Estimating the Cumulative Incidence of COVID-19 in the United States Using Four Complementary Approaches

Fred S. Lu[*,1,2] Andre T. Nguyen[*,1,3,4] Nicholas B. Link[*,1] Jessica T. Davis[7]
Matteo Chinazzi[7] Xinyue Xiong[7] Alessandro Vespignani[7] Marc Lipsitch[5]
Mauricio Santillana[1,5,6,†]

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA
[2]Department of Statistics, Stanford University, Stanford, CA
[3]University of Maryland, Baltimore County, Baltimore, MD
[4]Booz Allen Hamilton, Columbia, MD
[5]Department of Epidemiology, Harvard T.H. Chan School of Public Health
[6]Department of Pediatrics, Harvard Medical School, Boston, MA
[7]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA USA
[*]These authors contributed equally to this study.
[†]Correspondance to: Mauricio Santillana (msantill@fas.harvard.edu)

August 6, 2020

Supplementary Materials

1 Divergence by Location

Figures 1 and 2 show the *Divergence* method model fits for all available locations. COVID-19 is treated as an intervention, and we measure the impact of COVID-19 on observed CDC ILI, using predictions of ILI from the IDEA model and the virology model as counterfactuals. The impact of COVID-19 is calculated as the difference between the higher observed CDC ILI and the lower IDEA model predicted ILI and virology predicted ILI. The impact directly maps to an estimate of COVID-19 ILI-symptomatic case counts. Virology-predicted ILI is omitted when virology data is not available. We note that model fit quality varies by location. CDC reported ILI activity is plotted in blue, IDEA model predicted ILI is plotted in orange, and virology predicted ILI is plotted in green. We note that this method is meaningful only at the beginning of the outbreak (March 2020), while ILI surveillance systems are still fully operational and before they are impacted by COVID-19. The disappearance of the divergence does not mean that the outbreak is over, but rather that the ILI signal is no longer reliable.

Figure 1: Divergence model fits for first half of locations.

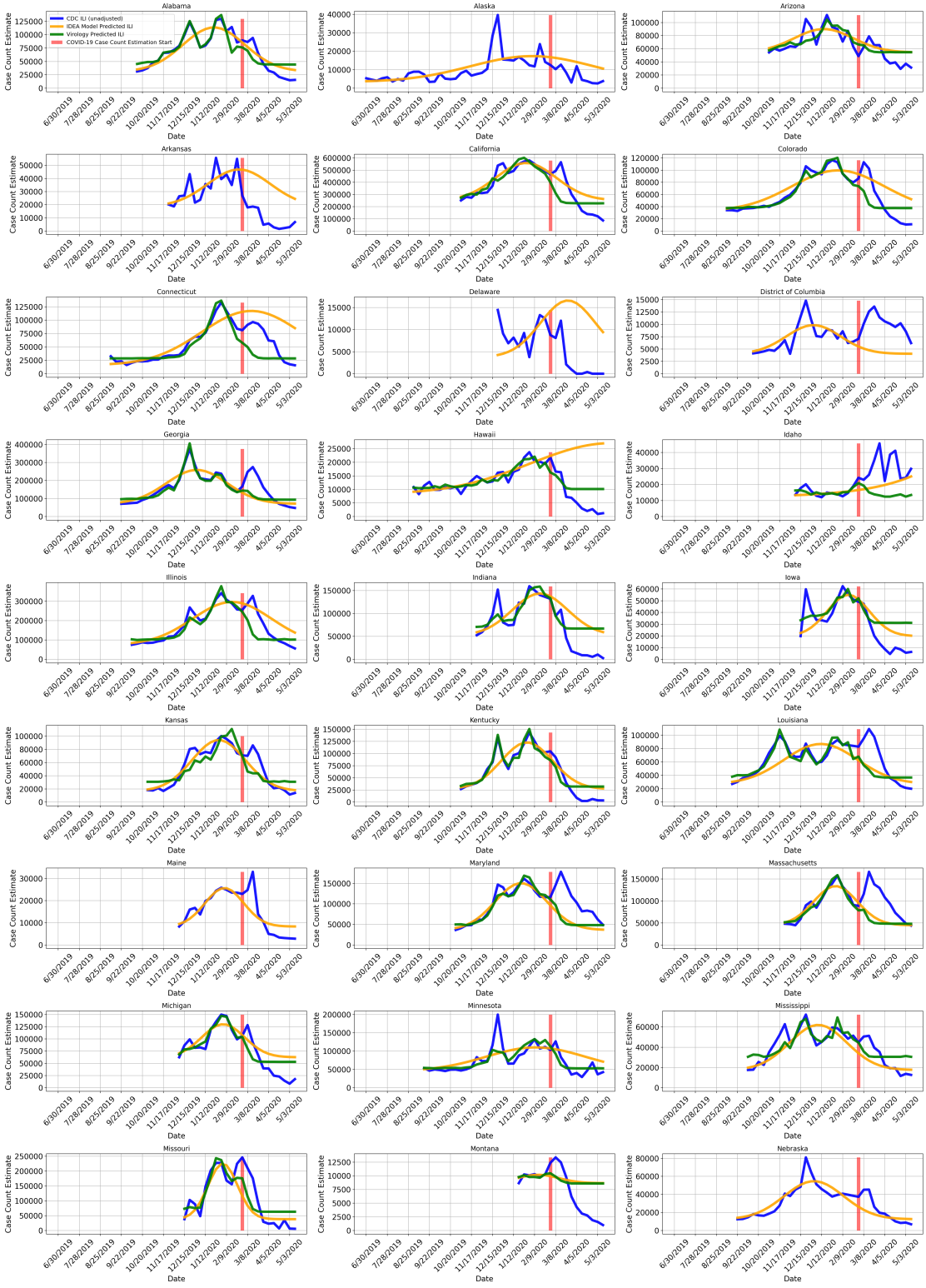
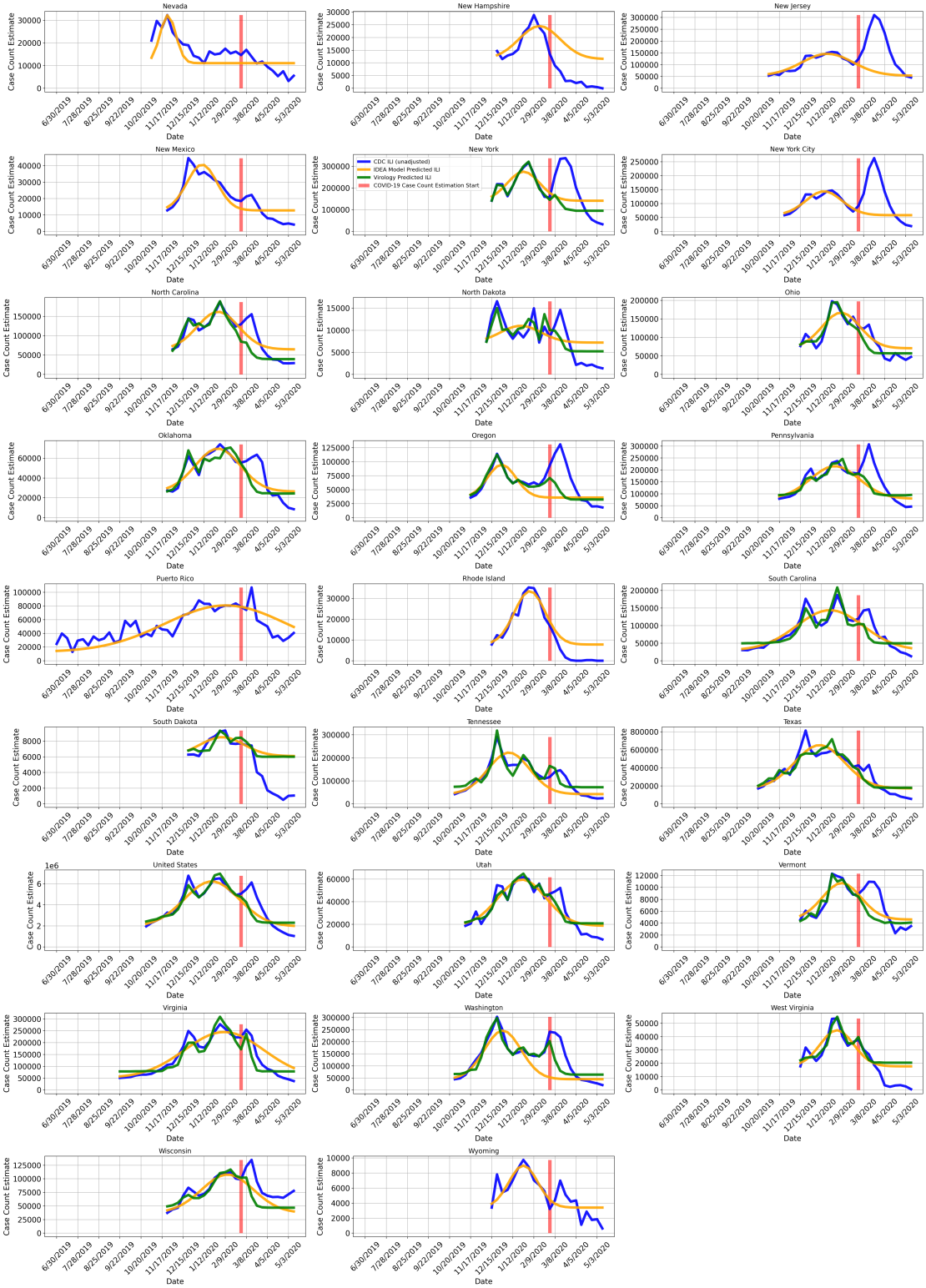


Figure 2: Divergence model fits for second half of locations.



2 Time Series Plots for All Methods

Figures 3 and 4 show the cumulative estimated counts for each week over the entire study period of March 1, 2020 to May 16, 2020, compared with cumulative reported counts, in each location in the United States. The solid and dotted lines indicate adjusted and unadjusted methods, respectively. Due to the seasonal nature of ILI information, estimates from all methods besides $mMAP$ are limited to April 4, 2020.

Figure 3: Cumulative case time series for first half of locations.

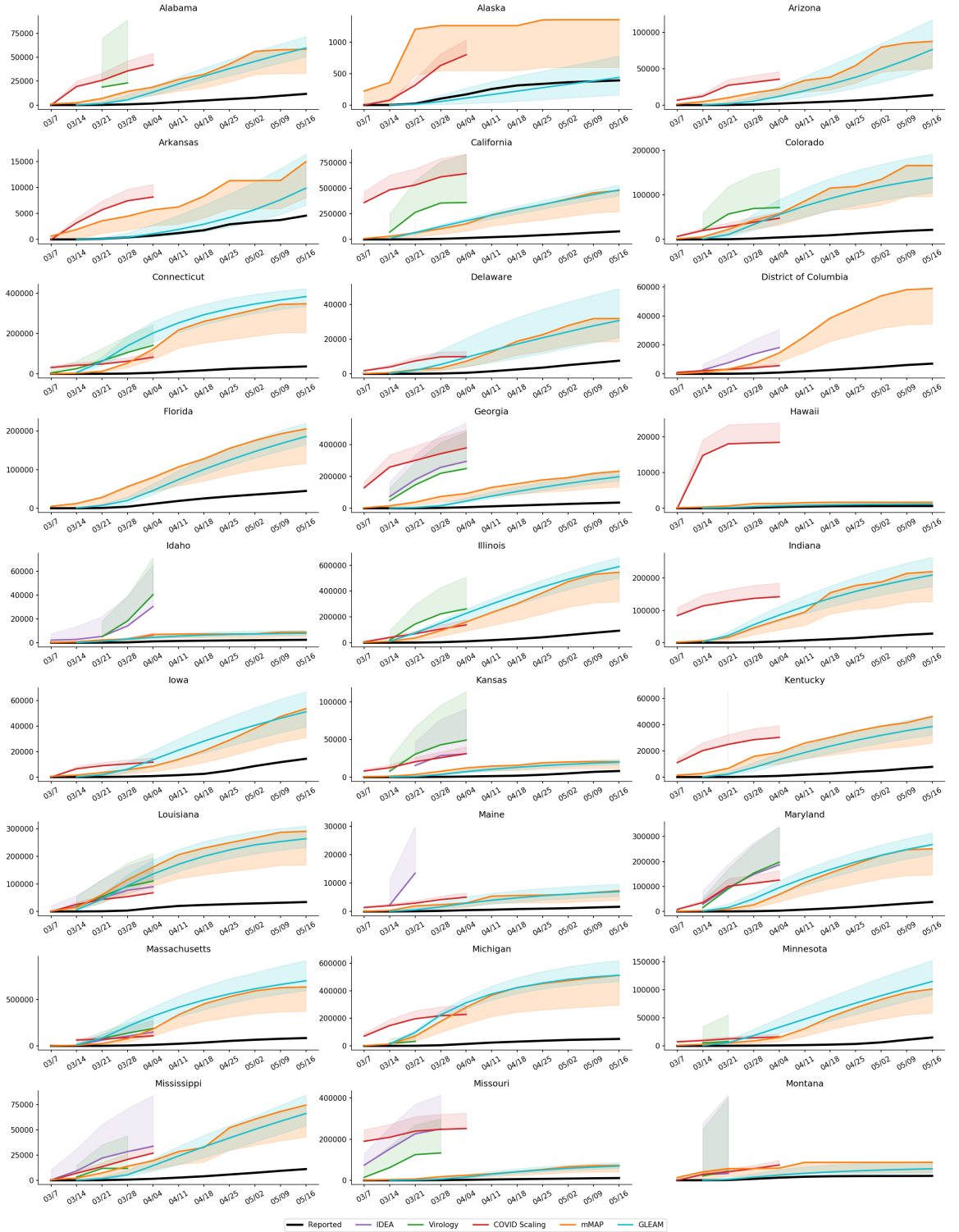
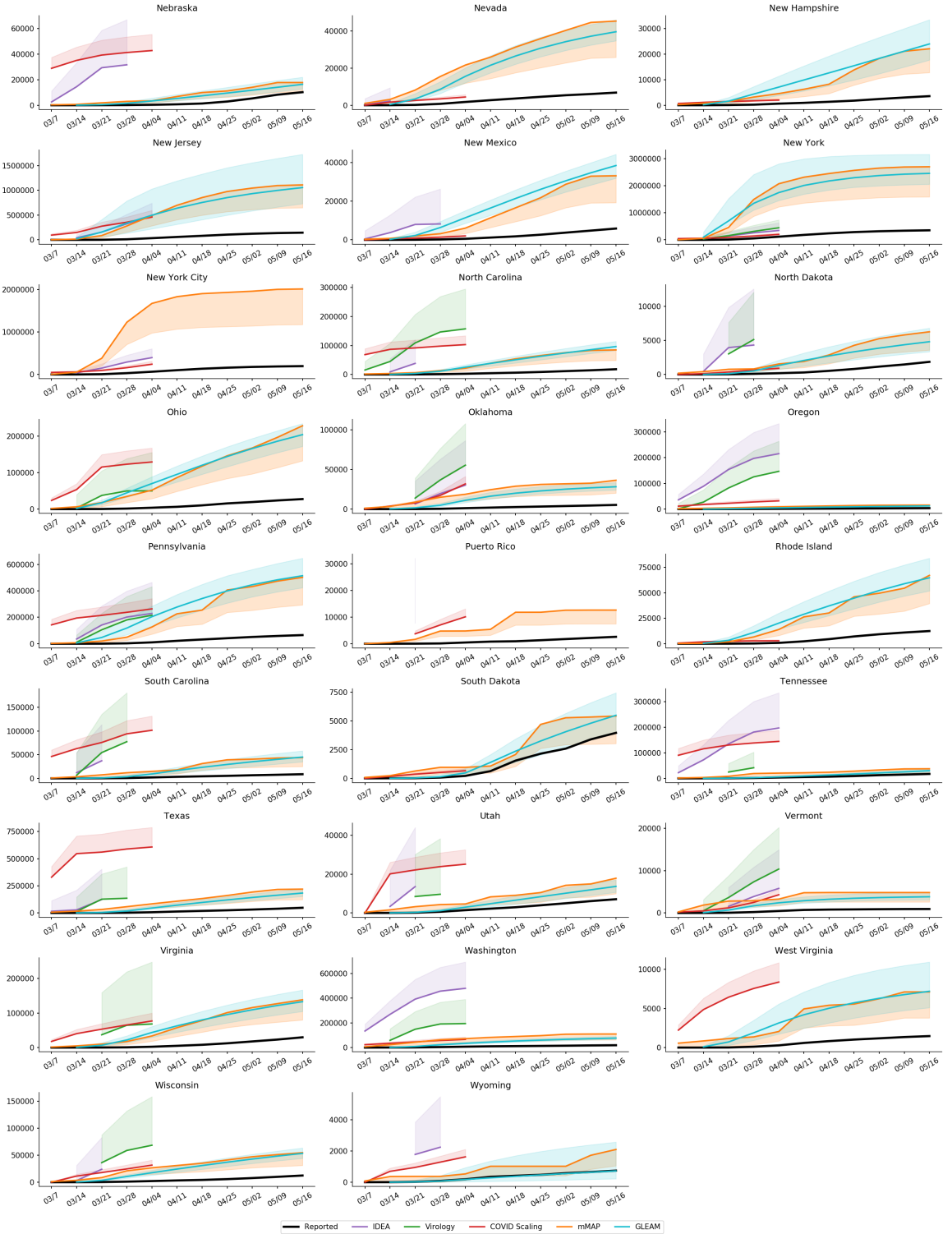


Figure 4: Cumulative case time series for second half of locations.



3 Virology-based Estimation

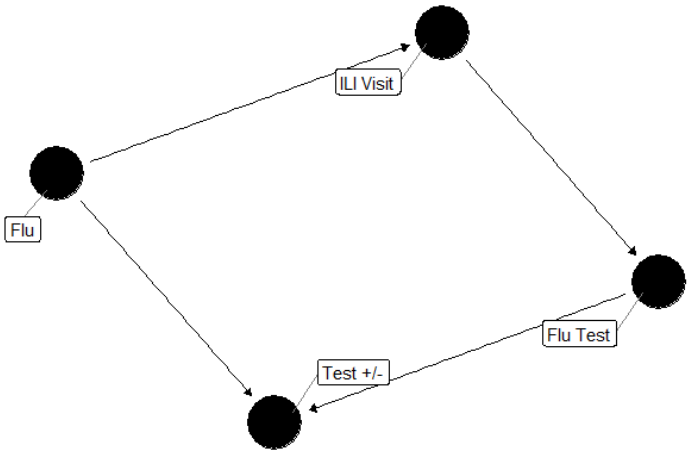


Figure 5: Causal DAG affecting flu positive results.

Both the virology-based *Divergence* model and the *COVID Scaling* method rely on the extrapolation of positive testing data to the actual symptomatic incidence of the disease. The causal diagram shown in Fig. 5 shows that an individual's flu test result depends on whether they have the disease, but also whether they receive a test in the first place (by going through the ILI visit path). More broadly, the relationship between test positive results and true disease counts are influenced by testing availability. We approximate the availability using the total administered tests divided by ILI cases. Identical reasoning applies for analysis of COVID-19 cases, as done in the COVID Scaling method.

We formulate a valid control as having the following two properties:

1. The control produces a reliable estimate of ILI activity.
2. The control is not affected by the COVID-19 intervention (that is, the model of ILI conditional on any relevant predictors is independent of COVID-19).

In Table 1, we show that the total positive tests divided by the availability satisfies both properties and successfully estimates the true flu counts (in the perfectly distributed case) even when a surge of COVID-19 cases is added.

	Baseline cases			With COVID-19 cases		
Data	1	2	3	1	2	3
Flu (F)	*20*	*20*	*40*	*20*	*20*	*20*
ILI (I)	100	100	200	200	200	400
Test (N)	10	50	50	10	50	50
Positive (F^+)	2	10	10	1	5	2.5
Availability (N/I)	0.1	0.5	0.25	0.05	0.25	0.125
Predict $\hat{F}$	**20**	**20**	**40**	**20**	**20**	**20**
Predict $\hat{I}$	**100**	**100**	**200**	**100**	**100**	**100**

Table 1: Series of examples showing that the proposed estimator predicts flu cases correctly even when potential COVID-19 is added.

4 Mortality-MAP Analysis

4.1 Proof of Case Recovery Given Convergence

In this section we will prove that if *mMAP* converges, which it does for every location in this analysis, the cases predicted by *mMAP*, C_d, fully recover deaths. That is that

$$D(t) = \sum_{\tau=1}^{t-1} p(T = t - \tau) \cdot C_d(\tau) \; \forall \; t \; \in 1..t_{max} \tag{1}$$

First, note that

$$\begin{aligned} C_d^{(i)}(t) &= \frac{C_{d^*}^{(i)}(t)}{p(T \leq (t_{max} - t))} \\ &= \frac{1}{p(T \leq (t_{max} - t))} \sum_{\tau=t+1}^{t_{max}} D(\tau) \cdot \frac{p(T = (\tau - t)) \cdot \frac{C_d^{(i-1)}(t)}{\sum C_d^{(i-1)}(t)}}{\sum_{s=1}^{\tau-1} p(T = (\tau - s)) \cdot \frac{C_d^{(i-1)}(s)}{\sum C_d^{(i-1)}(t)}} \\ &= \frac{1}{p(T \leq (t_{max} - t))} \sum_{\tau=t+1}^{t_{max}} \frac{D(\tau) \cdot p(T = (\tau - t)) \cdot C_d^{(i-1)}(t)}{\sum_{s=1}^{\tau-1} p(T = (\tau - s)) \cdot C_d^{(i-1)}(s)} \end{aligned} \tag{2}$$

Assuming mortality-MAP converges, $C_d(t) = C_d^{(i)}(t) = C_d^{(i-1)}$, so

$$\begin{aligned} C_d(t) &= \frac{1}{p(T \leq (t_{max} - t))} \sum_{\tau=t+1}^{t_{max}} \frac{D(\tau) \cdot p(T = (\tau - t)) \cdot C_d(t)}{\sum_{s=1}^{\tau-1} p(T = (\tau - s)) \cdot C_d(s)} \\ \implies p(T \leq (t_{max} - t)) &= \sum_{\tau=t+1}^{t_{max}} \frac{D(\tau) \cdot p(T = (\tau - t))}{\sum_{s=1}^{\tau-1} p(T = (\tau - s)) \cdot C_d(s)} \end{aligned} \tag{3}$$

(1) can be shown by induction. First, we will show that it holds for $t = t_{max} - 1$ and then show that if it is true for t_{i+1} then it must be true for t_i.

Setting $t = t_{max} - 1$, from (3) we see that

$$\begin{aligned} P(T{\leq}1) &= \frac{D(t_{max}) \cdot P(T=1)}{\sum_{s=1}^{t_{max}-1} P(T=(t_{max}-1-s)) \cdot C_d(s)} \\ \implies & \sum_{s=1}^{t_{max}-1} P(T=(t_{max}-1-s)) \cdot C_d(s) = D(t_{max}) \end{aligned} \tag{4}$$

since $P(0) = 0$, $P(T = 1) = P(T{\leq}1)$. Thus, (1) holds for $t = t_{max} - 1$. Now, assume (1) is true for all $t > t_i$. From (3),

$$\begin{aligned} P(T{\leq}(t_{max}-(t_i-1))) &= \sum_{\tau=t_i}^{t_{max}} \frac{D(\tau) \cdot P(T=(\tau-(t_i-1)))}{\sum_{s=1}^{\tau-1} P(T=(\tau-s)) \cdot C_d(s)} \\ &= \left[\frac{D(t_i) \cdot P(T=1)}{\sum_{s=1}^{t_i-1} P(T=t_i-s) \cdot C_d(s)} + \sum_{\tau=t_i+1}^{t_{max}} \frac{D(\tau) \cdot P(T=(\tau-(t_i-1)))}{\sum_{s=1}^{\tau-1} P(T=(\tau-s)) \cdot C_d(s)} \right] \\ &= \left[\frac{D(t_i) \cdot P(T=1)}{\sum_{s=1}^{t_i-1} P(T=t_i-s) \cdot C_d(s)} + \sum_{\tau=t_i+1}^{t_{max}} P(T=(\tau-(t_i-1))) \right] \end{aligned} \tag{5}$$

In the final step, $D(\tau)$ and the denominator cancel out because (1) is true for all $t > t_i$. Subtracting probabilities from both sides we end up with.

$$P(T=1) = \frac{D(t_i) \cdot P(T=1)}{\sum_{s=1}^{t_i-1} P(T=t_i-s) \cdot C_d(s)} \implies \sum_{s=1}^{t_i-1} P(T=t_i-s) \cdot C_d(s) = D(t_i) \tag{6}$$

Therefore, (1) is true for t_i and by induction is true for all $t < t_{max}$. Note that C_d is not a unique solution to the equation; since there are more potential days of cases than reported deaths this system is not full rank and there are infinite solutions (if C_d is allowed to be continuous). This result shows that at least the current estimate of C_d sensibly predicts the reported deaths. The next section demonstrates that this estimate of C_d does seem to be accurate for simulated and empirical data.

4.2 Satisfying Case Fatality Ratio Calculation

The authors of [1] propose an unbiased estimator of the case fatality rate as the following. In this study we are using the symptomatic case fatality ratio (sCFR), so here we define C as the total

symptomatic infections.

$$sCFR = \frac{\sum_{t=1}^{t_{max}} D(t)}{\sum_{t=1}^{t_{max}} \sum_{\tau=1}^{t-1} C(t-\tau) \cdot p(T=\tau)} \tag{7}$$

Note that the notation from the paper referenced is adapted to match the notation here, and that here P(T=0) so the summation limits are adjusted. We can show that the results from (1) satisfy this calculation of sCFR by showing that from our estimates of C, the RHS above equals the LHS. Note that in our formulation of C, $C_d = sCFR \cdot C$, since C_d is the time series of cases that end up in death, and C is the time series of all symptomatic cases.

$$\begin{aligned}
&\frac{\sum_{t=1}^{t_{max}} D(t)}{\sum_{t=1}^{t_{max}} \sum_{\tau=1}^{t-1} C(t-\tau) \cdot p(T=\tau)} \\
=&\frac{\sum_{t=1}^{t_{max}} \sum_{\tau=1}^{t-1} C_d(\tau) \cdot p(T=t-\tau)}{\sum_{t=1}^{t_{max}} \sum_{\tau=1}^{t-1} C(t-\tau) \cdot p(T=\tau)} \\
=&\frac{\sum_{t=1}^{t_{max}} \sum_{\tau=1}^{t-1} sCFR \cdot C(\tau) \cdot p(T=t-\tau)}{\sum_{t=1}^{t_{max}} \sum_{\tau=1}^{t-1} C(t-\tau) \cdot p(T=\tau)} \\
=&sCFR
\end{aligned} \tag{8}$$

To see that the numerator and denominator cancel out, substitute $j = t - \tau$ into the denominator. This demonstrates that our method converges to solutions that match previously researched formulations. Dependent on assumptions of accurate death reporting, the sCFR, and distribution of time from symptom onset to death, this method can accurately predict the unobserved symptomatic case time series.

4.3 Simulated and Empirical Validation

To validate *mMAP*, it was analyzed using simulated and real death data until June 7 from six countries: United States, China, Italy, Spain, Germany, and South Korea. Figure 6 compares *mMAP* predicted cases with reported cases. To visually scale the reported cases, the following equation is used:

$$reported\text{-}scaled = reported \cdot \frac{\sum predicted}{\sum reported}$$

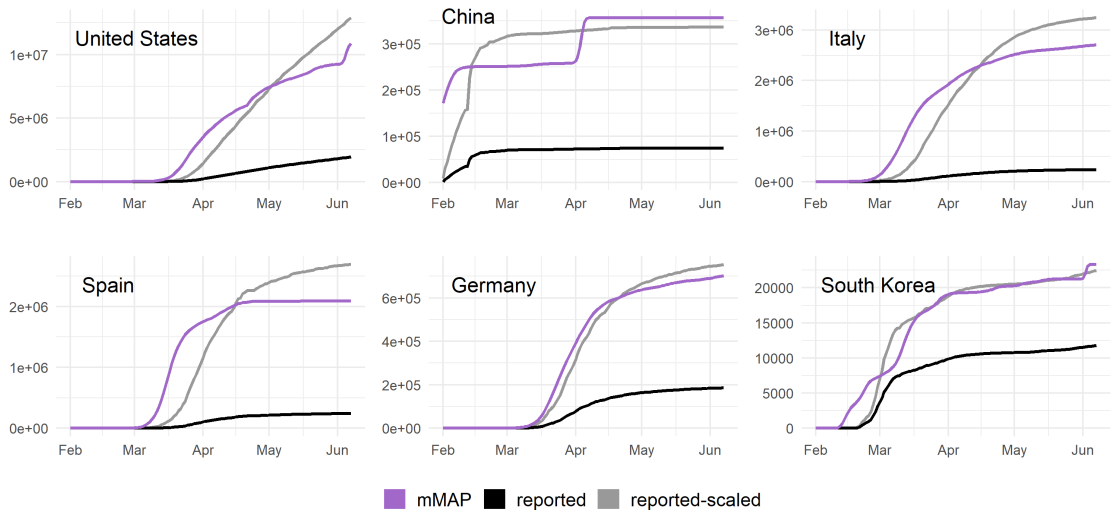
While the scales differ, the trends of predicted cases generally follow the trends of reported cases, especially in Italy, Germany, and South Korea. In the United States, Italy, and Spain, the differences could be a result of increasing case detection around the start of April; as testing increases,

we would expect to see more of a relative increase in reported cases than in reported deaths (because we would likely be picking up more of the less severe cases), which would cause the reported cases to increase more steeply than *mMAP* predictions.

In figure 7, the deaths for each country are simulated from the reported cases. Deaths are stochastically simulated from the reported cases using the log-normal distribution from symptom onset to death and an sCFR of 0.01. From the simulated deaths, *mMAP* predicts the original cases. As demonstrated by the proof in section 2.1, *mMAP* recovers cases on convergence (note it does not completely recover cases here because of the randomness of the simulation).

Both plots offer validation that *mMAP* can successfully predict the trend of the reported cases. However, these plots do not demonstrate if the scale of *mMAP* predictions are on target, as this is influenced by the under-reporting of deaths and the sCFR.

Figure 6: *mMAP* predictions compared to reported cases.



References

[1] Hiroshi Nishiura, Don Klinkenberg, Mick Roberts, and Johan AP Heesterbeek. Early epidemiological assessment of the virulence of emerging infectious diseases: a case study of an influenza pandemic. *PLoS One*, 4(8), 2009.

Figure 7: Simulated *mMAP* predictions compared to reported cases.