Challenges in Estimating Time-Varying Epidemic Severity Rates from Aggregate Data

Jeremy Goldwasser* Addison J. Hu† Alyssa Bilinski‡ Daniel J. McDonald§
Ryan J. Tibshirani*

Abstract

Severity rates like the case-fatality rate and infection-fatality rate are key 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. In practice, time-varying severity rates are typically estimated using a ratio of aggregate counts. We demonstrate that these estimators are capable of exhibiting large statistical biases, with concerning implications for public health practice, as they may fail to detect heightened risks or falsely signal nonexistent surges. We supplement our mathematical analyses with experimental results on real and simulated COVID-19 data. Finally, we briefly discuss strategies to mitigate this bias, drawing connections with effective reproduction number (R_t) estimation.

1 Introduction

Several public health metrics of interest express the probability that a second, often more serious outcome will follow a primary event. For example, the case-fatality rate (CFR) and infection-fatality rate (IFR) are commonly used to assess the deadliness of an epidemic (Garske et al., 2009; Russell et al., 2020b; Challen et al., 2021; Luo et al., 2021; COVID-19 Forecasting Team, 2022). Another central example of a "severity rate", which is a term that we use for metrics of this general form, is the hospitalization-fatality rate (HFR) (Bellan et al., 2020; Roth et al., 2021; Xie et al., 2024).

In an ideal setting, severity rates can be obtained directly from a comprehensive line-list or claims data set containing individual patient outcomes (Bellan et al., 2020; Challen et al., 2021; Roth et al., 2021; Xie et al., 2024). However, in fast-moving epidemics like COVID-19, large-scale tracking of individuals has been infeasible, especially in real-time. Instead, severity rates are routinely estimated from aggregate count data. While it is common to assume they are constant in time (Ghani et al., 2005; Jewell et al., 2007; Reich et al., 2012; Baud et al., 2020), consequential shifts in severity rates can occur in response to factors such as new therapeutics, vaccines, and variants (McNeil, 2020). Time-varying severity rates are often estimated with a ratio of primary and secondary aggregate data streams. For example, aggregate case and death counts were widely used to estimate and report COVID-19 CFRs, both in the academic literature (Yuan et al., 2020; Luo et al., 2021; Horita and Fukumoto, 2022; Liu et al., 2023; Wjst and Wendtner, 2023) and also in major news outlets like The Atlantic (Madrigal and Moser, 2020) and The Wall Street Journal (Kamp and Krouse, 2020). In fact, ratio estimators are so common that CFR is often (mis)labeled the case-fatality *ratio*.

In this work, we show that these ratio estimators are prone to nontrivial statistical bias. Bias arises as a consequence of changing severity rates—precisely when time-varying estimates should be most useful. It also arises due to misspecification of the delay distribution, which relates events, like cases and deaths in CFR. This is particularly troublesome for the popular lagged ratio estimator, which divides values of two aggregate data streams. We validate these findings empirically, tracking the hospitalization-fatality rate (HFR) during COVID-19. The ratio estimators failed to quickly signal increased risk in the onset of the Delta wave; later, in the aftermath of the inital Omicron wave, they surged while the true HFR fell. We provide heuristics for when to expect this bias in practice, and discuss ideas for alternative methodology which may avoid it.

*Department of Statistics, University of California, Berkeley
†Department of Statistics, Carnegie Mellon University
‡Departments of Health Policy and Biostatistics, Brown University
§Department of Statistics, University of British Columbia

2 Methods

In this section, we introduce the main estimators we study, and analyze their bias. Subsequently, we detail the data used for empirical study and validation.

2.1 Severity rate estimators

Severity rates convey the probability that a primary event will result in a secondary event in the future. In the case of CFR, for example, a primary event is a positive COVID-19 case and a secondary event is a death with a positive test result. Formally, a time-varying severity rate at time t is generally defined as:

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

Here, t may represent a discrete interval of time, such as a given day or week. It also may be understood in a continuous-time fashion. Although this will not be our focus in this paper, the same general principles apply in the continuous-time case. For simplicity, we will consider only the discrete-time setting, and we index time steps via integers, as in $t = 0, 1, 2, \ldots$.

Throughout, we denote by $\{X_t\}$ and $\{Y_t\}$ the aggregate time series of new primary and secondary events, respectively. These are often counts, and we will generally refer to them as such. At time t, we assume data for all past $s \le t$ is available, but future data is not. Therefore real-time estimates of p_t can only rely on past counts $\{X_s\}_{s \le t}$ and $\{Y_s\}_{s \le t}$. In practice, to stabilize estimates, smoothed counts are often used in place of raw counts. This may be simply absorbed into the notation for X_t and Y_t, and we do not address smoothing explicitly in the main text, but refer back to this issue in the appendix.

Lagged estimator. The canonical estimator for time-varying severity rates is a ratio between the counts of primary and secondary events, offset by a lag ℓ. This estimator is widely-used in epidemiology, both in the academic literature and in public health practice and communication (e.g., Kamp and Krouse, 2020; Madrigal and Moser, 2020; Yuan et al., 2020; Luo et al., 2021; Thomas and Marks, 2021; Horita and Fukumoto, 2022; Liu et al., 2023; Wjst and Wendtner, 2023). For concreteness, we define the *lagged ratio* at time t as:

$$\widehat{p}_t^{\ell} = \frac{Y_t}{X_{t-\ell}}, \tag{2}$$

where $\ell \ge 0$ is a given parameter (often chosen to maximize cross-correlation between $\{X_t\}$ and $\{Y_t\}$).

Convolutional estimator. Alternative methods for estimating severity rates utilize a delay distribution, which relates the two time series. The delay distribution at time t and lag k is defined as:

$$\pi_k^{(t)} = \mathbb{P}(\text{secondary event at } t+k \,|\, \text{primary event at } t, \text{ secondary event occurs}).$$

Throughout, we assume that the delay distribution has a finite support of d time steps. We also assume the delay distribution is stationary: $\pi_k^{(t)} = \pi_k$ for all k and t. (In reality, delay distributions may themselves be time-varying, and this creates its own set of challenges, which only exacerbate the ones we highlight in this paper for a fixed delay distribution.) While the delay distribution is generally unknown, several tools exist to estimate them from aggregate or line-list data (see Charniga et al., 2024 for a nice review). Given π, we can express the expected number of secondary events at time t as follows:

$$\begin{aligned}
\mathbb{E}[Y_t \,|\, \{X_s\}_{s \le t}] &= \sum_{k=0}^{d} X_{t-k} \mathbb{P}(\text{secondary at } t \,|\, \text{primary at } t-k) \\
&= \sum_{k=0}^{d} X_{t-k} \mathbb{P}(\text{secondary after } k \text{ time steps} \,|\, \text{secondary occurs, primary at } t-k) \\
&\qquad \times \mathbb{P}(\text{secondary occurs} \,|\, \text{primary at } t-k) \\
&= \sum_{k=0}^{d} X_{t-k} \pi_k p_{t-k}.
\end{aligned} \tag{3}$$

This is a convolution of the delay distribution against the product of primary incidence and the severity rate. If the severity rate remains constant, $p_{t-k} = p_t$ for all k between 0 and d, then the expression in (3) simplifies to $\mathbb{E}[Y_t \,|\, \{X_s\}_{s \le t}] = p_t \sum_{k=0}^{d} X_{t-k} \pi_k$. As studied in Overton et al. (2022), we can rearrange this relationship in order to estimate the severity rate at t, after plugging-in an estimate γ of the delay distribution π:

$$\widehat{p}_t^{\gamma} = \frac{Y_t}{\sum_{k=0}^{d} X_{t-k} \gamma_k}. \tag{4}$$

We call this the *convolutional ratio* estimator of the severity rate; in the notation here, the superscript in $\widehat{p}_t^{\gamma}$ emphasizes that γ is the distribution used in the definition of the estimator (4). To reiterate, in moving from (3) to (4), we are implicitly assuming that the severity rate p_t is constant over the interval of time from $t - d$ and t. Of course, this runs in contradiction to the fact that we are trying to estimate a time-varying severity rate in the first place. As we will see shortly, this can create significant bias in the convolutional ratio.

Some further comments are in order. The convolutional ratio has a longer history of study and use in the literature on estimating stationary severity rates. Indeed, Nishiura et al. (2009) developed the estimator in this setting, and used it to analyze the CFR in the H1N1 influenza pandemic of 2009. (The only difference to (4) is that in the stationary case we aggregate both the numerator and denominator over all past data.) This estimator, which is sometimes called the *delay-adjusted* estimator of the severity rate, is popular in the academic literature and public health practice (e.g., Garske et al., 2009; Russell et al., 2020b,a; Unnikrishnan et al., 2021), though arguably less popular than the lagged ratio. The package `cfr` (Gupte et al., 2024) gives an R implementation, for both the stationary and time-varying cases.

Furthermore, we note that (4) can be seen as a generalization of the lagged ratio estimator (2): when we take γ to be a point mass at lag ℓ, i.e., $\gamma_\ell = 1$ and $\gamma_k = 0$ for all $k \neq \ell$, then (4) reduces to (2).

Connection with reproduction numbers. Severity rates bear a natural connection with reproduction numbers. Both the true severity rate as defined in (1) and the case reproduction number R_t are defined as the average number of secondary events produced by a single primary event at t, but in contrast to severity rates, reproduction numbers have infections as the primary and secondary events, where a single infection can produce more than one follow-on infection. Playing the role of the delay distribution π in our setting is the generation interval distribution in reproduction numbers, which measures the time in between primary and secondary infections.

Severity rates and reproduction numbers can also be estimated similarly. Because primary events at t produce secondary events after t, their effect is not observed in real time. Therefore, standard real-time estimators for severity rates (2), (4) and for R_t both analyze the number of secondary events at t produced by relevant primary events. In words, this adopts a "backwards-looking" perspective (possible in real time), rather the the "foward-looking" perspective inherent to the definition in (1).

For reproduction numbers, the backwards-looking quantity has its own name: *instantaneous* R_t, which is the average number of secondary infections at time t produced by a single primary infection in the past. One of the most popular traditional estimators of instantaneous R_t is based on an essentially identical idea to the convolutional ratio (Fraser, 2007; Wallinga and Lipsitch, 2007):

$$\widehat{R}_t = \frac{I_t}{\sum_{k=1}^{d} I_{t-k} g_k}. \tag{5}$$

Here, I_t denotes the number of new infections at t, and g is the generation interval distribution. Note that the difference between (4) and (5) is that the latter uses the same aggregate time series, infections, in both the numerator and denominator. While more modern frameworks for estimating instantaneous R_t are often Bayesian (e.g., Cori et al., 2013), the underlying basic point estimates are still the same.

2.2 Well-specified analysis

First we analyze the bias of the convolutional ratio (4) in what we call the *well-specified* case, where the true delay distribution π is known. Formally, for an estimator $\widehat{p}_t$ of p_t, we define its bias as

$$\text{Bias}(\widehat{p}_t) = \mathbb{E}[\widehat{p}_t \,|\, \{X_s\}_{s \le t}] - p_t.$$

Proposition 1. *The bias of the convolutional ratio $\widehat{p}_t^{\pi}$ (where the true delay distribution π is known) is*

$$\text{Bias}(\widehat{p}_t^{\pi}) = \sum_{k=0}^{d} \left[\frac{X_{t-k}\pi_k}{\sum_{j=0}^{d} X_{t-j}\pi_j} (p_{t-k} - p_t) \right]. \tag{6}$$

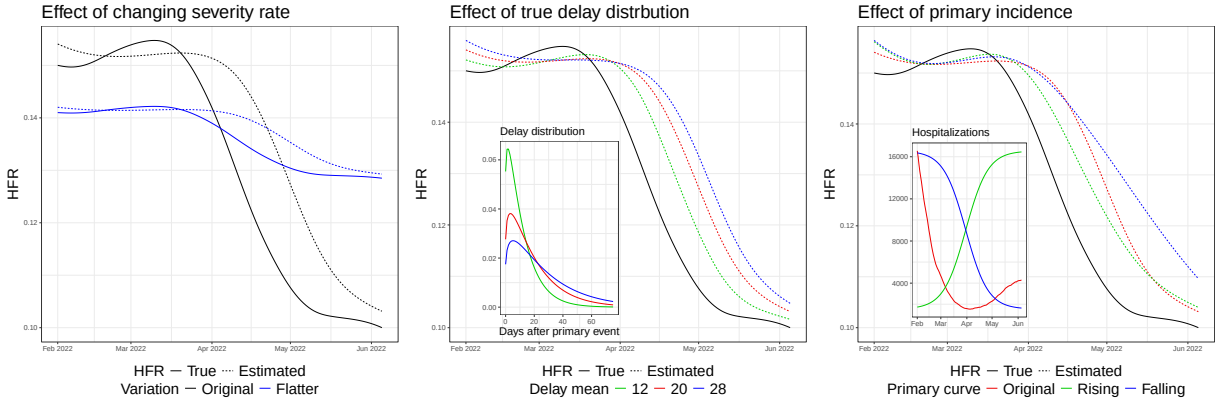
The proof of Proposition 1 is elementary and given in Appendix A.1. The bias (6) of the convolutional ratio estimator of the severity rate depends on three factors, which we discuss next. Figure 1 provides an accompanying illustration.

1. **Changes in severity rate**. The central component of this bias expression is the difference $p_{t-k} - p_t$. When the severity rate is constant over the d preceding time points, the convolutional ratio is unbiased (because this difference is zero). This falls in line with the motivation used to derive this estimator, as explained in the last subsection. But when severity rates change before t, these difference terms will be nonzero, in which case the estimator will be generally biased. Figure 1a shows a simple example of this: the estimated severity rates are most inaccurate in periods where the true rate is changing quickly.

   To make matters worse, the bias is in the opposite direction of the trend we want to detect; for example, suppose the severity rate is monotonically falling, with $p_t < p_{t-1} < \cdots < p_{t-d}$. The bias will then be positive, meaning the ratio estimates do not decline with the true rate. In fact, the estimated severity may even rise, not fall. Conversely, when true severity rates are rising, the 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 π. In general, the bias will be greatest when the delay distribution has a long enough tail to upweight significant differences in severity rate. While this distinction may appear subtle, Section 3 highlights its surprisingly large effects. The simple example in Figure 1b also shows significant differences in bias between shorter and longer delay distributions.

3. **The primary incidence curve.** Changing primary incidence X_t will also affect the bias, presuming the severity rate changes roughly monotonically in the recent past. Intuitively, this up- or down-weights the terms $X_{t-k}\pi_k(p_{t-k} - p_t)$ at times further from the present, which are likely to contribute the most bias. In general, falling primary incidences will amplify the bias, whereas rising events will minimize it. Figure 1c provides an illustration of this phenomenon.



(a) (b) (c)

Figure 1: Simple examples which illustrate the effects of the three factors explained above on the bias (6) in estimating the severity rate. We take the primary incidence curve to be COVID-19 hospital admissions, as reported to the HHS in early 2022. We then simulate a secondary incidence curve, COVID-19 deaths, from (3) without noise. The underlying HFR curve p_t and delay distribution π used in the simulation were derived from external data sources, as explained in more detail in Section 2.4.

Appendix A.3 provides further analysis, by discussing a simplified setting in which the bias (6) described in Proposition 1 itself simplifies in elucidating ways.

2.3 Misspecified analysis

We now analyze the bias of the convolutional ratio (4) for an arbitrary distribution γ. Recall that π denotes the true delay distribution in (3). We refer to the present as the *misspecified* case, as γ may be (arbitrarily) different from π.

Proposition 2. *The bias of the convolutional ratio $\widehat{p}_t^{\gamma}$ (where the true delay distribution π is unknown, and the working delay distribution γ is arbitrary, but also supported on d time steps) is*

$$\mathrm{Bias}(\widehat{p}_t^{\gamma}) = A_t^{\gamma} \, \mathrm{Bias}(\widehat{p}_t^{\pi}) + p_t(A_t^{\gamma} - 1), \tag{7}$$

where $A_t^{\gamma} = \sum_{j=0}^{d} X_{t-j}\pi_j \,/\, \sum_{j=0}^{d} X_{t-j}\gamma_j$. This compares how the delay distributions convolve against the most recent primary incidence levels.

The proof of Proposition 2 is again elementary, and given in Appendix A.2. Under misspecification, the proposition gives an additive decomposition (7) of the convolutional ratio bias, based on the well-specified bias $\mathrm{Bias}(\widehat{p}_t^{\pi})$ (as studied in Proposition 1), and a misspecification factor A_t^{γ}. At the outset, we note that if $\pi = \gamma$ (no misspecification), we have $A_t^{\gamma} = 1$, and (7) reduces to the well-specified bias. Generally, values of $A_t^{\gamma} > 1$ amplify the oracle bias and add positive misspecification bias $p_t(A_t^{\gamma} - 1) > 0$; meanwhile, values of $A_t^{\gamma} < 1$ shrink the oracle bias and add negative misspecification bias $p_t(A_t^{\gamma} - 1) < 0$.

Whether or not misspecification contributes a larger magnitude of bias overall hence depends on whether or not the sign of the misspecification term $p_t(A_t^{\gamma} - 1)$ agrees with the sign of the oracle bias $\mathrm{Bias}(\widehat{p}_t^{\pi})$. This need not always be the case, though in our experience, it is often true in both real and simulated experiments, as we will see in Section 3. Here, to gain more insight, we study the behavior of the bias in three settings:

1. Smooth γ with a lighter tail and smaller mean than π (more mass concentrated at recent time points).
2. Smooth γ with a heavier tail and larger mean than π (less mass concentrated at recent time points).
3. Nonsmooth γ, with a point mass at lag ℓ; we reiterate that in this case the convolutional ratio reduces to the lagged ratio $\widehat{p}_t^{\ell}$ in (2). We also note that its misspecification factor A_t^{γ} reduces to a quantity we similarly denote $A_t^{\ell} = \sum_{j=0}^{d} X_{t-j}\pi_j \,/\, X_{t-\ell}$, and its bias (7) reduces to

$$\mathrm{Bias}(\widehat{p}_t^{\ell}) = \frac{\sum_{j=0}^{d} X_{t-j}\pi_j}{X_{t-\ell}} \, \mathrm{Bias}(\widehat{p}_t^{\pi}) + p_t\left(\frac{\sum_{k=0}^{d} X_{t-k}\pi_k}{X_{t-\ell}} - 1\right). \tag{8}$$

We discuss the behavior of the bias in these three settings, as a function of the primary incidence curve (which drives bias through the misspecification factor A_t^{γ}). Figure 2 provides an accompanying illustration.

Primary incidence rising. Consider the case where primary events are rising—first slowly, then rapidly before leveling off. A lighter-tailed γ will place more weight on recent time points, with higher counts, than π; thus $\sum_{j=0}^{d} X_{t-j}\gamma_j > \sum_{j=0}^{d} X_{t-j}\pi_j$, so $A_t^{\gamma} < 1$. The opposite occurs for a heavier-tailed γ: this places more weight on distant low-count time points and less on the ongoing surge, so $A_t^{\gamma} > 1$. A point mass distribution γ places all of its mass on $X_{t-\ell}$, and during the steepest phase of the rise, this is considerably less than X_t. The true delay π distributes mass across the d most recent time points, and a large fraction of its mass will be convolved against the last $\ell - 1$ time points, whose counts exceed $X_{t-\ell}$. The times before $t - \ell$ have less of an offsetting effect, because incidence has risen at a growing rate. Hence, during a surge we will see $X_{t-\ell} < \sum_{j=0}^{d} X_{t-j}\pi_j$, and $A_t^{\ell} > 1$. However, the behavior of A_t^{ℓ} will be generally more erratic than A_t^{γ} for a smooth distribution γ, as the denominator in the former is less smooth as t varies.

Figure 2 visualizes this as hopitalizations (primary events) rise between December 2021 and mid-January 2022. Throughout this period A_t^{γ} is below/above 1 for the light-tailed/heavy-tailed γ. Meanwhile, A_t^{ℓ} spikes to 1.25 in early January. Correspondly, the lagged HFR rises to 20% when the true one drops to 15%.

Primary incidence falling. Next assume primary incidence reaches a maximum and begins to fall. The smooth distributions behave much the same as when incidence was rising. The light-tailed distribution has more mass around the peak than π, so $A_t^{\gamma} < 1$. Conversely, $A_t^{\gamma} > 1$ for the heavier-tail distribution because

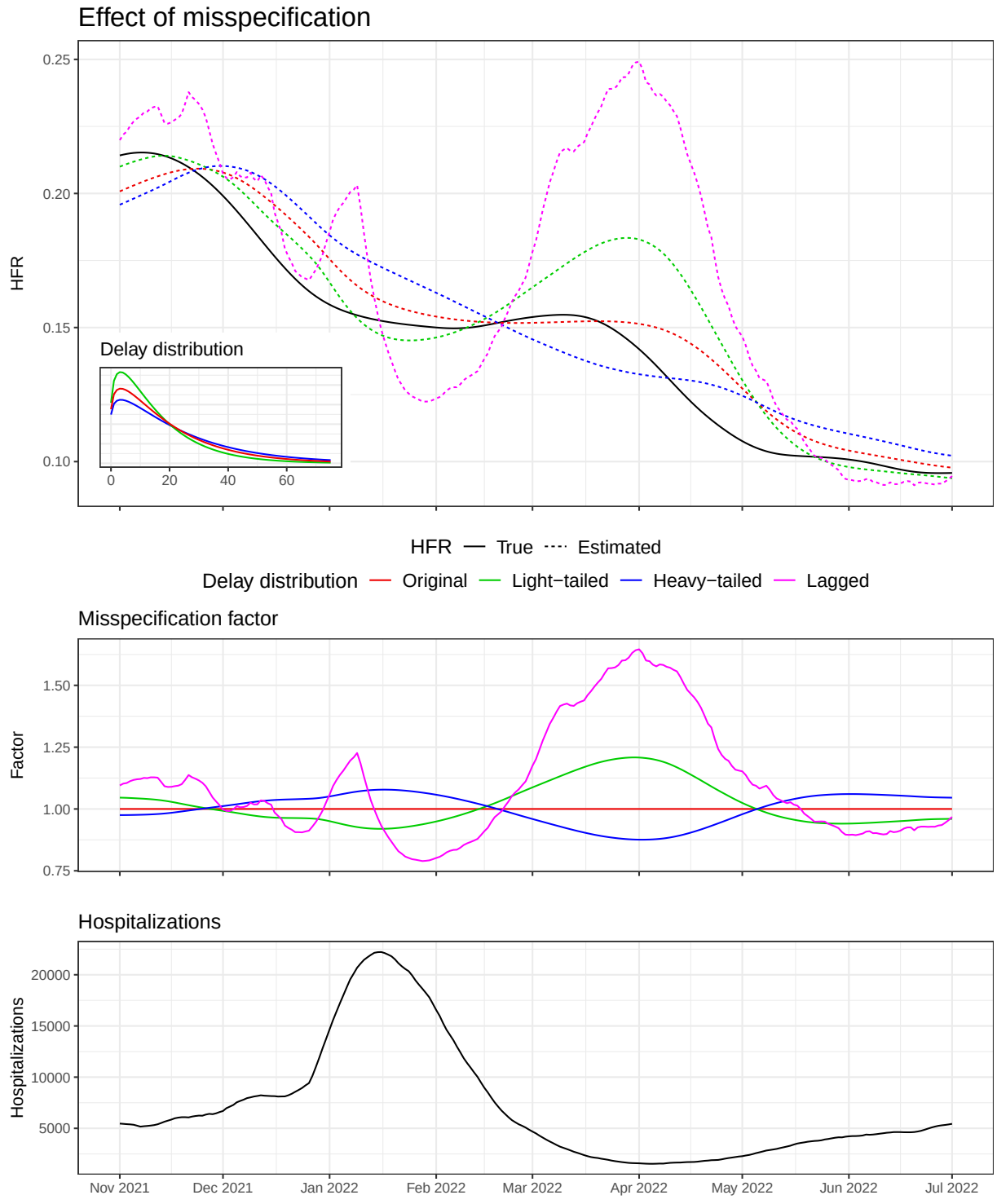


Figure 2: Examples of convoluational ratio estimates under misspecification. As in Figure 1, the primary events are COVID-19 hospitalizations, as reported to the HHS, and secondary events are deaths simulated noiselessly from (3). The underlying HFR curve p_t and delay distribution π used in the simulation were fit using external data sources detailed shortly in Section 2.4. The lagged ratio estimator used $\ell = 16$, chosen to maximize cross-correlation (between hospitalizations and deaths).

it convolves more mass before the top of the rise. The lagged bias changes its behavior in this period; while A_t^ℓ had exceeded 1 before the peak, it quickly plunges below 1. Exactly ℓ time points after the peak, the lagged estimator attains the smallest possible value of A_t^ℓ, as $X_{t-\ell}$ maximizes its denominator. Again, the lagged ratio is likely to have larger fluctuations of A_t^ℓ, since its denominator reaches extremes that are not witnessed in A_t^γ (the convolution in the denominator of A_t^γ acts as a smoother).

In Figure 2, we can see A_t^ℓ drop below 0.8 near the start of February 2022; this happens precisely $\ell = 16$ days after daily new hospitalizations peak above 20,000 in mid-January. The lagged ratio falls from 20% to 12.5% accordingly, with the true HFR remains roughly constant, hovering around 15%. In the same period, the convolutional ratios (with light- or heavy-tailed γ) stay quite close to the true HFR.

Primary incidence levels out from a fall. The most jarring instance of misspecification bias occurs as primary incidence levels out. The true delay distribution π has a heavier tail than the low-mean, light-tailed distribution γ. It also has a heavier tail than the point mass distribution, which has no tail at all. This has important implications as the peak of the surge fades into the past. Compared to π, the light-tailed γ and point mass distribution convolve little to no mass with the high-count period of the wave. As a result, both A_t^γ and A_t^ℓ rise above 1, and severity rate estimates spike. The magnitude of this spike depends how quickly primary incidence is changing.

Figure 2 displays this false spike. Around the start of April 2022, we see A_t^γ (for light-tailed γ) and A_t^ℓ reach maximums near 1.25 and 1.68, respectively. Their corresponding HFR estimates reach 18% and 25% while the true HFR has fallen below 14%. This is of course highly problematic as it signals a rise in severity at a very counterintuitive time, when hospitalizations are at their lowest. The heavy-tailed delay γ has the opposite trend and underestimates the true HFR at this time, but by a smaller amount.

2.4 Experimental setup

Here we describe the data and general experimental setup used in Figures 1 and 2, and in Section 3.

Hospitalization-fatality rate. Our experiments analyze the HFR throughout the COVID-19 pandemic. While HFR may be less common as an object of study compared to CFR, it has a few advantages. First and foremost, hospitalization reporting was much more complete than case reporting throughout the pandemic. Hospitals were mandated to report new daily COVID-19 admissions to the Department of Health and Human Services (HHS) (Department of Health and Human Services, 2023). Due to changes in case ascertainment over time (cases as a fraction of infections), it is harder to interpret the CFR in a time-varying fashion, i.e., harder to understand what precisely this is reflecting over the course of the pandemic.

A second advantage is that hospitalization counts published by the HHS are aligned by admission date. This makes it more meaningful to interpret the HFR as a reflection of severity, especially as a time-varying quantity. In comparison, case counts as aggregated by John Hopkins University (JHU) (Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, 2023) (the central resource for comprehensive COVID-19 case data in the US) are aligned by report date. Extreme reporting delays (sometimes cases were reported 45 days after infections, see, e.g., Jahja et al., 2022) make the CFR less meaningful to study as a time-varying quantity, even outside from ascertainment issues.

Lastly, we were able to find a good "ground truth proxy" for the national HFR during the COVID-19 pandemic, as published by the National Hospital Care Survey (NHCS). This helps guide our simulations and also serves as validation data for us, as we describe in more detail below.

Aggregate data streams. To estimate the real-time HFR, we use aggregate counts of daily COVID-19 hospitalizations and deaths as made available in the Epidata API (Farrow et al., 2015), developed by the Delphi Group. Like HHS for hospitalizations, the JHU Center for Systems Science and Engineering (CSSE) provided the definitive resource for real-time death counts during the pandemic. These counts reflect times at which deaths were reported to health authorities, not necessarily when they actually happened. Hence raw JHU death counts are highly volatile due to reporting idiosyncrasies like day-of-week effects and data dumps. Hospitalizations are also subject to strong day-of-week effects. We therefore smoothed all data with a 7-day trailing average, for both hospitalizations and deaths.

Our real-time estimates of HFR actually use data that was available two days after the date in question. This was done to account for a typical two-day latency in the most recent data available. In this sense, one can actually view our real-time estimates as a two-day backcast of the HFR. In the rare event that counts were still unavailable at a two-day lag, we imputed their values with the most recently observed data (this is a common scheme, called last-observation-carried-forward or LOCF).

Hyperparameters. The ratio estimators of the HFR require choices of the lag ℓ and delay distribution γ. The experiments in Section 3 use a lag of $\ell = 20$ days, which roughly maximizes the cross-correlation between hospitalizations and deaths over the entire pandemic. For γ, we use a discrete gamma distribution, and set its support length to be $d = 75$ days, a conservative choice. For its mean, we use 20 again; this agrees nicely with a UK study which finds a median hospitalization-to-death time of 11 days (Ward and Johnsen, 2021),[1] and a CDC report that 63% of COVID-19 deaths are reported within 10 days (National Center for Health Statistics (NCHS) at Centers for Disease Control and Prevention, 2023). 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. Appendix C evaluates the robustness of findings against different hyperparameter values.

Validation data. While the true HFR curve is unknown, there are sound ways to approximate it; one way is to use estimates from the National Hospital Care Survey (NHCS) (National Hospital Care Survey (NHCS) at Centers for Disease Control and Prevention, 2023), which records weekly HFRs based on a representative subset of 601 hospitals across the US. These estimates end up being consistently biased downwards because they are only based on deaths which occur in the hospital. A CDC analysis (Ahmad et al., 2023) found that roughly 60% of COVID-19 deaths occurred in hospitals in 2022, down from nearly 70% in 2021 and 2022. To account for non-inpatient deaths, we divide the NHCS estimates by these percentages. Lastly, we smooth the resulting HFR estimates with a spline, via the `smooth.spline` function in R (which chooses the smoothness hyperparameter by minimizing generalized cross-validation error). This results in our proxy for the ground truth HFR curve p_t.

This ground truth proxy is a useful benchmark to judge the fidelity of our HFR estimates; of course, it is not perfect and it is derived from a relatively small subset of hospitals. Results in Section 3 suggest it may be too high in late 2022. Appendix B discusses alternative approximations of ground truth HFR.

3 Results

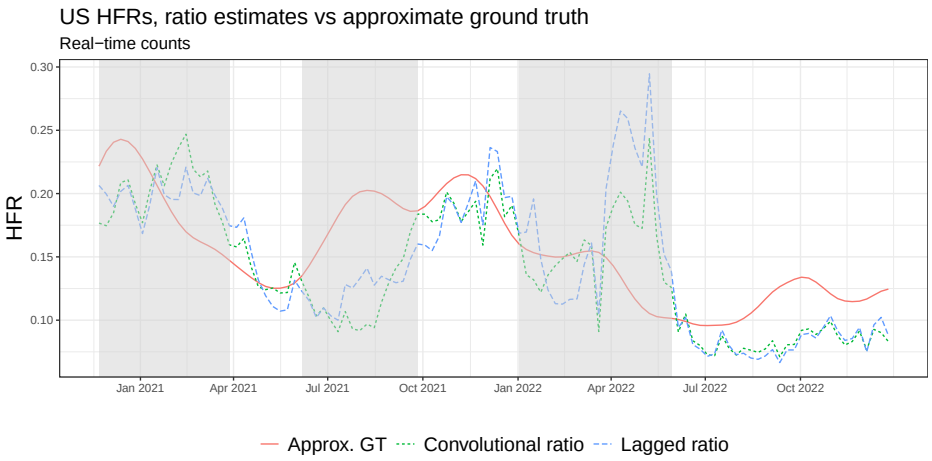
In this section, we explore the performance of the ratio estimators in greater depth. We analyze HFR estimates on real data in Section 3.1, and simulated data in Section 3.2. Throughout both, we continue to use aggregate hospitalization counts from HHS throughout the COVID-19 pandemic. The code to reproduce all results in this paper is available at `https://github.com/jeremy-goldwasser/Severity-Bias`.

3.1 COVID-19 data

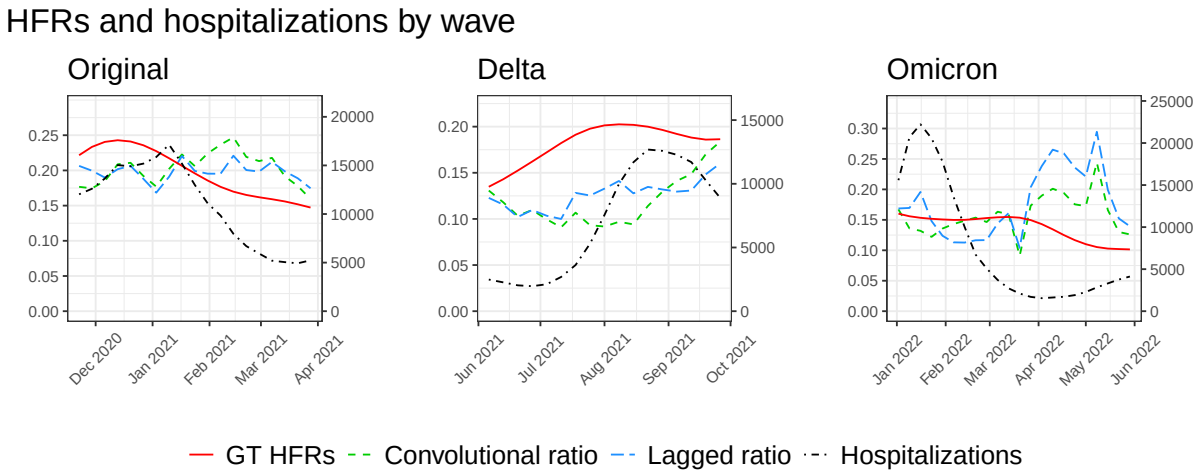
Figure 3 displays real-time HFR estimates on the data described in Section 2.4. We computed HFRs from November 2020 to December 2022, spanning the major COVID variants. The real-time hospitalization and death counts exhibit a fair degree of instability, so we preprocessed them with seven-day smoothing. Even then, the basic HFRs (4) and (2) still had a few days with wild spikes. To smooth these artifacts away, we further applied a seven-day trailing window to the ratio estimators, as in (10) and (11). Figure 9 in Appendix C examines how trailing window length affects HFR estimates.

Overall, both ratio estimators perform very poorly. As described in Propositions 1 and 2, the bias is consistent and nontrivial, especially for the lagged estimator. Both the lagged and convolutional ratios respond very slowly to changes in the HFR. As the HFR declines following the wave in winter 2021, both ratios remain near 0.2 for several months. More troublingly, they are very slow to detect the rising HFR in the early Delta period (summer 2021). If the purpose of these estimators is to inform stakeholders of increased risks in real time, they failed during the Delta surge.

[1] Of course, conditions in the UK may be quite different from the US. However, we rely on the UK study because it provides the most comprehensive information on COVID-19 hospitalization-to-death delay distributions.



(a) Comparing convolutional and lagged ratios against approximate ground truth.



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

Figure 3: HFR estimates from real-time aggregates, Nov. 2020 — Dec. 2022. Biased periods of major waves are highlighted.

The most significant bias comes in the middle of the Omicron wave in spring 2022. In this period, the HFR remains around 15% until April, then sharply declines to 9% two months later. The lagged ratios first fluctuate above and below the true HFRs. Subsequently, both estimates surge as the true HFR nears its nadir, with the lagged ratio nearing 30%. This dramatic upswing signals a serious false alarm. The analysis in Sections 2.2 and 2.3 explain each of these failure cases.

Well-specified analysis. We start by analyzing the convolutional ratio with respect to the well-specified bias expression in Proposition 1. While this expression assumes that the true delay distribution is known, we found that different choices of delay distribution generally yield similar bias (Appendix C). This indicates that our estimates may not be far from the oracle ratio.

Proposition 1 indicates that the bias moves in the opposite direction of the true severity rate. This occurs during the Delta wave, when the HFRs rise well before the ratio estimates do. Conversely, falling HFRs produce positive bias, as observed in the original and Omicron waves.

The enormity of the bias during Omicron can partially 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. 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.

Misspecified analysis. The misspecification analysis explains central discrepancies between the convolutional and lagged ratios. Section 2.3 discusses how $A_t^\ell < 1$ in the initial weeks of declining primary incidence. As a result, the lagged ratio should incur negative misspecification error. We observe this when hospitalizations with the original variant fall from their peak in mid-January 2021. Throughout February, lagged estimates are about 2% below the positively biased convolutional ratios.

When primary incidence rises, $A_t^\ell > 1$, contributing positive bias relative to the oracle convolutional ratio. Correspondingly, as hospitalizations surge due to the Delta variant in August 2021, the lagged ratio is less negatively biased.

Lastly, we explained that $A_t^\ell > 1$ after a fall in primary incidence. This accounts for the lagged ratio having higher bias in April 2022, when hospitalizations level out from the Omicron surge. Figure 2 visualizes this same period, where the true HFRs and delay distribution are known. There, A_t^ℓ reaches 1.5, by far its maximum — validating the magnitude of the difference between the convolutional and lagged ratios.

We performed several robustness checks to assess the stability of these findings. Appendix C explores the effect of different hyperparameters and locations. By and large, the ratio estimators yield roughly the same bias regardless of these considerations. It also compares HFR estimates using finalized counts, rather than the data available in real time. This exploration finds that the observed biases could not be attributed to real-time reporting issues.

3.2 Simulated data

We further evaluated these methods in a variety of simulation settings. Given a series of time-varying HFRs p_t and delay distribution π, deaths are defined without noise from (3):

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

Like the experiments in Sections 2.2 and 2.3, we used real-time HHS hospitalization counts. The simulations in this section evaluate performance over a two-year period, with a broader range of underlying HFR curves. To supplement the NHCS HFRs, we mimicked the opposite trend by inverting and rescaling them. We also modeled a stationary HFR of 10% over all time. As in Section 2.4, the delay distributions were again gamma with standard deviation 0.9 of their mean. We experimented with means of 12 and 24 to illustrate a short and long delay distribution.

To elucidate the oracle bias in Proposition 1, we let the convolutional ratio use the true delay distribution. For the lagged ratio, ℓ was again chosen to maximize the cross-correlation between hospitalizations and

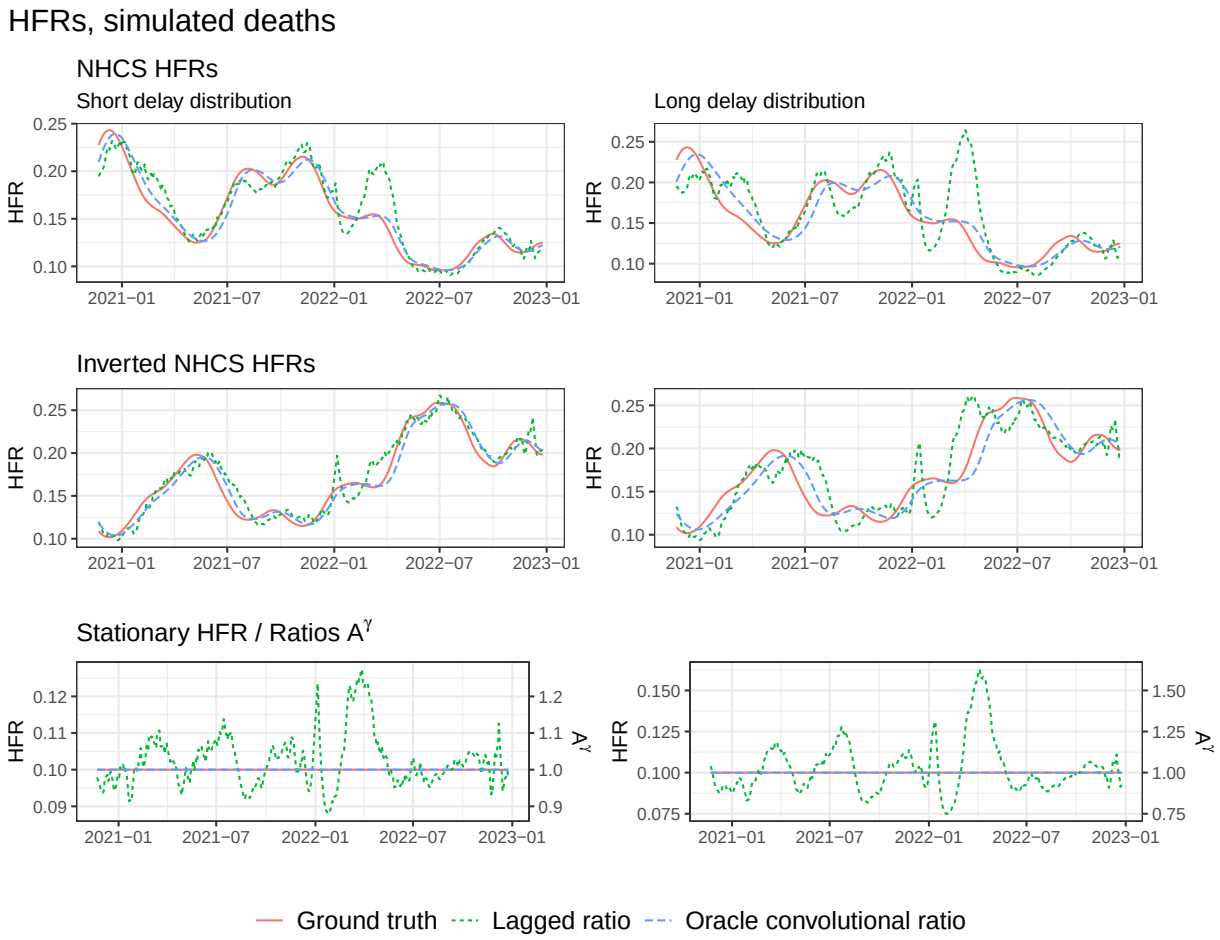


Figure 4: True and Estimated HFRs from Simulated Deaths. First column has short delay distribution, second has long.

deaths. We also experimented with the mean of the delay distribution, as advocated by Feng et al. (2023). Figure 13 in Appendix D shows the results are similarly unstable.

Figure 4 displays the results on the six settings of delay distribution and HFR. Matching expectations, HFRs are significantly more biased given the longer underlying delay distribution. In all three HFR settings, the lagged ratios swing more widely. For example, when the true HFR is a constant 10%, they peak at 12% under the light-tailed distribution, compared to 15% with the heavy tail. The oracle convolutional ratio does not share the lagged estimator's dramatic oscillations. Rather, it tracks the general shape of true curve, albeit at a delay. For the NHCS HFRs, the average delay was 5 days for the light-tailed distribution, and 12 days with heavy tail; these delays were 6 and 14 days for the inverted curve. (To compute the average delay, we again took the maximal cross-correlation between the two series.)

The analysis in Section 2.3 accounts for the wide gap in performance between the two estimators. Proposition 2 expresses bias under misspecification as a function of the ratio A_t^γ. These ratios A_t^γ are visualized in Figure 4 as rescaled HFR estimates in the stationary case. Given a constant rate p, the oracle convolutional ratio is unbiased, so Proposition 2 reduces to $\mathbb{E}[\widehat{p}_t^\gamma] = pA_t^\gamma$. Furthermore, $\mathbb{E}[\widehat{p}_t^\gamma] = \widehat{p}_t^\gamma$, as our setup simulates deaths without noise. Consequently, $A_t^\gamma = \frac{\widehat{p}_t^\gamma}{p}$.

In expectation, the lagged ratio is higher than the oracle convolutional ratio when $A_t^\gamma > 1$, and lower when $A_t^\gamma < 1$. Comparing the A_t^γ curves to the estimated HFRs, the bias moves very similarly. For example, during the Delta and Omicron waves, rapid rises in hospitalizations produced high values of A_t^ℓ. This accounts for the spikes in August 2021 and January 2022. When hospitalizations level out from the Omicron surge, A_t^ℓ

spikes to 1.2 and 1.5 for the short and long distributions — hence the positive bias in spring 2022. Lastly, the lagged estimator should have negative bias as primary events fall. We observe this in Delta (September 2021) and Omicron (Februrary 2022).

The misspecified bias (Proposition 2) rescales the oracle bias and adds a misspecification term. Studying Figure 4, we observe the misspecification term tends to dominate when A_t^γ strays away from 1. To understand this, consider periods in which the oracle bias is negative. As introduced in Section 2.3, the oracle and misspecification terms are at odds with each other when this is the case.

Invariably, the lagged ratio moves in the direction of the misspecification term $p_t(A_t^\gamma - 1)$. Under the true NHCS HFRs, for example, the lagged estimates spike with A_t^ℓ in August 2021. In the inverted setting, the lagged bias tracks the down-up-down motion of A_t^ℓ during the first five months of 2021. That the misspecification term wins out in these conflicting settings indicates it comprises a disproportionate amount of the bias. Indeed, the oracle bias is generally low enough that multiplicative rescaling may not have a large effect.

For a further example, consider the bias at April 2022 on the middle right. The true HFR is 14%, with the convolutional ratio nearby at 15%. Meanwhile, the lagged ratio peaks at 22.5%, driven upwards by an A_t^ℓ of 1.5. Decomposing the lagged bias of 8.5% with Proposition 2, the oracle term $A_t^\ell \text{Bias}(\widehat{p}_t^\pi)$ equals only 1.5%; meanwhile, the misspecification term $p_t(A_t^\ell - 1) = 7\%$, accounting for the majority of the bias.

4 Discussion

Our analyses illustrate that practitioners should take caution when using time-varying severity ratio estimators. They exhibit considerable bias when severity rates change, particularly the popular lagged ratio estimator. 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.

Analyzing the lagged ratio enables us to make real-time heuristics about its performance in practice. Proposition 2 decomposes its bias into oracle and misspecification terms, the latter of which has been shown to dominate. Based solely on the primary incidence curve, we can expect the lagged ratio to make the following errors:

1. Unreasonably high severity estimates when primary incidence is rising quickly;
2. Rapid declines when primary incidence is falling quickly;
3. Unexpected surges when primary incidence has leveled out after falling.

Practitioners can adjust their reactions accordingly when these bias patterns occur in real time. For example, if the lagged HFR spikes shortly after hospitalizations reach a stable low, a savvy epidemiologist can temper her alarm with the knowledge it may well be spurious.

While the lagged ratio is ubiquitous in practice, our analysis of its drawbacks suggests other aggregate estimators should be favored. Figure 4 showed the oracle convolutional ratio is much more accurate. It still outperformed the lagged ratio given a misspecified delay distribution, though its bias was also large (Figures 2 and 3).

Qu et al. (2022) proposed an approach that differs considerably from the ratios discussed in this paper. The method estimates all historical severity rates at once, using the relation in (3) to fit a fused lasso model. This estimator is inherently forward-looking, where rates at t are exclusively used to produce secondary events after t. Given regularization parameter λ, current time T, and start time t_0, the fused lasso estimates

$$\widehat{p}^{\text{FL}} = \text{argmin}_{p \geq 0} \sum_{t=t_0}^{T} (Y_t - \sum_{j=0}^{d} X_{t-j} \gamma_j p_{t-j})^2 + \lambda \sum_{t=t_0-d+1}^{T} |p_t - p_{t-1}|.$$

This estimator is not succumb to the issues of the backward-looking ratios. 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. Thorough investigation of its performance is a promising object for future study.

Future work could generalize the above approach beyond piecewise constants. The fused lasso is a special case of trend filtering, a nonparametric regression technique that fits piecewise polynomials Tibshirani (2014). Higher-order curves may better model trends and improve performance. (Jahja et al., 2022) applies trend filtering to a similar deconvolution problem, reconstructing latent infections from case reports. Its insights on tail regularization may be useful to stabilize severity estimates.

Overton et al. (2022) 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 (Figures 4 and 6). 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.

Severity rates may be biased in ways beyond the statistical bias our work focuses on. In Section 2.4, we mentioned that HFR estimation from aggregates is subject to survivorship bias — the failure to account for deaths that occurred outside the hospital (Lipsitch et al., 2015). Under-reporting is another central challenge, particularly for CFR. Not all infections are reported, reporting rates change across time, and severe cases are more likely to be reported than mild cases. Reich et al. (2012) proposed 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). Angelopoulos et al. (2020) applied this in the context of COVID-19, analyzing how the chosen delay distribution affects its results. The work also identifies other sources of bias, like differences in case definition and testing eligibility.

As discussed in Section 2.1, severity rates may be understood in connection with reproduction numbers. This connection extends to their bias as well. For example, we demonstrated that the convolutional ratio is unbiased if the severity rate and delay distribution in the d days before t are stationary. In a similar vein, Fraser (2007) noted that instantaneous R_t is equal to case R_t if conditions remain unchanged. Future work along the lines of Eales and Riley (2023) could apply our analytical framework to R_t bias, examining the fidelity of instantaneous R_t as a proxy for case R_t.

References

Adjei, S., Hong, K., Molinari, N.-A. M., Bull-Otterson, L., Ajani, U. A., Gundlapalli, A. V., Harris, A. M., Hsu, J., Kadri, S. S., Starnes, J., Yeoman, K., and Boehmer, T. 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. *Morbidity and Mortality Weekly Report*, 71(37):1182–1189.

Ahmad, F. B., Cisewski, J. A., Xu, J., and Anderson, R. N. (2023). COVID-19 mortality update — United States, 2022. *Morbidity and Mortality Weekly Report*, 72(18):493–496.

Angelopoulos, A. N., Pathak, R., Varma, R., and Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harvard Data Science Review*, Special Issue 1.

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

Bellan, M., Patti, G., Hayden, E., Azzolina, D., Pirisi, M., et al. (2020). Fatality rate and predictors of mortality in an Italian cohort of hospitalized COVID-19 patients. *Scientific Reports*, 10:20731.

Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (2023). COVID-19 data repository. `https://github.com/CSSEGISandData/COVID-19`.

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. *British Medical Journal*, 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. arXiv:2405.08841.

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. *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. https://www.hhs.gov/sites/default/files/covid-19-faqs-hospitals-hospital-laboratory-acute-care-facility-data-reporting.pdf.

Eales, O. and Riley, S. (2023). Differences between the true reproduction number and the apparent reproduction number of an epidemic time series. arXiv:2307.03415.

Farrow, D. C., Brooks, L. C., Tibshirani, R. J., and Rosenfield, R. (2015). Delphi Epidata API. https://github.com/cmu-delphi/delphi-epidata.

Feng, J., Luo, H., Wu, Y., Zhou, Q., and Qi, R. (2023). A new method for accurate calculation of case fatality rates during a pandemic: Mathematical deduction based on population-level big data. *Infectious Medicine*, 2(2):96–104.

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. *British Medical Journal*, 339:b2840.

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.

Gupte, P. R., Kucharski, A. J., Russell, T. W., Lambert, J. W., Gruson, H., Taylor, T., Azam, J. M., Degoot, A. M., and Funk, S. (2024). cfr: Estimate disease severity and case ascertainment. https://cran.r-project.org/package=cfr.

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*, 95(1):e28231.

Jahja, M., Chin, A., and Tibshirani, R. J. (2022). Real-time estimation of covid-19 infections: Deconvolution and sensor fusion. *Statistical Science*, 37(2):207–228.

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). *Statistics in Medicine*, 26(9):1982–1998.

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) at Centers for Disease Control and Prevention (2023). Deaths by select demographic and geographic characteristics. https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm.

National Hospital Care Survey (NHCS) at Centers for Disease Control and Prevention (2023). In-hospital mortality among hospital confirmed COVID-19 encounters by week from selected hospitals. https://www.cdc.gov/nchs/covid19/nhcs/hospital-mortality-by-week.htm.

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. E., 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 Computational Biolology*, 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. *Scientific Reports*, 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.

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.

Russell, T. W., Golding, N., Hellewell, J., Abbott, S., Wright, L., Pearson, C. A. B., van Zandvoort, K., Jarvis, C. I., Gibbs, H., Liu, Y., Eggo, R. M., Edmunds, W. J., Kucharski, A. J., and CMMID COVID-19 working group (2020a). Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Medicine*, 18(332).

Russell, T. W., Hellewell, J., Jarvis, C. I., van Zandvoort, K., Abbott, S., Ratnayake, R., CMMID COVID-19 working group, 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):2000256.

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

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323.

Unnikrishnan, J., Mangalathu, S., and Kutty, R. V. (2021). Estimating under-reporting of COVID-19 cases in Indian states: An approach using a delay-adjusted case fatality ratio. *BMJ Open*, 11(1):e042584.

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. *Journal of the American Medical Association*, 331(22):1963–1965.

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

A Proofs and further analysis

A.1 Proof of Proposition 1

Here and henceforth, we abbreviate $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \{X_s\}_{s \le t}]$. Observe that

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

The well-specified bias can be understood as a weighted average of $\{p_{t-k} - p_t\}_{k=0}^{d}$. The attainable absolute bias ranges between $\min_{k=0,\dots,d} |p_{t-k} - p_t| = 0$, achieved by $k = 0$, and $\max_{k=0,\dots,d} |p_{t-k} - p_t|$. This maximal bias is achieved by setting one of the weights $X_{t-k}\pi_k / (\sum_{j=0}^{d} X_{t-j}\pi_j)$ to 1 and the rest to zero, either through the delay distribution π or through the primary incidence curve X. Hence, the explanations for delay distribution and primary incidence are aligned: They inflate the bias by upweighting distant timepoints for which the severity rate was different. If severity rates are monotonically changing, for example, then the maximal bias occurs at $k = d$.

A.2 Proof of Proposition 2

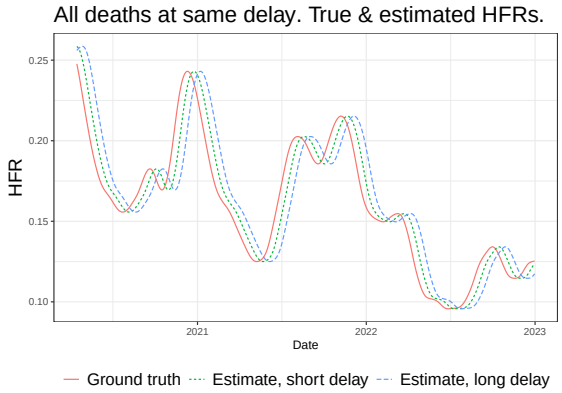
Observe that

$$\begin{aligned}
\text{Bias}(\widehat{p}_t^{\gamma}) &= \frac{\mathbb{E}_t[Y_t]}{\sum_{k=0}^{d} X_{t-k}\gamma_k} - p_t \\
&= \frac{\sum_{k=0}^{d} X_{t-k}\pi_k p_{t-k}}{\sum_{k=0}^{d} X_{t-k}\gamma_k} - \frac{\sum_{k=0}^{d} X_{t-k}\gamma_k p_t}{\sum_{k=0}^{d} X_{t-k}\gamma_k} \\
&= \sum_{k=0}^{d} \frac{X_{t-k}}{\sum_{j=0}^{d} X_{t-j}\gamma_j}(\pi_k p_{t-k} - \gamma_k p_t) \\
&= \sum_{k=0}^{d} \frac{X_{t-k}}{\sum_{j=0}^{d} X_{t-j}\gamma_j}(\pi_k p_{t-k} - (\pi_k + (\gamma_k - \pi_k))p_t) \\
&= \frac{\sum_{j=0}^{d} X_{t-j}\pi_j}{\sum_{j=0}^{d} X_{t-j}\gamma_j} \sum_{k=0}^{d} \frac{X_{t-k}\pi_k}{\sum_{j=0}^{d} X_{t-j}\pi_j}(p_{t-k} - p_t) - \\
&\qquad p_t \sum_{k=0}^{d} \frac{X_{t-k}}{\sum_{j=0}^{d} X_{t-j}\gamma_j}(\gamma_k - \pi_k) \\
&= \frac{\sum_{j=0}^{d} X_{t-j}\pi_j}{\sum_{j=0}^{d} X_{t-j}\gamma_j} \text{Bias}(\widehat{p}_t^{\pi}) + p_t \Big[\frac{\sum_{k=0}^{d} X_{t-k}\pi_k}{\sum_{j=0}^{d} X_{t-j}\gamma_j} - 1 \Big]
\end{aligned}$$

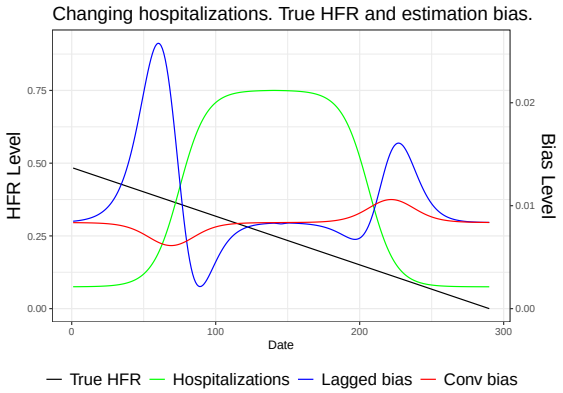
A.3 Further analysis of (6)

In this section, we present examples that further explain the well-specified bias. These are more contrived that the ones in Section 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, $\pi_k = \mathbb{1}\{k = \ell\}$,



(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 5: Toy examples of biased severity rates.

so the convolutional and lagged ratios are both $\widehat{p}_t = \frac{X_{t-\ell}p_{t-\ell}}{X_{t-\ell}} = p_{t-\ell}$ presuming both have access to the oracle delay distribution. Figure 5a 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 5a). 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 Proposition 1:

$$\text{Bias}(\widehat{p}_t^{\gamma}) = \text{Bias}(\widehat{p}_t^{\ell}) = \Big(\sum_{k=0}^{d} \pi_k p_{t-k}\Big) - p_t.$$

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 d days. Further assume $p_{t-d} \neq p_t$, so there is some degree of bias. Then

$$\begin{aligned} |\text{Bias}(\widehat{p}_t^{\gamma})| &= \frac{\frac{1}{2}\big|X_t(p_t - p_t) + X_{t-d}(p_{t-d} - p_t)\big|}{\frac{1}{2}(X_t + X_{t-d})} \\ &= \frac{X_{t-d}|p_{t-d} - p_t|}{X_{t-d}(1 + \frac{X_t}{X_{t-d}})} = \frac{|p_{t-d} - p_t|}{1 + \frac{X_t}{X_{t-d}}} \end{aligned}$$

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

Figure 5b 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.

The figure also plots with lagged ratio with $\ell = \frac{d}{2}$, the mean of the delay distribution. When daily hospitalizations are close to constant, the two estimators converge towards the same ratio. During periods of change, however, the lagged estimator has different bias. It first moves upwards — the opposite direction as

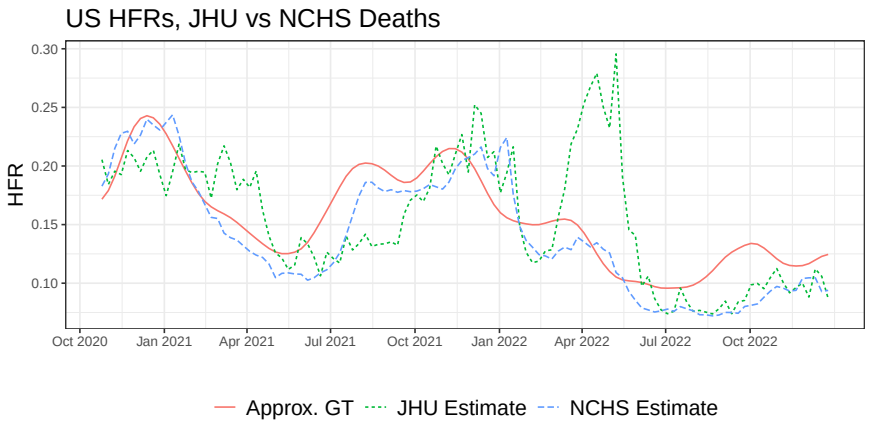


Figure 6: Real-time lagged ratios, JHU vs NCHS deaths. Seven-day smoothing with 19- and 11-day lags, respectively.

the convolutional bias — with far greater magnitude. This can be explained by the ratio $A_t^\ell = \frac{X_{t-2\ell}+X_t}{2X_{t-\ell}}$ from Proposition 2. As hospitalizations begin to steeply rise, $X_{t-2\ell}$ and $X_{t-\ell}$ are similar, but $X_t > X_{t-\ell}$. Hence, $A_t^\ell > 1$, contributing positive bias to both the oracle and misspecification terms. As hospitalizations level out near the top, $A_t^\ell < 1$, hence the bias falling lower. The opposite pattern occurs as hospitalizations fall.

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 6 shows this minor change has a significant effect on the bias. It compares the real-time lagged ratios with deaths sourced from JHU and NCHS. JHU is much more biased during the variant periods discussed. For example, NCHS only rises from 12% to 14% as Omicron falls, far below JHU's surge above 25%. As analyzed in Section 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 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.

$$\widehat{p}_t^{\text{LaggedRetro}} = \frac{Y_{t+L}}{X_t} \tag{9}$$

The second approach computed a single HFR for each major variant, then mixing by the proportions of variants in circulation. Formally, let $\widehat{p}_j$ 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

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

Each variant's HFR $\widehat{p}_j$ 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

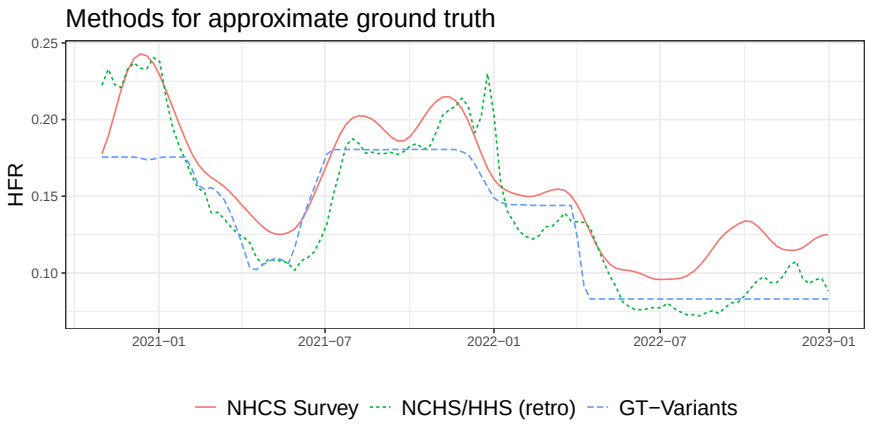


Figure 7: Methods for retrospective ground truth HFRs.

`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 et al., 2022).

Figure 7 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 Data source

The results in Section 3.1 use hospitalization and death counts available in real time. To investigate the sensitivity of our findings, we recomputed the lagged and convolutional ratios, this time using the finalized aggregates. Figure 8a shows the estimates with real-time and finalized counts track very closely to one another. Therefore, the observed bias in 3 cannot be attributed to reporting quirks.

The one period where the curves are significantly different from one another is in March 2022. While the HFRs from finalized counts steadily rise, the real-time estimates sharply fall then immediately bounce back. This sudden drop is due to a brief period in which reported death counts were suddenly too low (Fig. 8b. This is corrected in the finalized counts, hence their smooth HFRs. Removing this artifact further reinforces the bias trends described in Section 2.2.

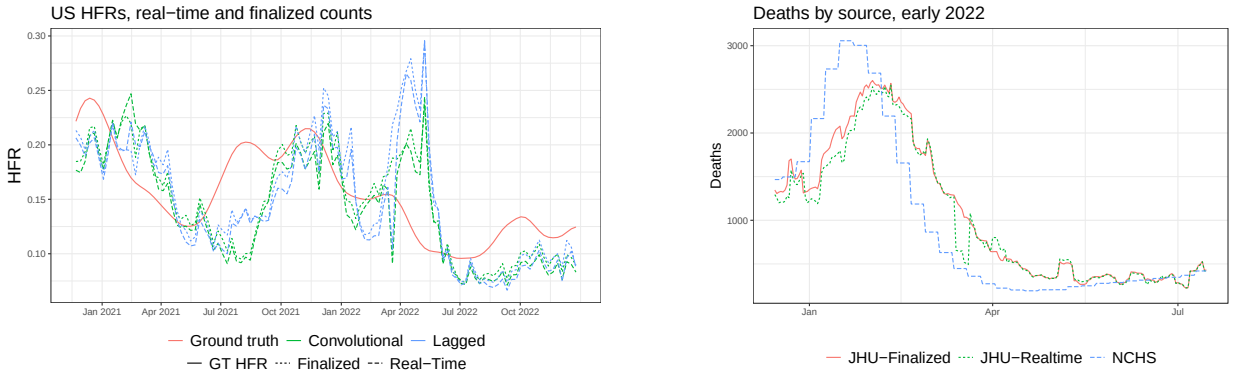
C.2 Hyperparameters

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 9 plots performance over choices of window size parameter. We analyze smoothed versions of the lagged estimator

$$\widehat{p}_t^{\ell,W} = \frac{\sum_{s=t-w+1}^{t} Y_s}{\sum_{s=t-w+1}^{t} X_{s-\ell}}, \tag{10}$$

as well as the convolutional estimator

$$\widehat{p}_t^{\gamma,W} = \frac{\sum_{s=t-w+1}^{t} Y_s}{\sum_{s=t-w+1}^{t} \sum_{k=0}^{d} X_{s-\ell-k}\gamma_k}. \tag{11}$$



(a) HFR estimates with real-time and finalized counts.

(b) Aggregate deaths by data source and report time.

Figure 8: Convolutional ratio estimates are biased regardless of which delay distribution is selected.

Figure 9: The length of the trailing window bears little impact on the findings.

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 10 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 in winter 2021 and rises less slowly during Delta. However, all are biased to varying degrees, most notably the huge spurious surge in spring 2022.

Figure 11 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 11a kept the standard deviation to 90% of the mean, per Ward and Johnsen (2021). We also evaluated with a more compact delay distribution in 11b.

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.3 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

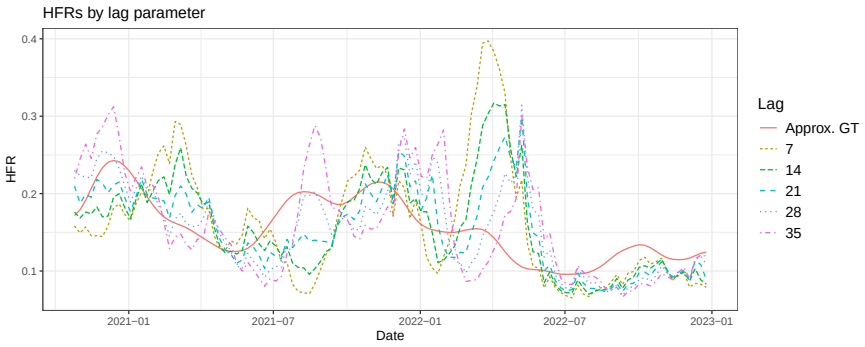
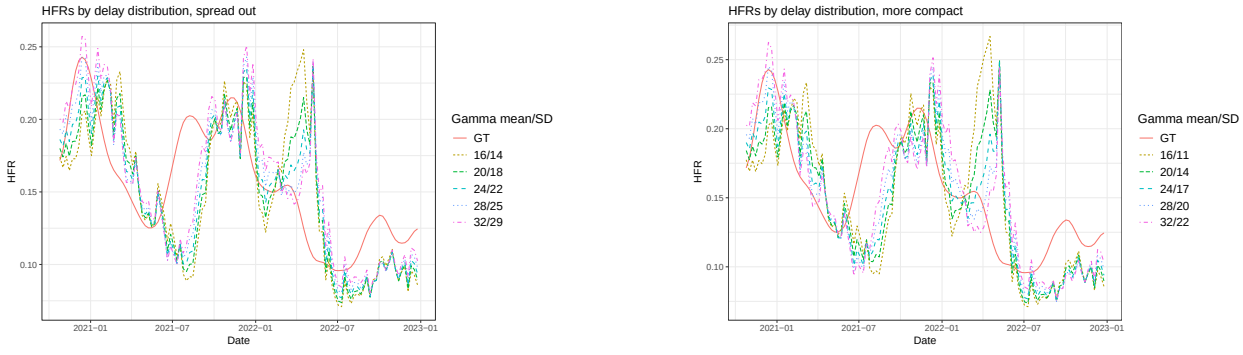


Figure 10: HFRs are biased regardless of what lag parameter is selected.



(a) SD is 0.9×mean.

(b) SD is 0.7×mean.

Figure 11: Convolutional ratio estimates are biased regardless of which delay distribution is selected.

poorly approximate the HFRs for individual states. A better state-level source is the retrospective lagged ratio (9) using NCHS deaths. Figure 12 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. 3a). Ratios in California, Texas, and Florida all are slow to detect the uptick in HFR during Delta; they also spike during Omicron in California, and to a lesser extent Florida. 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 17. While their HFRs are 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.

D Miscellaneous results

D.1 Alternative lag on simulation

Figure 13 presents the simulation results from Section 3.2 when the lag is the mean of the delay distribution. It clearly indicates the lagged ratio is not markedly better under this lag. This is in step with Figure 10, which demonstrated that its performance on real data are robust to choice of lag.

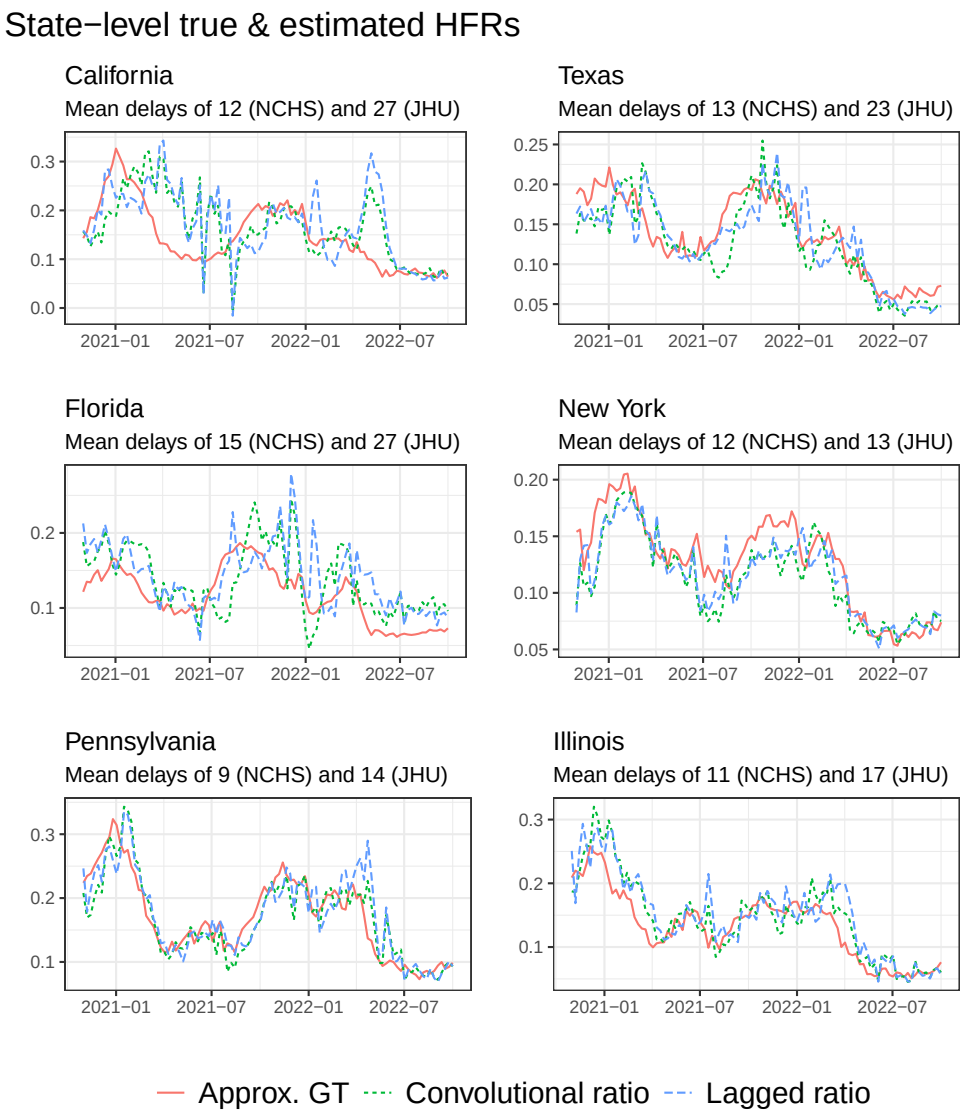


Figure 12: HFRs by individual states. Comparing retrospective estimates with NCHS against real-time estimates with JHU.

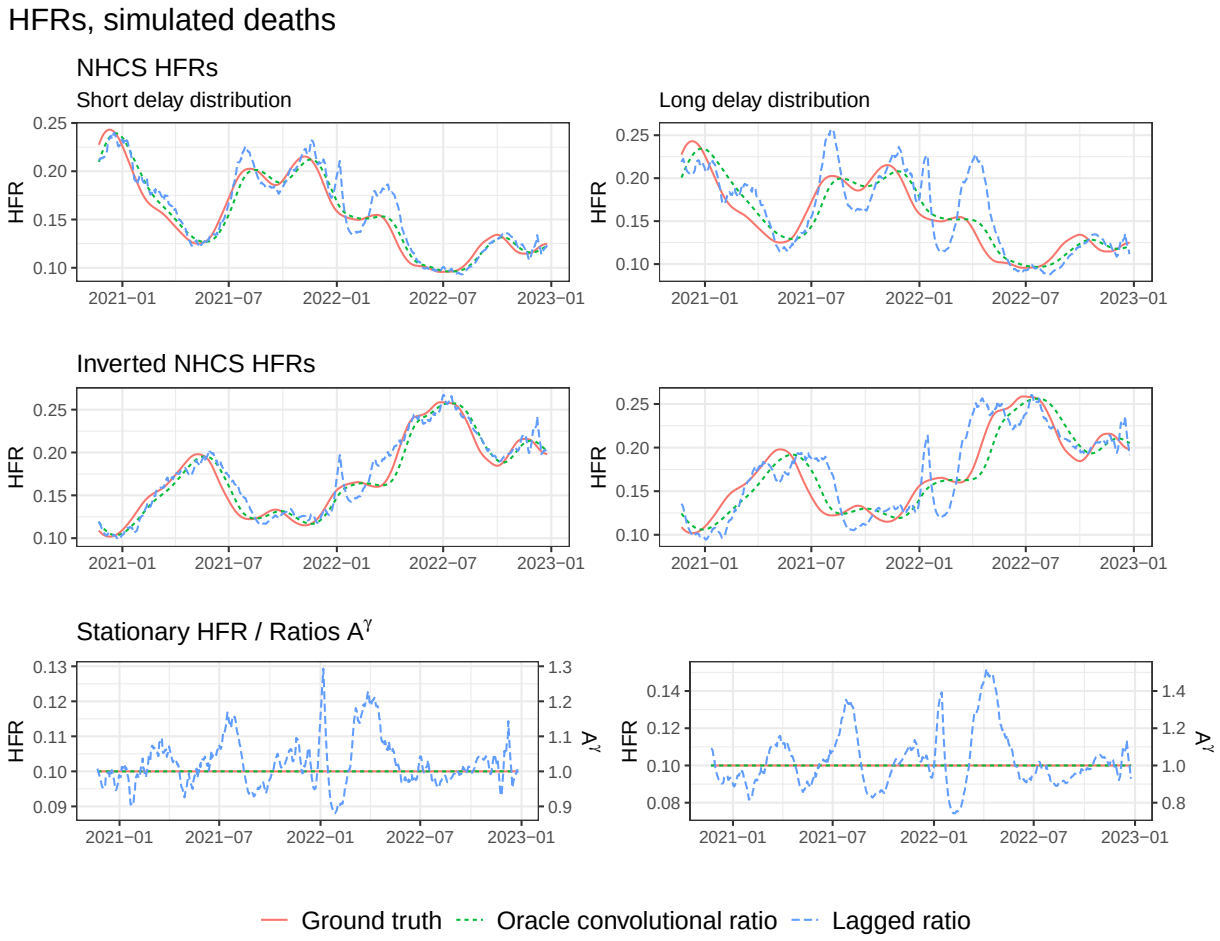


Figure 13: Lag chosen as mean of delay distribution. Same simulated data as Section 3.2.