^{d} X_{t-k} \pi_k p_{t-k}.$$

To mimic the real data, we first used the same HFRs from NHCS used for validation in 3.1. We also inverted them and rescaled in order to simulate the opposite trend. Lastly, we explored a stationary HFR of 10% over all time. The delay distributions were again gamma with standard devation 0.9 of their

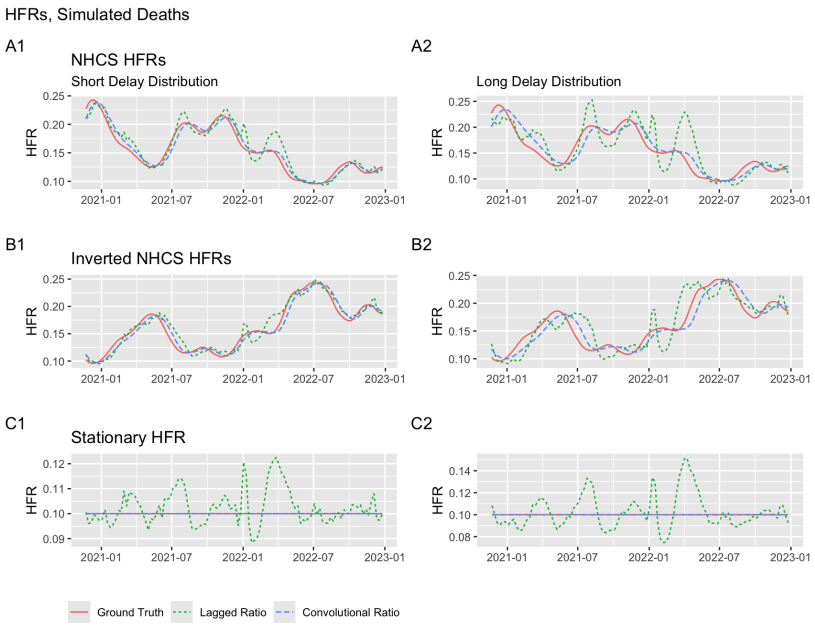


Figure 3: True and Estimated HFRs from Simulated Deaths. First column has short delay distribution, second has long.

mean. We experimented with means of 12 and 24 to illustrate a short and long delay distribution. The ratio estimators used the oracle hyparameters: The true delay distribution for the convolutional ratio, and its mean for the lagged method. Estimates were not smoothed over a trailing window.

Figure 3 displays the results on the 2 delay distributions and 3 HFR settings. The bias was significantly more pronounced with the longer delay distribution. To assess the relationship between the estimated and true HFRs, we identified the offset that maximized the cross-correlation between the two series. On both the true and inverted NHCS HFRs, this was 7 days for the short distribution, a relatively innocuous gap. However, the offset is 21 days with the longer distribution - concerningly slow during a rapidly changing severity rate like the Delta surge.

Interestingly, the lagged estimator performed considerably worse than the convolutional ratio. When the true HFR changed, the lagged estimates oscillated while the convolutional ratio followed the general trend. In the stationary HFR case, its bias reached as high as 50%; in contrast, the convolutional estimator was unbiased as anticipated (Eq. 6). This discrepancy is striking, as the lagged ratio is the most commonly-used time-varying estimator.

4 Discussion

Our analyses illustrate that practitioners should take caution when using these time-varying severity ratios. They exhibit considerable bias when severity rates change, particularly the popular lagged ratio. A major purpose of these estimators is to inform stakeholders of changing risks in real time; this bias indicates they may fail to do so in a reliable manner.

Given the drawbacks of these methods, alternative approaches may be preferable when there is reason to believe the true rate is changing. If possible, severity rates can be obtained from line-list data after accounting for right censoring. These rates can then be scaled to a broader population with careful demographic adjustment Verity et al. (2020).

More commonly, only aggregate data is available, especially in real time. In that case, other methods may outperform these ratios. Qu et al. propose estimating all severity rates at once with a Fused Lasso model, using the relation in Eq. 3. Unlike the other approaches, this method is inherently forward-looking, where rates at t are exclusively used to produce secondary events after t. However, it may suffer from other sources of bias. It is inclined to estimate smoothly-changing severity rates as piecewise constant, and may yield unstable real-time estimates due to scarce data at the tail.

Overton et al. also proposed a forward-looking method, this one a ratio between relevant primary and secondary events. However, this method is not applicable in real time, as it uses secondary events after t to compute the severity rate. Nevertheless, it is a useful tool for retrospective estimation.

Another retrospective tool is aggregate COVID deaths from NCHS, a resource that was not available in real time (Appendix B.1). Unlike JHU, whose aggregates align deaths by report date, NCHS counts deaths on the day the actually occurred. As a result, the mean of its delay distribution is considerably lower, so it produces more accurate ratio estimates (Fig. 5, 3). Analogously, bias is a more serious issue with earlier primary events. For example, case- or infection-fatality ratios may be more biased than hospitalization-fatality ratios.

While still biased, the convolutional ratio generally outperformed the lagged method (Fig. 2a, 3). While this estimator is widely used for overall HFRs with cumulative counts, we have not come across any applications for the time-varying case. This further suggests the lagged ratio is overused in practice, though the convolutional ratio may not be the best existing alternative Overton et al. (2022); Qu et al. (2022).

Severity rates may be biased in ways beyond the statistical bias our work focuses on. Section 2.4 mentioned, for example, the fact that estimating HFR from aggregates fails to address the large proportion of deaths occur outside the hospital; Lipsitch et al. refer to this as "survivorship bias." A central challenge for CFR estimation is under-reporting: Not all events are reported, reporting rates change across time, and severe cases are more likely to be reported than mild cases. Reich et al. propose an estimator for a time-invariant *relative* CFR - the ratio of CFRs between groups - that learns these latent reporting rates via the EM algorithm Dempster et al. (1977). Jordan applied this in the context

of COVID-19, analyzing how the chosen delay distribution affects its results. They also identify other sources of bias, like differences in case definition and testing eligibility.

An interesting connection exists between estimating severity rates and reproduction numbers. A central metric in epidemiology is *case* R_t, the average number of secondary infections produced by a single infection at time t. Typically estimated in real-time is the closely-related *instantaneous* R_t, average number of secondary infections at time t produced by a single primary infection in the past. Comparable to the delay distribution π is the renewal equation g, which measures the time between primary and secondary infections.

As defined in 1, the severity rate is analogous to case R_t. Both concern the average number of secondary events produced by a primary event at time t. Moreover, the real-time severity ratios we study are analogous to instantaneous R_t, both of which measure how primary events in the past contribute to secondary events at t. Indeed, one of the most popular frameworks for estimating instantaneous R_t is strikingly similar to the convolutional ratio Cori et al. (2013); Fraser (2007); Wallinga and Lipsitch (2007):

$$\hat{R}_t = \frac{I_t}{\sum_{k=0}^{d} I_{t-k} g_k}. \tag{9}$$

Fraser notes that instantaneous R_t is equal to case R_t if conditions remain unchanged. Similarly, we demonstrated that the convolutional ratio 5 is unbiased if the severity rate and delay distribution in the d days before t are stationary. Bias arises as a consequence of changing conditions. Future work could apply this same analytical framework to R_t, examining the fidelity of instantaneous R_t as a proxy for case R_t.

References

Adjei, Stacey and Hong, Kai and Molinari, Noelle-Angelique M and Bull-Otterson, Lara and Ajani, Umed A and Gundlapalli, Adi V and Harris, Aaron M and Hsu, Joy and Kadri, Sameer S and Starnes, Jon and Yeoman, Kristin and Boehmer, Tegan K (2022). Mortality risk among patients hospitalized primarily for covid-19 during the omicron and delta variant pandemic periods - united states, april 2020-june 2022. *MMWR Morb Mortal Wkly Rep*, 71(37):1182–1189.

Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., and Favre, G. (2020). Real estimates of mortality following COVID-19 infection. *Lancet Infect Dis*, 20(7):773. Epub 2020 Mar 12.

Bellan, M., Patti, G., Hayden, E., et al. (2020). Fatality rate and predictors of mortality in an italian cohort of hospitalized covid-19 patients. *Sci Rep*, 10:20731.

Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (2023). Covid-19 data repository. GitHub repository.

Challen, R., Brooks-Pollock, E., Read, J. M., Dyson, L., Tsaneva-Atanasova, K., and Danon, L. (2021). Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ*, 372:n579.

Charniga, K., Park, S. W., Akhmetzhanov, A. R., Cori, A., Dushoff, J., Funk, S., Gostic, K. M., Linton, N. M., Lison, A., Overton, C. E., Pulliam, J. R. C., Ward, T., Cauchemez, S., and Abbott, S. (2024). Best practices for estimating and reporting epidemiological delay distributions of infectious diseases using public health surveillance and healthcare data.

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512.

COVID-19 Forecasting Team (2022). Variation in the COVID-19 infection–fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. *The Lancet*, 399(10334):1469–1488.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Department of Health and Human Services (2023). Covid-19 guidance for hospital reporting and faqs for hospitals, hospital laboratory, and acute care facility data reporting.

Fraser, C. (2007). Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. *PLOS ONE*, 2(8):1–12.

Garske, T., Legrand, J., Donnelly, C. A., Ward, H., Cauchemez, S., Fraser, C., Ferguson, N. M., and Ghani, A. C. (2009). Assessing the severity of the novel influenza a/h1n1 pandemic. *BMJ*, 339.

Ghani, A. C., Donnelly, C. A., Cox, D. R., Griffin, J. T., Fraser, C., Lam, T. H., Ho, L. M., Chan, W. S., Anderson, R. M., Hedley, A. J., and Leung, G. M. (2005). Methods for Estimating the Case Fatality Ratio for a Novel, Emerging Infectious Disease. *American Journal of Epidemiology*, 162(5):479–486.

Horita, N. and Fukumoto, T. (2022). Global case fatality rate from COVID-19 has decreased by 96.8% during 2.5 years of the pandemic. *Journal of Medical Virology*.

Jewell, N. P., Lei, X., Ghani, A. C., Donnelly, C. A., Leung, G. M., Ho, L.-M., Cowling, B. J., and Hedley, A. J. (2007). Non-parametric estimation of the case fatality ratio with competing risks data: an application to Severe Acute Respiratory Syndrome (SARS). *Stat Med*, 26(9):1982–1998.

Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the covid-19 case fatality rate. *Harvard Data Science Review*.

Kamp, J. and Krouse, S. (2020). Case-Fatality Metric Points to Increase in December Deaths. *Wall Street Journal*.

Lipsitch, M., Donnelly, C. A., Fraser, C., Blake, I. M., Cori, A., Dorigatti, I., Ferguson, N. M., Garske, T., Mills, H. L., Riley, S., Van Kerkhove, M. D., and Hernán, M. A. (2015). Potential biases in estimating absolute and relative case-fatality risks during outbreaks. *PLoS Neglected Tropical Diseases*, 9(7):e0003846.

Liu, J., Wei, H., and He, D. (2023). Differences in case-fatality-rate of emerging sars-cov-2 variants. *Public Health in Practice*, 5:100350.

Luo, G., Zhang, X., Zheng, H., and He, D. (2021). Infection fatality ratio and case fatality ratio of covid-19. *International Journal of Infectious Diseases*, 113:43–46.

Madrigal, A. C. and Moser, W. (2020). How Many Americans Are About to Die? *The Atlantic*.

McNeil, D. G. J. (2020). The Pandemic's Big Mystery: How Deadly Is the Coronavirus? *New York Times*.

National Center for Health Statistics (NCHS) (2023). In-hospital mortality among hospital confirmed covid-19 encounters by week from selected hospitals. National Hospital Care Survey (NHCS).

Nishiura, H., Klinkenberg, D., Roberts, M., and Heesterbeek, J. A. P. (2009). Early Epidemiological Assessment of the Virulence of Emerging Infectious Diseases: A Case Study of an Influenza Pandemic. *PLoS One*, 4(8):e6852.

Overton, C., Webb, L., Datta, U., Fursman, M., Hardstaff, J., Hiironen, I., Paranthaman, K., Riley, H., Sedgwick, J., Verne, J., Willner, S., Pellis, L., and Hall, I. (2022). Novel methods for estimating the instantaneous and overall COVID-19 case fatality risk among care home residents in England. *PLoS Comput Biol*, 18(10):e1010554.

Qu, Y., Lee, C. Y., and Lam, K. F. (2022). A novel method to monitor covid-19 fatality rate in real-time, a key metric to guide public health policy. *Sci Rep*, 12:18277.

Reich, N. G., Lessler, J., Cummings, D. A. T., and Brookmeyer, R. (2012). Estimating Absolute and Relative Case Fatality Ratios from Infectious Disease Surveillance Data. *Biometrics*, 68(2):598–606. Published online 2012 Jan 25.

Roth, G. A., Emmons-Bell, S., Alger, H. M., Bradley, S. M., Das, S. R., de Lemos, J. A., Gakidou, E., Elkind, M. S. V., Hay, S., Hall, J. L., Johnson, C. O., Morrow, D. A., Rodriguez, F., Rutan, C., Shakil, S., Sorensen, R.,

Stevens, L., Wang, T. Y., Walchok, J., Williams, J., and Murray, C. (2021). Trends in Patient Characteristics and COVID-19 In-Hospital Mortality in the United States During the COVID-19 Pandemic. *JAMA Network Open*, 4(5):e218828–e218828.

Russell, T. W., Hellewell, J., Abbott, S., Jarvis, C. I., van Zandvoort, K., Ratnayake, R., CMMID nCov working group, Flasche, S., Eggo, R., Edmunds, W. J., and Kucharski, A. J. (2020a). Using a Delay-Adjusted Case Fatality Ratio to Estimate Under-Reporting. *Fondazione Cerm.*

Russell, T. W., Hellewell, J., Jarvis, C. I., van Zandvoort, K., Abbott, S., Ratnayake, R., working group, C. C.-., Flasche, S., Eggo, R. M., Edmunds, W. J., and Kucharski, A. J. (2020b). Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Eurosurveillance*, 25(12).

Thomas, B. S. and Marks, N. A. (2021). Estimating the case fatality ratio for covid-19 using a time-shifted distribution analysis. *Epidemiol Infect*, 149:e197.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G. T., Fu, H., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 20(6):669–677. Open Access.

Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604.

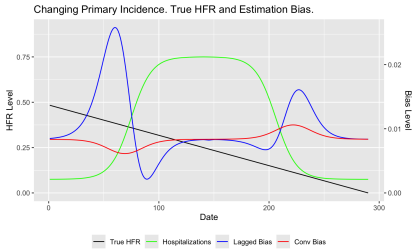
Ward, T. and Johnsen, A. (2021). Understanding an evolving pandemic: An analysis of the clinical time delay distributions of covid-19 in the united kingdom. *PLoS One*, 16(10):e0257978.

Wjst, M. and Wendtner, C. (2023). High variability of COVID-19 case fatality rate in Germany. *BMC Public Health*, 23:416.

Xie, Y., Choi, T., and Al-Aly, Z. (2024). Mortality in Patients Hospitalized for COVID-19 vs Influenza in Fall-Winter 2023-2024. *JAMA*, 331(22):1963–1965.

Yuan, J., Li, M., Lv, G., and Lud, Z. K. (2020). Monitoring transmissibility and mortality of COVID-19 in Europe. *Int J Infect Dis*, 95:311–315.

(a) All deaths after ℓ days. HFR ratios equivalent; plotting delays of $\ell = 14$ and 28 days.



(b) Changing primary incidence. Plotting bias of lagged and convolutional ratios.

Figure 4: Simple examples of severity rate bias.

A Analysis

In this section, we present two examples that further explain the bias. These are more contrived that the ones in 2.2, for example using unrealistic delay distributions. Nevertheless, their bias can be simplified to simple analytic formulas, isolating the three contributing factors.

To elucidate the relationship between changing severity rates and the ratio estimators' bias, consider the trivial case where all secondary events occur after exactly ℓ days with no noise. By definition, $\pi_k = \mathbb{1}\{k = \ell\}$, so the convolutional and lagged ratios are both $\hat{p}_t = \frac{X_{t-\ell} p_{t-\ell}}{X_{t-\ell}} = p_{t-\ell}$ presuming both have access to the oracle delay distribution. Figure 4a displays this with the approximate ground truth HFRs from NHCS.

In this case, the bias is the change in the true severity rate $p_{t-\ell} - p_t$. The estimator is unbiased only when the severity rate is stationary. Otherwise, for example, the ratio will be 20% too low if the true severity rate was 20% lower ℓ days ago.

Intuitively, severity rates may be less similar to the present value p_t further back in time. In this simple example, the bias $p_{t-\ell} - p_t$ is generally larger when $\ell = 28$ than $\ell = 14$ (Fig 4a). This expresses the observation that estimates with heavier-tailed delay distributions tend to have more bias.

Section 2.2 claims that changes in primary incidence levels affect the magnitude of bias for the convolutional ratio. Here, we present simple examples that formalize this claim. First assume primary incidence is constant, in which case the convolutional and lagged ratios are equal. The time series factors neatly out of the bias expression 6:

$$\text{Bias}(\hat{p}_t^{\text{Conv}}) = \text{Bias}(\hat{p}_t^{\text{Lagged}}) = \Big(\sum_{k=0}^{d} \pi_k p_{t-k} \Big) - p_t.$$

This is the difference between a weighted average of previous severity rates and the present. Weights for the historical rates are given by the delay distribution,

providing further justification for its central role in the bias.

Next, suppose half of the secondary events occur immediately after the primary event ($t = 0$), and the other half after ℓ days. Further assume $p_{t-\ell} \neq p_t$, so there is some degree of bias. Then

$$\begin{aligned} |\text{Bias}(\hat{p}_t^{\text{Conv}})| &= \frac{\frac{1}{2}\left|X_t(p_t - p_t) + X_{t-\ell}(p_{t-\ell} - p_t)\right|}{\frac{1}{2}(X_t + X_{t-\ell})} \\ &= \frac{X_{t-\ell}|p_{t-\ell} - p_t|}{X_{t-\ell}(1 + \frac{X_t}{X_{t-\ell}})} = \frac{|p_{t-\ell} - p_t|}{1 + \frac{X_t}{X_{t-\ell}}} \end{aligned}$$

The absolute bias is monotonically decreasing in $\frac{X_t}{X_{t-\ell}}$, the proportion change in primary incidence. Rising primary incidence ($\frac{X_t}{X_{t-\ell}} > 1$) yields less bias, while falling levels yield more.

Figure 4b displays this setting. Hospitalizations are defined as $X = \sigma(s) * 9000 + 1000$, where σ is the sigmoid function and s takes 300 evenly spaced steps from -9 to 7. The true HFRs fall from 0.5 to 0 over the same number of even steps. Indeed, the convolutional ratio's bias dips as hospitalizations rise, and rises as they fall.

When daily hospitalizations approach a constant level, the two estimators become the same ratio, so their biases converge. During periods of change, however, the lagged estimator has different bias. It oscillates up and down, reaching higher bias than the convolutional ratio.

TO DO: ANALYZE WHY THIS HAPPENS.

B Alternative Data Sources

B.1 Retrospective Deaths

JHU presented daily deaths in real time, aligned by the date they were reported. In contrast, the National Center for Health Statistics (NCHS) provided weekly totals for deaths aligned by occurrence, and were not available in real time. Thus, delay distributions with NCHS deaths have a lighter tail.

Figure 5 shows this minor change has a significant effect on the bias. It compares the real-time lagged ratios (Eq. 2) with deaths sourced from JHU and NCHS. JHU is much more biased during the Alpha, Delta, and Omicron periods discussed. For example, NCHS only rises from 12% to 14% as Omicron falls, far below JHU's surge above 25%. As analyzed in 2.2, JHU's heavier-tailed delay distribution inflates the influence of dates with higher HFRs than the present.

B.2 Alternative Ground Truth

We considered two retrospective approaches to approximate the ground truth national HFRs over time. The first approach took lagged ratios with aggregate deaths from NCHS. NCHS is a better resource than JHU because it uses death

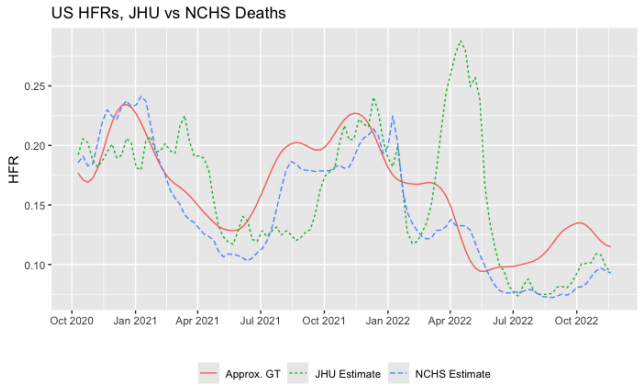


Figure 5: Real-time Lagged Ratios, JHU vs NCHS deaths. Seven-day smoothing with 19- and 11-day lags, respectively.

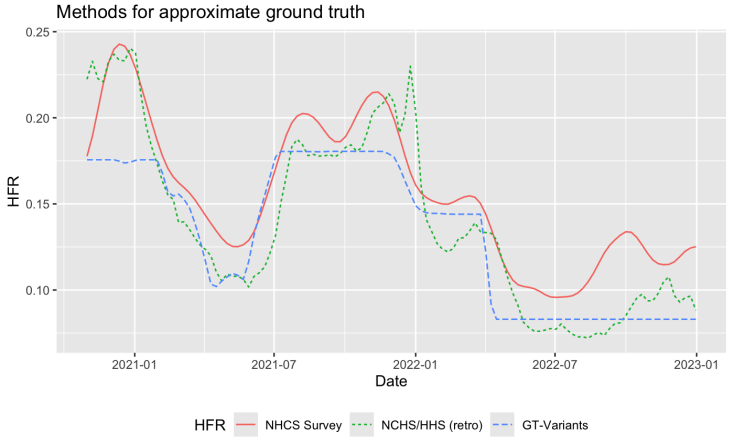


Figure 6: Methods for Retrospective Ground Truth HFRs.

counts from the date they actually occurred, not merely reported. In addition, we take a forward-looking ratio, which is retrospective insofar as it uses data after time t to estimate the HFR.

$$\hat{p}_t^{\text{LaggedRetro}} = \frac{Y_{t+L}}{X_t} \tag{10}$$

The second approach computed a single HFR for each major variant, then mixing by the proportions of variants in circulation. Formally, let $\hat{p}_j$ approximate the HFR of variant j; let v_t^j be its proportion of cases at time t, where $\sum_j v_t^j = 1 \; \forall t$. The HFR estimate is

$$\hat{p}_t^{\text{Var}} = \sum_j v_t^j \hat{p}_j.$$

Each variant's HFR $\hat{p}_j$ was defined as the ratio of total NCHS deaths and HHS hospitalizations during the period where it accounted for over 50% of activate cases. The case proportions v_t^j were obtained from `covariants.org`. To ensure estimates were reasonable, we only considered the 4 largest variants: The original strain, Alpha, Delta, and Omicron. Because Omicron began with an enormous surge that quickly subsided, we split it into early and late periods at April 1, 2022, following Adjei, Stacey and Hong, Kai and Molinari, Noelle-Angelique M and Bull-Otterson, Lara and Ajani, Umed A and Gundlapalli, Adi V and Harris, Aaron M and Hsu, Joy and Kadri, Sameer S and Starnes, Jon and Yeoman, Kristin and Boehmer, Tegan K (2022).

Figure 6 displays the three curves approximating the true HFRs. They have nontrivial differences in magnitude, but move more or less in conjunction. To validate our results, we primarily used the rescaled NHCS HFRs as the least problematic of the three. The retrospective NCHS ratios are subject to statistical bias, expressed in 8. The variant-based HFRs are flatter, as they do not account for other sources of variability. Therefore, they do not explain for the statistical bias within each variant period, which arises due to changes in the underlying severity rate.

C Robustness Checks

C.1 Hyperparameters

In this section we demonstrate the robustness of our findings against choices of hyperparameters. (All results are with the finalized version of JHU deaths.) First, Figure 7 plots performance over choices of window size parameter. We analyze smoothed versions of the lagged estimator

$$\hat{p}_t^{\text{Lagged,W}} = \frac{\sum_{s=t-w+1}^{t} Y_s}{\sum_{s=t-w+1}^{t} X_{s-\ell}}, \tag{11}$$

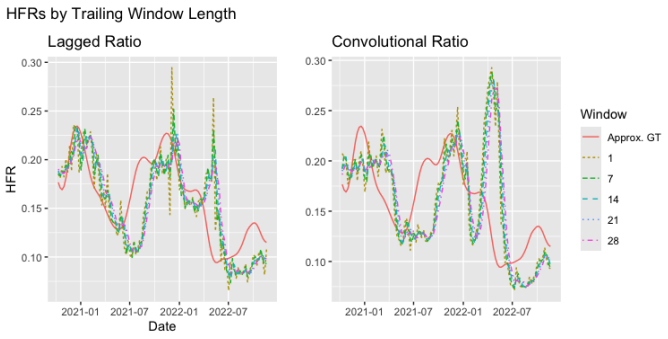


Figure 7: The length of the trailing window bears little impact on the findings.

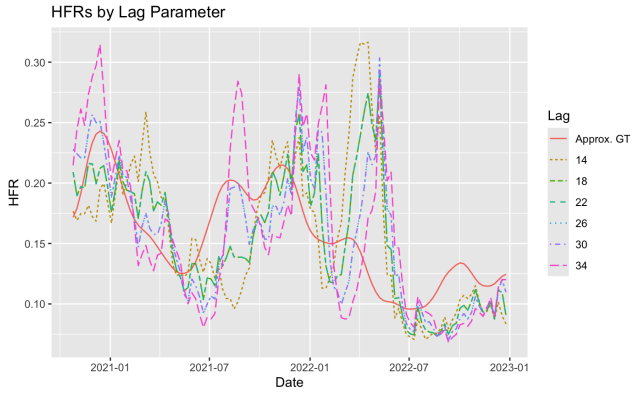


Figure 8: HFRs are biased regardless of what lag parameter is selected.

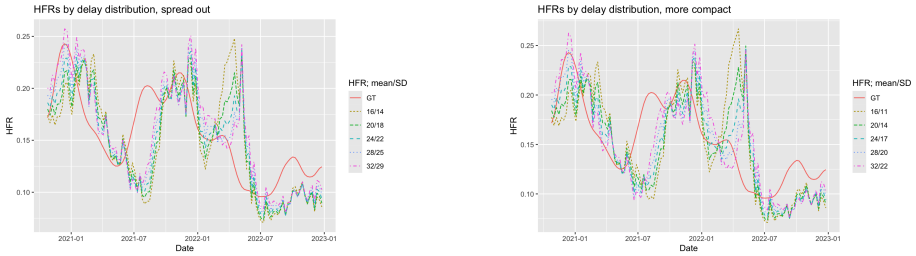
as well as the convolutional estimator

$$\hat{p}_t^{\text{Conv,W}} = \frac{\sum_{s=t-w+1}^{t} Y_s}{\sum_{s=t-w+1}^{t} \sum_{k=0}^{d} X_{s-\ell-k} \hat{\pi}_k}. \tag{12}$$

Results are very similar, indicating the bias does not disappear when smoothing over a longer history.

We next examine the time-to-death hyperparameters: The lag ℓ for the lagged ratio and delay distribution π for the convolutional ratio. Figure 8 displays HFR estimates with lags ranging from 2 to 5 weeks. Unlike the window size, changing this parameter leads to different behavior across lags. Some choices are better than others; a 28-day lag, for example, falls appropriately during Alpha and rises less slowly during Alpha. However, all are biased to varying degrees, most notably the huge spurious surge in spring 2022.

Figure 9 compares the performance of the convolutional ratio across different choices of delay distribution. We kept the discrete gamma shape for each, but varied the mean and standard deviation. As before, Figure 9a kept the standard



(a) SD is 0.9×mean.

(b) SD is 0.7×mean.

Figure 9: Convolutional ratio estimates are biased regardless of which delay distribution is selected.

deviation to 90% of the mean, per Ward and Johnsen. We also evaluated with a more compact delay distribution in 9b.

All HFR estimates in the figures are significantly biased. Regardless of delay distribution, the ratios are negatively biased during the onset of Delta, and surge after the peak of Omicron. This indicates the bulk of the error is fundamental to the estimator, and cannot be attributed to model misspecification.

Comparing to the approximate ground truth HFRs from NHCS, performance improved slightly with a longer delay distribution than the purported mean of 20 days. Its mean absolute error was 0.031, whereas the delay distribution with mean 28 and standard deviation 25 had a MAE of 0.27. Nevertheless, this difference is relatively small, with the alternative delay distribution still showing similar bias.

C.2 Geography

Next, we repeat our analysis on different geographies, finding similar trends. We repeated our computations on the 6 largest US states with the same lag and delay distribution, with finalized death counts from JHU. Because the NHCS survey was conducted on a subset of hospitals meant to represent the US at large, it may poorly approximate the HFRs for individual states. A better state-level source is the retrospective lagged estimate (10) using NCHS deaths. Figure 10 compares this rough ground truth with the real-time estimates. For both NCHS and JHU deaths, we again take the lag that maximizes cross-correlation with hospitalizations; the standard deviation of the delay distribution is 0.9 times the mean.

Several states have similar biases as the US results (Fig. 2a). Ratios in California, Texas, and Florida all are slow to detect the uptick in HFR during Delta; in California and Florida they also spike during Omicron. Note these states are the ones with the largest optimal lags, an estimate of the average time to death. As our simulated examples have shown, the shape of the delay distribution is a key factor behind the degree of bias. In contrast, New York, Pensylvania, and Illinois have mean delays of at most 20; while their HFRs are

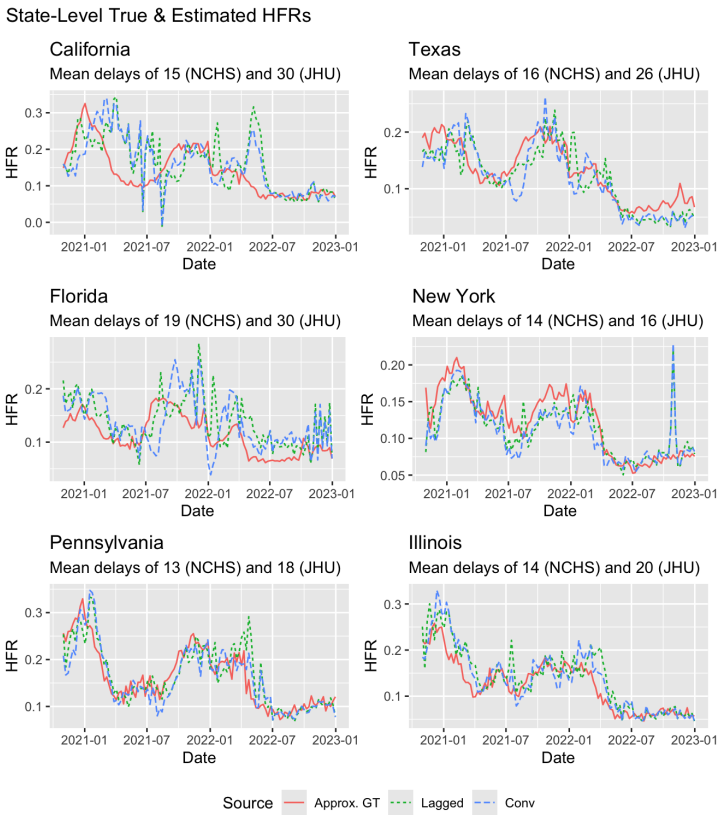


Figure 10: HFRs by individual states. Comparing retrospective estimates with NCHS against real-time estimates with JHU.

still biased, they are relatively close to the NCHS curve. This suggests that fatality ratios are generally less trustworthy in states that take longer to report deaths.