Severity Bias Paper Draft

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 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 (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 that these ratio estimators exhibit fundamental statistical bias. Bias arises when true severity rates change, precisely when time-varying estimates should be most useful. For example, the ratio estimators would have failed to quickly identify the rise in hospitalization-fatality rate (HFR) during 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. To stabilize estimates, smoothed counts are often used in practice Liu et al. (2023); Luo et al. (2021); Wjst and Wendtner (2023), but for the sake of simplicity of presentation we will ignore this.

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 prior works. The real-time estimator only uses data until the present timestep t:

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

Many 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. The expected number of secondary events at any given day can be expressed by convolving the delay distribution against hospitalizations and severity rates.[1]

[1] 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.

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

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. Overton et al. proposed the following convolutional ratio, using a plug-in estimate of the delay distribution:

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

This can be understood as a generalization of 2. If all secondary events occur after L days, then it reduces to the same ratio. Otherwise, it expresses a more accurate relation the two time series.

The ratios in Equations 2 and 4 track a time-varying severity rate. To estimate the average rate over all time, they are often used with cumulative counts. That is, each time series is the sum of all counts from the first timestep. This version of Eq. 4 is unbiased if the severity rate is stationary, and is widely used in practice Nishiura et al. (2009). For the time-varying setting, however, the lagged ratio is much more common.

2.2 Mathematical Analysis

In this section, we demonstrate that these time-varying severity ratios are biased when the true rates are changing. Assume the true delay distribution is a constant π over all time with maximum length d. We first analyze the convolutional ratio (Eq. 4), assuming the oracle delay distribution π is known.

$$
\begin{aligned}
\text{Bias}(\hat{p}_t^{\text{Conv}}) = E[\hat{p}_t^{\text{Conv}}] - p_t &= \frac{E[Y_t]}{\sum_{k=0}^{d} X_{t-k}\pi_k} - p_t \\
&= \frac{\sum_{k=0}^{d} X_{t-k}\pi_k p_{t-k}}{\sum_{k=0}^{d} X_{t-k}\pi_k} - \frac{p_t \sum_{k=0}^{d} X_{t-k}\pi_k}{\sum_{k=0}^{d} X_{t-k}\pi_k} \\
&= \sum_{k=0}^{d} \frac{X_{t-k}\pi_k}{\sum_{j=0}^{d} X_{t-j}\pi_j}(p_{t-k} - p_t). \qquad (5)
\end{aligned}
$$

The degree of bias in Eq. 5 depends on three factors.

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.

   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.

   Example. To elucidate the extent to which changing severity rates bias this estimator, consider the trivial case where all secondary events occur after exactly ℓ days. By definition, $\pi_k = \mathbb{1}\{k = \ell\}$, so the bias reduces to $p_{t-\ell} - p_t$. So if the true severity rate was 20% lower ℓ days ago, the ratio will be 20% 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.

   Example. Constructing another simple case, let the severity rate be changing monotonically before $t - m$, then constant until t. If the delay distribution assigns all probability mass within the first m days, the convolutional ratio will be unbiased. Otherwise, it will be biased, with longer delay distributions producing greater bias. [AH: In this and maybe the other examples, a figure would help the reader...

   (a) (For example) in black: the true p_t as outlined above

   (b) in blue (matplotlib C0): $\hat{p}_t^{conv}$ for a delay distribution that has all mass within first m days

   (c) in orange (C1) : $\hat{p}_t^{conv}$ with delay distribution that has mass outside first m days

   If we can also plot the convolutional kernels and show how they spread mass differently, in ways that produce more / less bias, that would be useful too..

   Especially given that there are plots in the results section that show evolving bias over a two-year period, showing how the bias can be constructed in these toy examples visually can prepare the reader to break down what they see in the subsequent sections more easily]

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.

   Example. Suppose half of the secondary events occur after exactly a days, and the other half after $b > a$. Further assume severity rate is the same a days before t, but different b days prior. Then

$$\begin{aligned}\text{Bias}(\hat{p}_t^{\text{Conv}}) &= \frac{\frac{1}{2}\big(X_{t-a}(p_{t-a} - p_t) + X_{t-b}(p_{t-b} - p_t)\big)}{\frac{1}{2}(X_{t-a} + X_{t-b})} \\ &= \frac{X_{t-b}(p_{t-b} - p_t)}{X_{t-b}(1 + \frac{X_{t-a}}{X_{t-b}})} = \frac{p_{t-b} - p_t}{1 + \frac{X_{t-a}}{X_{t-b}}}\end{aligned}$$

   This bias is monotonically decreasing in $\frac{X_{t-a}}{X_{t-b}}$, the rate of change in primary incidence. Rising primary incidence ($\frac{X_{t-a}}{X_{t-b}} > 1$) yields less bias, while falling levels yield more.

This bias expression for the lagged ratio obeys roughly the same structure:

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

Again, the bias will be positive when the HFR is falling, and vice versa. Indeed, empirically the two ratio estimators are shown to have similar bias (Section 3). However, Eq. 6 is more difficult to analyze due to its reliance on the lag L, hence this section's focus on the convolutional ratio.

[AH: Can we frame discussion of the bias for the lagged estimator through Bias Factor 2, "shape of the delay distribution", and use the fact that the lagged estimator is a convolution estimator with $\pi = 1\{k = L\}$? How does that interact with the Example given for Bias Factor 2?

In fact, in the spirit of Ryan's question, "if I were a referee, I would ask how the bias is affected by different choices of convolutional kernel in the estimator (if not the oracle), and by different choices of lag — would it make sense to do (5) without assuming the oracle delay distribution is known; obtaining a bias expression that has separate $\hat{\pi}$ from oracle π, and then specializing it to the cases where we know the oracle, so $\hat{\pi} = \pi$, and where $\hat{\pi}$ puts all mass on L?

I'm not specifically saying this should be done; just that we should consider doing it if it streamlines the exposition/clarifies the differences.]

[RJT: I realized the following: in (5), we're assuming that the convolutional model is well-specified (and that the estimator uses the oracle delay distribution) but in (6), we have not assumed that the lagged model is well-specified. In other words, we've looked at the well-specified bias of the convolutional estimator in (5), but the mis-specified bias of the lagged estimator in (6). I think this is why it's hard to interpret the latter, and creates some confusion.

If we assume that the lagged model is well-specified, the (5) becomes easy to interpret (just like AH says in his comment above ... because in this case can just interpret this in the context of the point #2 in the discussion of the convolutional estimator). This is a particular (point mass) delay distribution. If this aligns with a big change in the severity rate, then the bias will be high. Right?]

2.3 Experimental Setup

2.3.1 HFR Estimation

Our experiments focus on the Hospitalization-Fatality Rate, or HFR. While this is a less commonly used metric than the Case-Fatality Rate (CFR), the two should exhibit the same statistical biases. Moreover, CFR may slightly misrepresent our analyses because the model relating cases to deaths may be slightly different. Case reporting during Covid suffered significant time delays, with some cases being reported *after* the death. Therefore, the delay distribution for CFR may not have non-negative support. Unlike case data, the Department of Health and Human Services (HHS) reported hospitalizations in real time based on the date they occurred. This makes hospitalizations a more suitable choice of primary event for study.

During the pandemic, data sources counted deaths in different ways. John Hopkins University (JHU) provided the definitive resource for real-time death counts, releasing figures daily. However, these counts reflected the times at which deaths were reported to health authorities, not necessarily when they actually happened. In contrast, the National Center for Health Statistics (NCHS) provide weekly totals for deaths aligned by occurrence. Unlike JHU, the NCHS deaths were not available in real time.

To compute real-time HFRs, we use hospitalizations from HHS and deaths from JHU. Data was pulled from the `epidatr` API, developed by the Delphi Group. HHS reported hospitalizations weekly, so we smoothed them with a spline to produce daily counts. Raw JHU death counts exhibit high volatility due to reporting idiosyncrasies like day-of-week effects and data dumps. For that reason, we used a 7-day trailing average of counts. For each dataset, we used the finalized counts, not the versions that would have been available in real time. Doing so would have likely resulted in even more volatile HFR estimates. We also obtained finalized weekly deaths from NCHS, again smoothing with a spline.

The two ratio estimators (Eq. 2 and 4) require choices of lag and delay distribution. We tested the robustness of findings against different hyperpa-

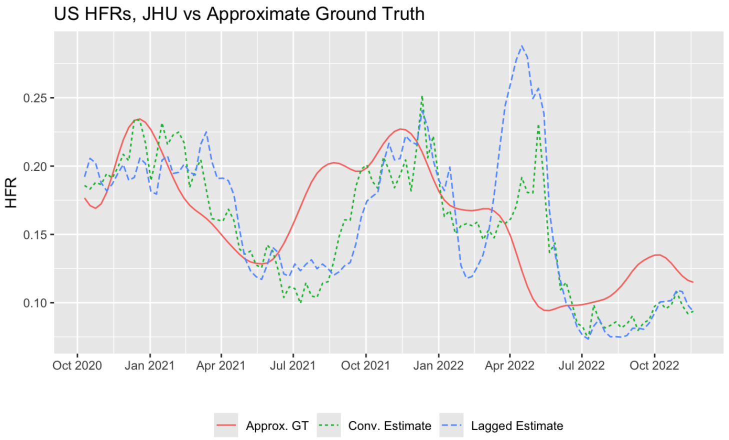

Figure 1: Time-varying HFR estimates are highly biased. Terms in both ratio estimates are smoothed over a 7-day window.

rameter values. A common choice for delay distribution is a discrete gamma distribution, which we use for the convolutional ratio.

2.3.2 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. The NHCS, a representative subset of hospitals across the US, recorded weekly HFRs from in-patient deaths. We smoothed these HFRs with a spline to account for inter-week variability.

HFRs obtained 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 in-patient 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 after again smoothing with a spline.

We considered two other sources for ground truth HFRs, discussed in Appendix A. These are fairly consistent with the rescaled NHCS data, but entail making worse assumptions.

3 Results

3.1 National Data

Figure 1 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 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. [AH: This is great — I think we would really drive the point home if in addition to Figure 1, we had a three-panel figure that "zooms in" on each of these cases, and specifically labels them with each of the three failure modalities, so that the reader can see exactly how they correspond to the conditions under which bias occurs outlined in Section 2. (Even better if there are toy examples in Section 2, so the reader can see the "idealized version" of the failure, and a "real world" manifestation of that failure mode.]

Lastly, the shape of the delay distribution plays a significant role in the bias. Figure 2 compares the real-time lagged ratios 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%. The difference can be primarily attributed to JHU having a much heavier-tailed delay distribution. As analyzed in 2.2, this inflates the influence of dates with higher HFRs than the present.

We performed several robustness checks to assess the stability of these findings. Figure 1 compares the convolutional and lagged ratio estimators, finding both exhibit bias. The convolutional estimator is slightly better, but still very problematic. Appendix B 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. For a series of time-varying HFRs p_t and delay distribution π, deaths are defined as

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

To mimic the real data, we used the same HFRs from NHCS used for validation in 3.1. We also inverted them and rescaled in order to simulate the opposite

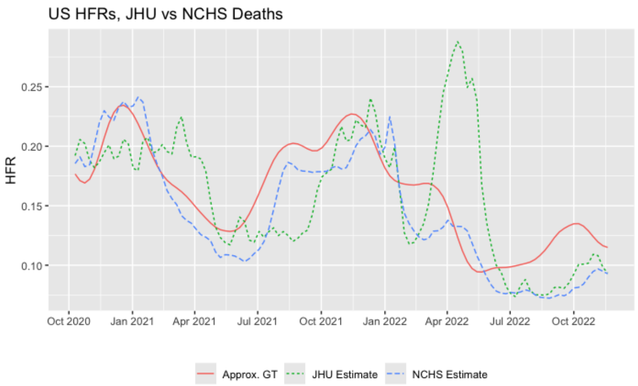

Figure 2: Real-time Lagged Ratios, JHU vs NCHS deaths. Seven-day smoothing with 19- and 11-day lags, respectively.

trend. Lastly, we explored a stationary HFR of 10% over all time. In addition, we used the delay distributions that produced the most reasonable convolutional HFRs on the observed death counts. For NCHS, this was a gamma distribution with mean 11 and standard deviation 10; for JHU, these quantities were 28 and 21, respectively.

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. 5). Nevertheless, the lagged ratio is the standard time-varying estimator in practice.

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 estima-

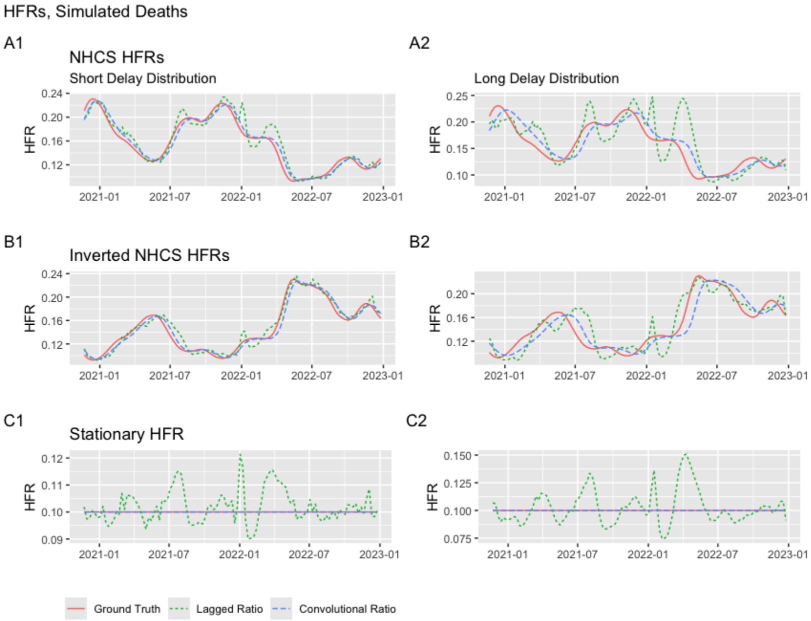

Figure 3: True and Estimated HFRs from Simulated Deaths. First column has short delay distribution, second has long.

tors is to inform stakeholders of changing risks in real time; this bias indicates they may fail to do so in a prompt 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).

When only aggregate data is available, 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.

In a similar vein, deaths from NCHS should be used for retrospective analysis, not JHU. Longer delay distributions have been shown to produce significantly more bias (Fig. 2, 3). Therefore estimates with data that counts secondary events by report date will always be worse. 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. 1, 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).

TO DO: Discuss other sources of bias in severity rates. e.g. Anastasios, Nick Reich papers.

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):

$$\hat{R}_t = \frac{I_t}{\sum_{k=0}^{d} I_{t-k} g_k}. \tag{7}$$

Fraser notes that instantaneous R_t is equal to case R_t if conditions remain unchanged. Similarly, we demonstrated that the convolutional ratio 4 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.

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.

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

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.

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

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*.

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.

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.

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.

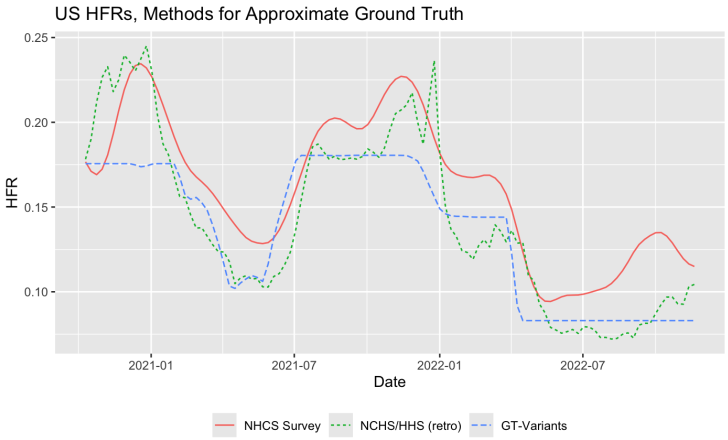

Figure 4: Methods for Retrospective Ground Truth HFRs.

A 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 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.

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

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).

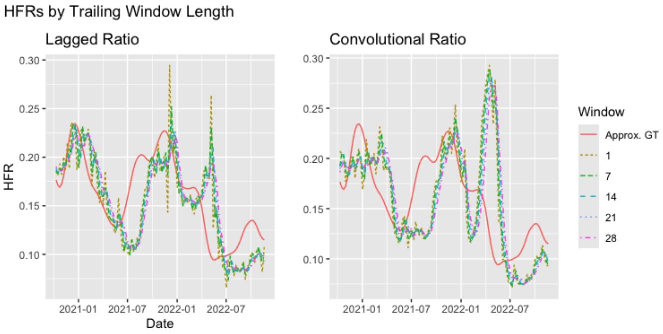

Figure 5: The length of the trailing window bears little impact on the findings.

Figure 4 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 6. The variant-based HFRs are unreasonably flat, as they do not account for other sources of variability. Therefore, they fail to explain for the bias of the estimated HFRs, which arises due to changes in the underlying rate.

B Robustness Checks

In this section we demonstrate the robustness of our findings against choices of hyperparameters and geography. First, Figure 5 plots performance over choices of lag parameter in the lagged ratio (Eq. 2). 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 6 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.

TO DO: Different delay distribution

TO DO: Different state

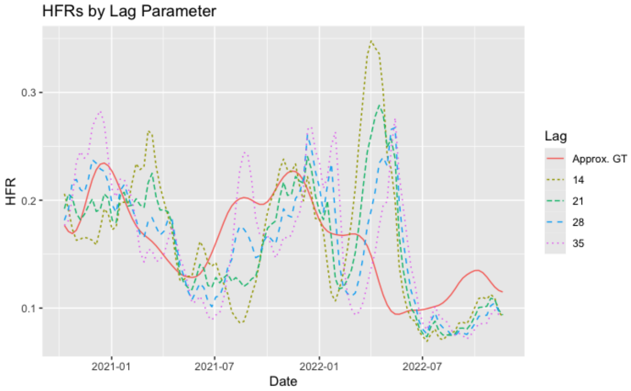

Figure 6: HFRs are biased regardless of what lag parameter is selected.