Estimating Epidemic Severity Rates

Jeremy Goldwasser

Time-varying severity rates in epidemiology

- Severity rates express the probability that a primary event at time t will result in serious secondary event, e.g.
  - Case-fatality rate (CFR)
  - Hospitalization-fatality rate (HFR)
- Time-varying or stationary?
  - Most academic work on estimating severity rates assumes stationarity over time.
  - Severity rates constantly change due to new variants, therapeutics, etc.
  - Epidemiologists at the CDC use time-varying rates to analyze new risks.

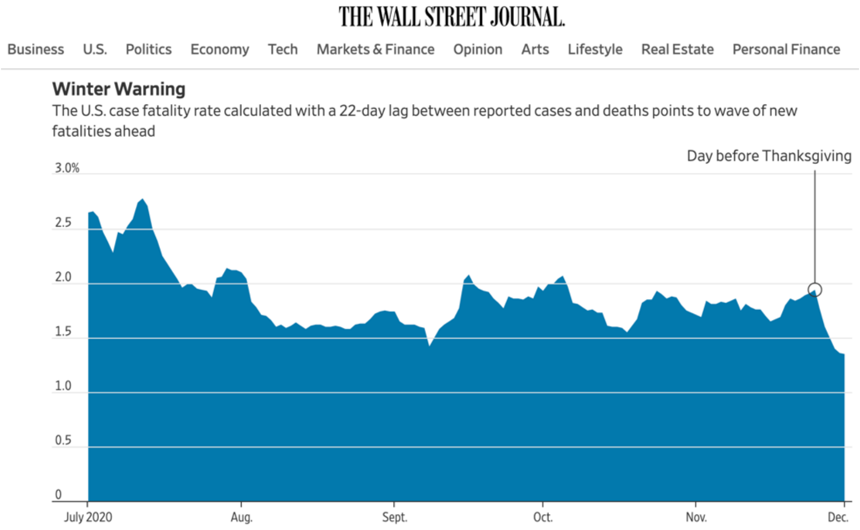

Newsletters The Atlantic Saved

HEALTH

How Many Americans Are About to Die?

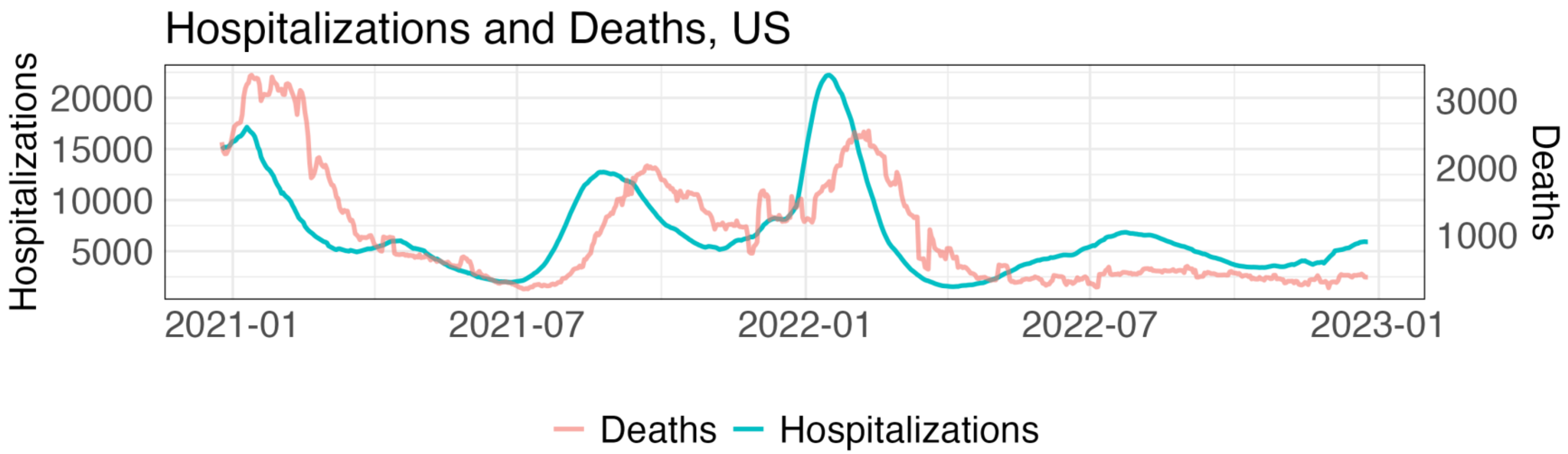
A new analysis shows that the country is on track to pass spring's grimmest record.

By Alexis C. Madrigal and Whet Moser



Often estimate severity from aggregate data

- Calculating severity rates is straightforward with a line list of patient outcomes.
  - CFR: Observe fraction of patients that tested positive at t who ultimately die.
- Maintaining such a line list may be unrealistic or impossible
  - In this case, severity rates must be estimated from aggregate count data.



Standard ratio estimators

- Most estimators for severity rates are simple ratios ("case fatality ratio") between secondary events and at-risk primary events
- The standard time-varying approach is a lagged ratio of aggregate counts:

$$\widehat{\mathrm{CFR}}_t = \frac{\text{Deaths at } t}{\text{Cases at } t - \ell}$$

- A more principled generalization uses the delay distribution:

$$\widehat{\mathrm{CFR}}_t = \frac{\text{Deaths at } t}{\sum_k \{\text{Cases at } t - k\} \times \hat{\mathbb{P}}(\text{Death is at } k \text{ days})}$$

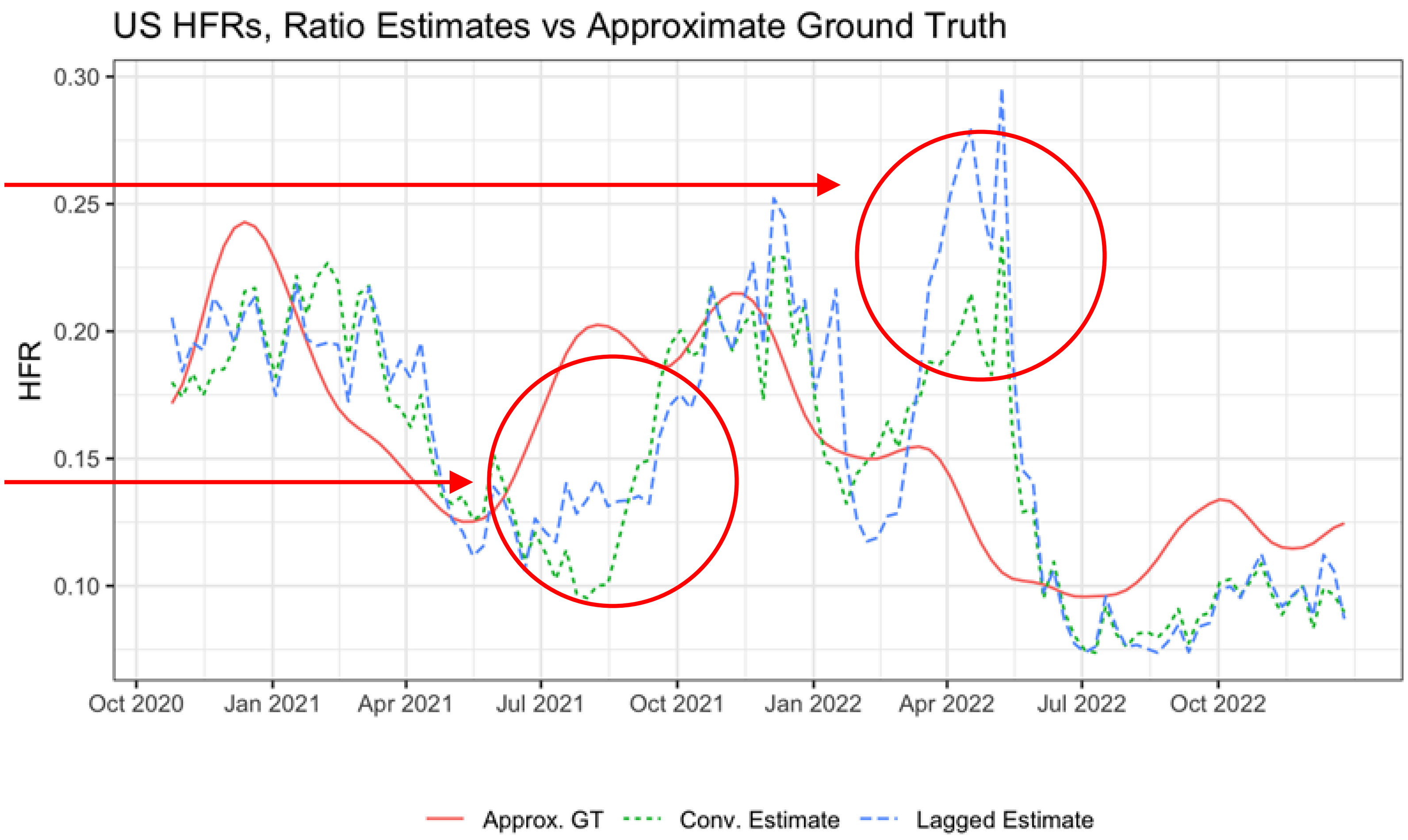
Our work: Understanding the bias of these ratios and proposing statistically sound alternatives.

Observed these ratios exhibit huge bias

Notable failures, HFR:

- Signaled enormous, nonexistent surge after Omicron peak – especially lagged ratio.
- Ignored higher risk as Delta took over

Findings robust across parameters, geography, etc.



Ingredients of Analysis: Data Streams

- Let X_t denote the primary incidence time series
- Let Y_t denote the secondary incidence time series.
    - We focus on HFR because there is decent ground truth data.
- In theory, they have the following relation:

$$Y_t | X_{s \leq t} = \sum_{k=0}^{d} \sum_{i=1}^{x_{t-k}} \mathbf{1}\{i^{\text{th}} \text{ case at } t-k \text{ died at } t\}$$

- In practice, real-world data may be messier due to e.g. day-of-week effects or data dumps.

Ingredients of Analysis: Statistical Model

- Given $Y_t | X_{s \le t} = \sum_{k=0}^{d} \sum_{i=1}^{x_{t-k}} \mathbf{1}\{i^{\text{th}} \text{ case at } t-k \text{ died at } t\}$
- Taking expectation reveals convolution of hospitalizations with delay distribution π and HFRs p:

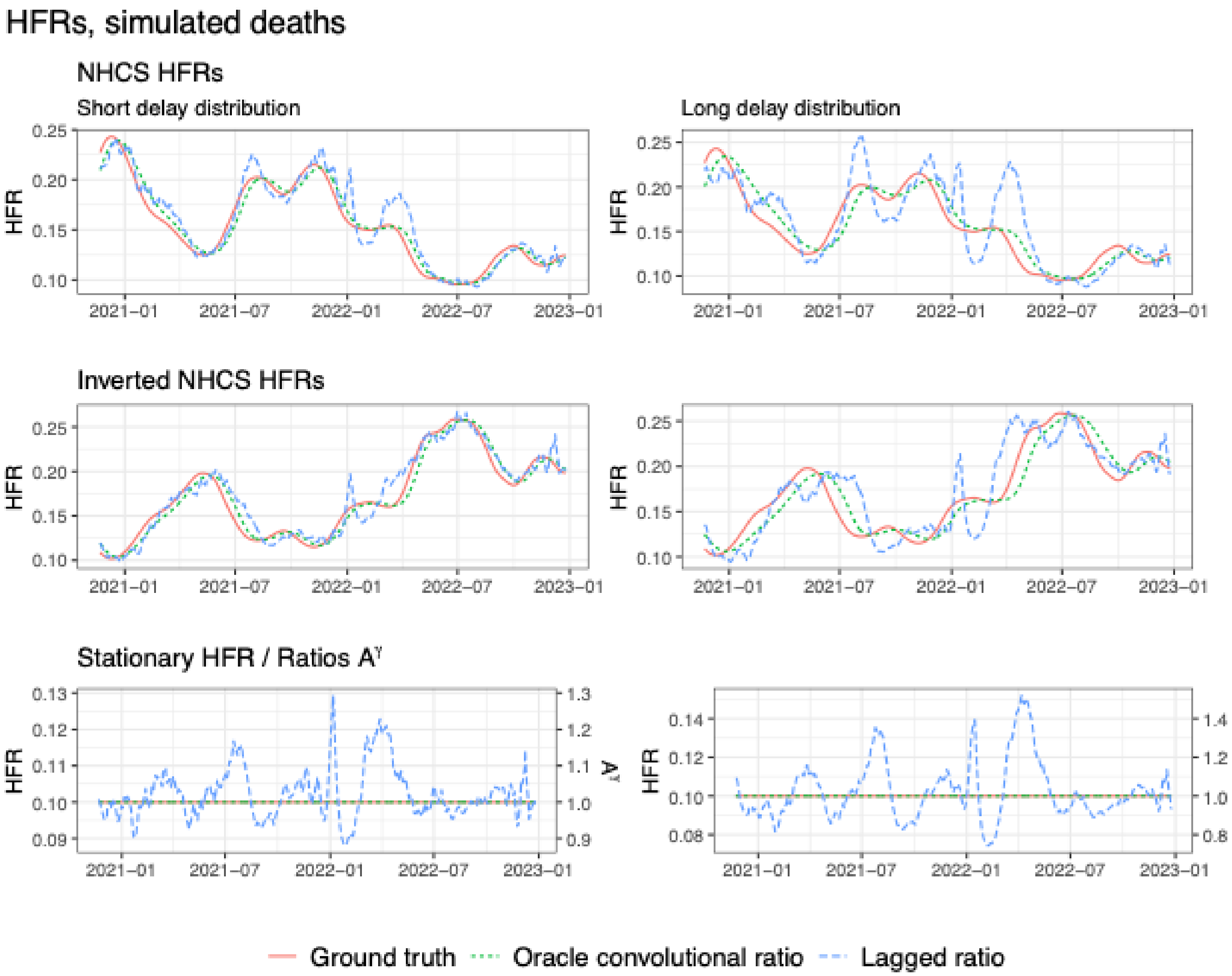
$$\begin{aligned}
\{\text{Deaths at } t\} &:= \sum_k \{\text{Hospitalizations at } t-k\} \\
&\qquad \times \mathbb{P}(\text{Die in } k \text{ days}) \\
&= \sum_k \{\text{Hospitalizations at } t-k\} \\
&\qquad \times \mathbb{P}(\text{Die in } k \text{ days} \mid \text{Die}) \\
&\qquad \times \mathbb{P}(\text{Die} \mid \text{Hospitalized at } t-k) \\
&= \sum_k X_{t-k} \pi_k p_{t-k}
\end{aligned}$$

Recreate bias on simulated data

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

$$\hat{p}_t^{\text{Conv}} = \frac{Y_t}{\sum_{k=0}^{d} X_{t-k}\pi_k}$$

- Noiseless simulation, so $Y_t = E[\ Y_t | X_{s \leq t}\ ]$ from the previous slide
- Even when hospitalizations are flat, the estimated HFR is up to 50% too high!



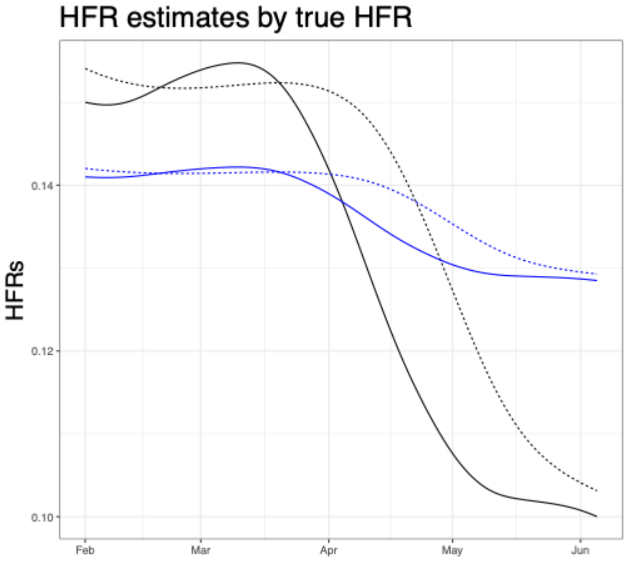
Well-specified analysis

For a stationary oracle delay distribution π,

$$\mathrm{Bias}(\hat{p}_t^{\pi}) = \sum_{k=0}^{d} \frac{\overset{B}{X_{t-k}}\overset{C}{\pi_k}}{\sum_{j=0}^{d} X_{t-j}\pi_j} \underbrace{(p_{t-k} - p_t)}_{A}.$$

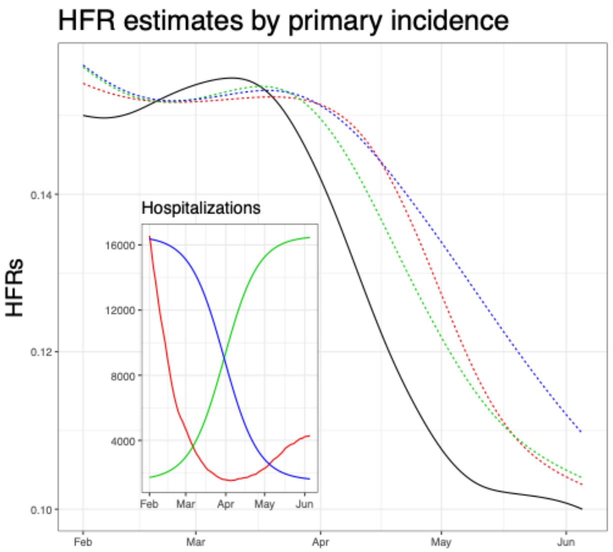
Bias of Convolutional Ratio with True Delay Distribution

A. Arises due to changing severity rates p
B. Affected by changing primary incidence X
   a. Usually falling → more bias
C. Exacerbated by heavy-tailed delay distr. π



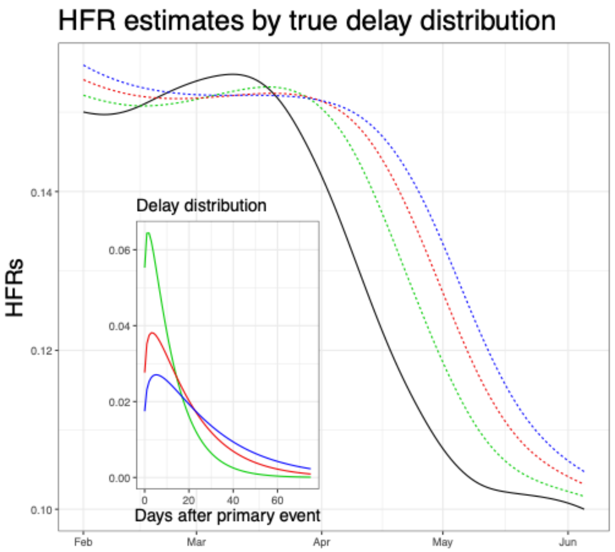
HFR — Ground truth ···· Estimated
Curve — True (NHCS) — Flatter

A



HFR — Ground truth ···· Estimated
Delay mean — True — Rising — Falling

B



HFR — Ground truth ···· Estimated
Delay mean — 12 — 20 — 28

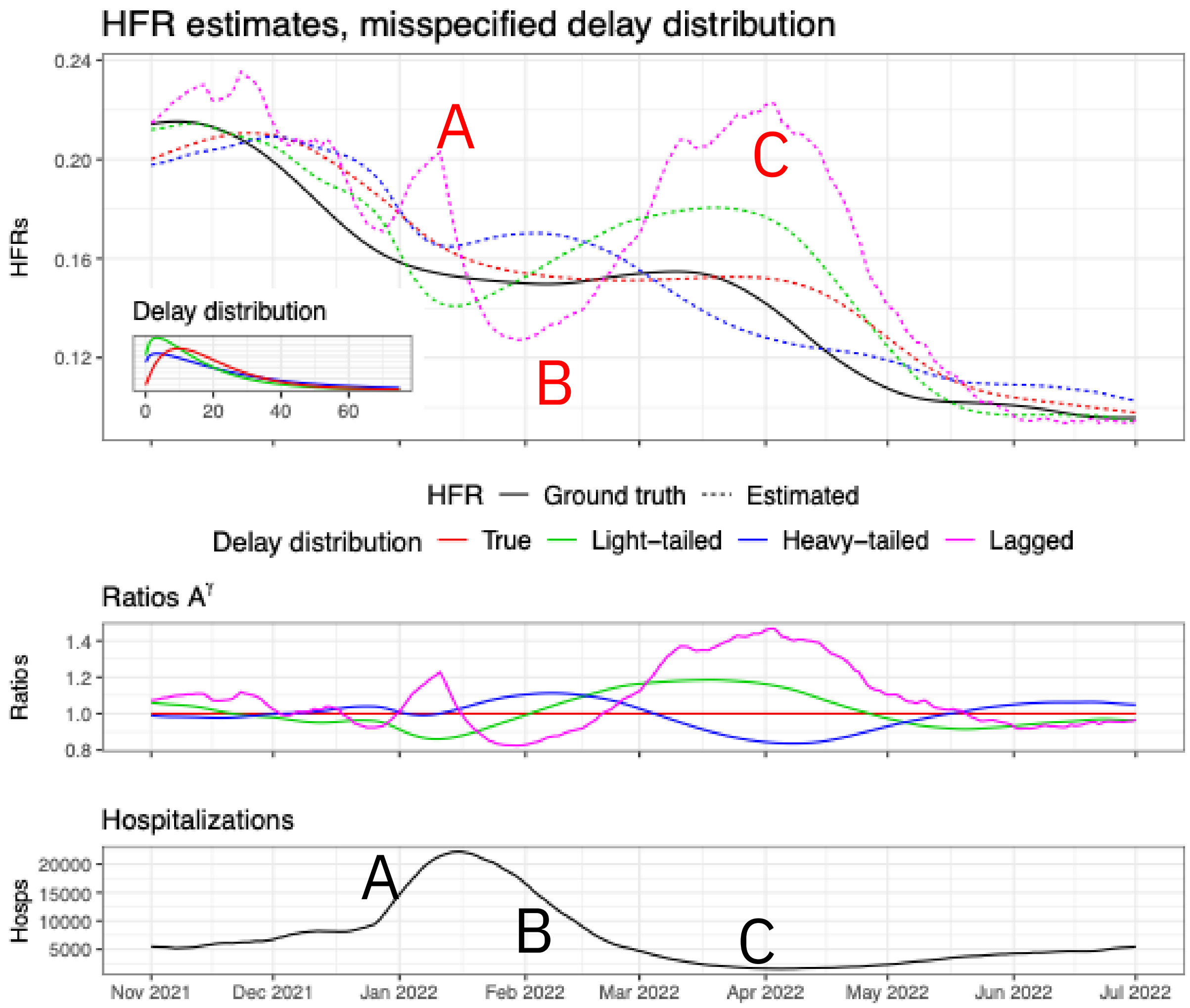
C

Misspecified analysis

For oracle delay distribution π, misspecified estimate γ, and

$$A_t^\gamma = \frac{\sum_{j=0}^{d} X_{t-j}\pi_j}{\sum_{j=0}^{d} X_{t-j}\gamma_j},$$

$$\boxed{\mathrm{Bias}\left(\hat{p}_t^\gamma\right) = A_t^\gamma \mathrm{Bias}\left(\hat{p}_t^\pi\right) + p_t\left(A_t^\gamma - 1\right)}$$

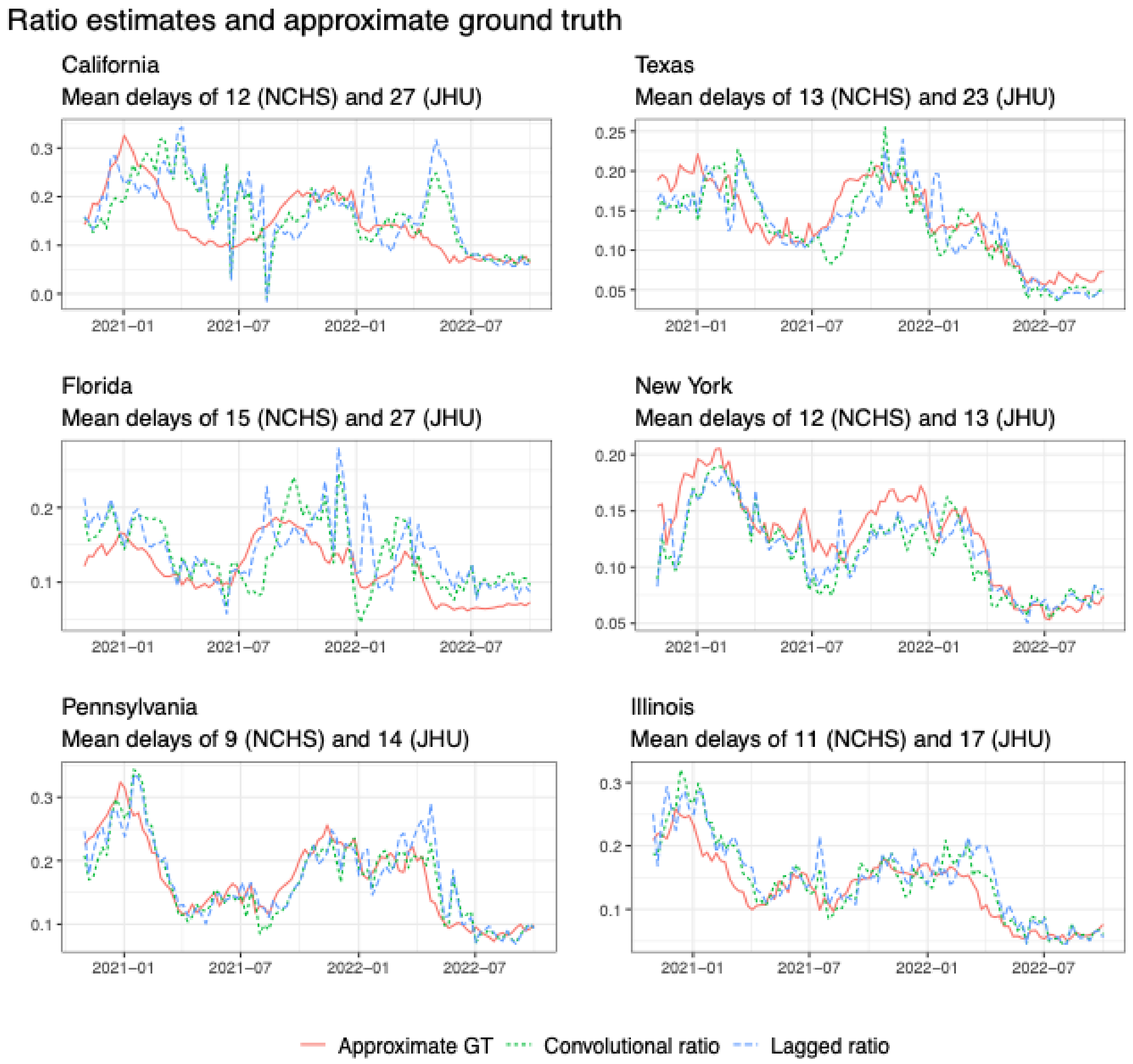


Bias of Convolutional Ratio with Misspecified Delay Distribution γ

- Arises as a consequence of changing primary incidence.
- Heuristics for lagged ratio:
  a. Too high during rise
  b. Too low during fall
  c. Too high after leveling out

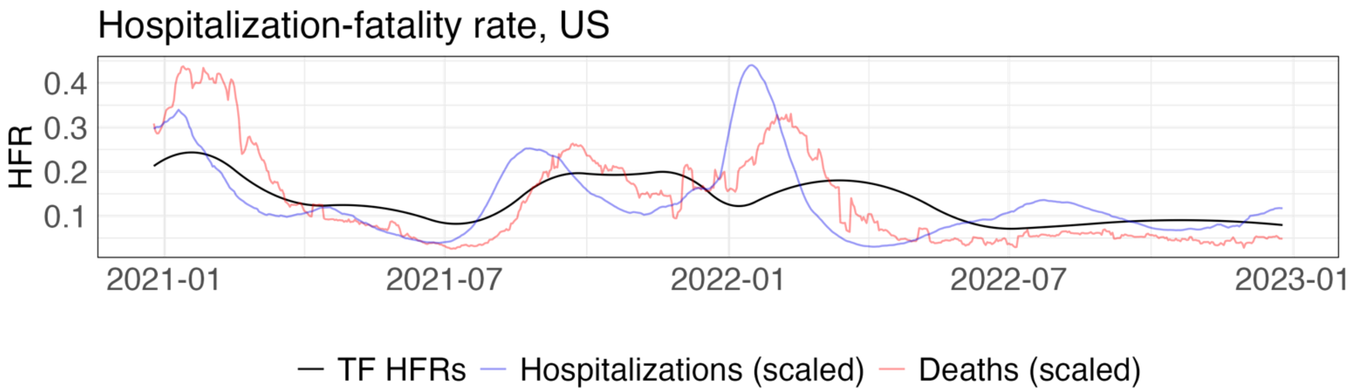
State-level results

- We estimate HFRs on JHU, which uses deaths aligned by report date – not the date the actually occurred.
- Longer reporting delays → heavier-tailed delay distribution → more bias (well-specified)
- Convolutional ratio consistently outperforms lagged ratio, which again is
    a. Too high during rise
    b. Too low during fall
    c. Too high after leveling out

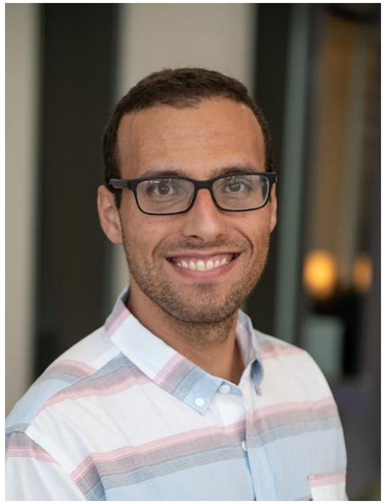


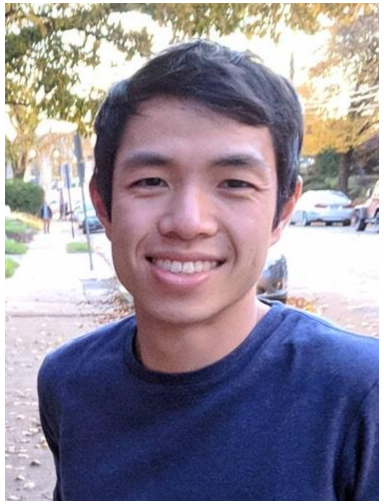
Follow-up work: Improving severity estimation

- Currently, we are developing a new method that avoids these biases.
- Instead of obtaining only the current severity rate, our approach estimates the curve *over all time*, then takes the most recent prediction.
  - We approximate maximum likelihood estimation on a faithful probabilistic model, using modern smoothing techniques for stability.
- Preliminary results demonstrate large improvements on retrospective analysis; we have yet to test its efficacy in the real-time setting.



Collaborators





Thanks for your attention!