Estimating the Prevalence of COVID-19 in the United States: Three Complementary Approaches

Estimating the Prevalence of COVID-19 in the United States: Three Complementary Approaches

Fred S. Lu[*,1,2] Andre T. Nguyen[*,1,3,4] Nick Link[*,1] Mauricio Santillana[1,5,†]

[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 Pediatrics, Harvard Medical School, Boston, MA
[*]These authors contributed equally to this study.
[†]Correspondance to: Mauricio Santillana (msantill@fas.harvard.edu)

April 18, 2020

Abstract

Effectively designing and evaluating public health responses to the ongoing COVID-19 pandemic requires accurate estimation of the week to week burden of COVID-19. Unfortunately, a lack of systematic testing across the United States (US) due to equipment shortages and varying testing strategies has hindered the usefulness of the available positive COVID-19 case counts. We introduce three complementary approaches aimed at estimating the prevalence of COVID-19 in each state in the US as well as in New York City. Instead of relying on an estimate from a single data source or method that may be biased, we provide multiple estimates, each relying on different assumptions and data sources. Across our three approaches, there is a consistent conclusion that estimated state-level COVID-19 case counts usually vary from 10 to 100 times greater than the official positive test counts. Nationally, our lowest and highest estimates of COVID-19 cases in the US from March 1, 2020 to April 4, 2020 are 2.7 and 8.3 million (9 to 27 times greater). These estimates are to be compared to the cumulative confirmed cases of about 311,000 as of April 4th. Our approaches demonstrate the value of leveraging existing influenza-like-illness surveillance systems for measuring the burden of new diseases that share symptoms with influenza-like-illnesses. Our methods may prove useful in assessing the burden of COVID-19 in other countries with comparable influenza surveillance systems.

1 Introduction

COVID-19 (SARS-CoV-2), is a coronavirus that was first identified in Hubei, China, in December of 2019. On March 11, due to its extensive spread, the World Health Organization (WHO) de-

clared it a pandemic [1]. As of April 8, 2020, COVID-19 has infected people in nearly every country globally and in all 50 states in the United States (US) [2]. This pandemic now poses a substantial public health threat with potentially catastrophic consequences. Reliable estimates of COVID-19 infections, particularly at the start of the outbreak, are critical for appropriate resource allocation, effective public health responses, and improved forecasting of disease burden [3].

A lack of widespread testing due to equipment shortages, varying levels of testing by region over time, and uncertainty around test sensitivity make estimating the point prevalence of COVID-19 difficult [4, 5]. In addition, it has been estimated that 18% [6] to 50% [7, 8] of people infected with COVID-19 do not show symptoms. Even in symptomatic infections, under-reporting can further complicate the accurate characterization of the COVID-19 burden. For example, one study estimated that in China, 86% of cases had not been captured by lab-confirmed tests [9], and it is possible that this percentage is even higher in the US [5]. Finally, it has been suggested that the available information on confirmed COVID-19 cases across geographies may be an indicator of the local testing capacity over time (as opposed to an indicator of the epidemic trajectory). Thus, solely relying on positive test counts to infer the total number of COVID-19 infections, and the epidemic trajectory, may not be sensible [10].

The aim of this study is to develop alternative methodologies, each with different sets of inputs and assumptions, to estimate the weekly prevalence of COVID-19 in each state in the US. One such approach is to analyze region-specific changes in the number of individuals seeking medical attention with influenza-like illness (ILI), defined as having a fever in addition to a cough or sore throat. The significant overlap in symptoms common to both ILI and COVID-19 suggests that leveraging existing disease monitoring systems, such as ILINet, a sentinel system created and maintained by the United States Centers of Disease Control and Prevention (CDC) [11, 12], may offer a way to estimate the prevalence of COVID-19 without needing to rely on COVID-19 testing results. Importantly, recent regional increases in ILI in conjunction with stable or decreasing influenza case numbers present a discrepancy (or an increase in ILI not explained by an increase in influenza) that can be used to impute COVID-19 cases.

A second and related approach uses ILI data to deconfound COVID-19 testing results from state-level testing capabilities. These two methods show that existing ILI surveillance systems provide a useful signal for measuring COVID-19 prevalence in the US, especially during the early stages of the outbreak. A third and final approach, which uses reported COVID-19-attributed deaths to estimate COVID-19 prevalence and improves upon previously introduced methodologies [13, 14, 15, 16] is presented. COVID-19 deaths may represent a lower-noise estimate of cases than surveillance testing given that patients who have died are sicker, more likely to be hospitalized, and thus more likely to be tested than the general infected population.

While previous work has attempted to quantify COVID-19 prevalence in the United States using discrepancies in ILI trends [17, 18], to the best of our knowledge this study is the first to offer a range of estimates at the state level, leveraging a suite of complementary methods based on different assumptions. We believe that this provides a more balanced picture of the uncertainty over COVID-19 prevalence in each state. While our results are approximations and depend on a variety of (likely time-dependent) estimated factors, we believe that our presented case counts better represent prevalence than simply relying on laboratory-confirmed COVID-19 tests. Providing such estimates for each state enables the design and implementation of more effective and efficient public health measures to mitigate the effects of the ongoing COVID-19 epidemic outbreak. While the scope of this paper is focused on the United States, the methods introduced here are general enough that they may prove useful to estimate COVID-19 burden in other locations with comparable disease (and death) monitoring systems.

2 Results

We implement four methods based on three complementary approaches to estimate the prevalence of COVID-19 within the US from March 1st to April 4th, 2020. These dates correspond to the early stages of the outbreak (with fewer than 50 confirmed cases in the US), up to the date of the most recent available CDC reports as of April 16th. The first two methods, labeled *div-IDEA* and *div-Vir*, fall under the *Divergence* approach, which first estimates what the level of ILI activity across the US would have been if the COVID-19 outbreak had not occurred. Each method develops a control time series and uses the unexpected increase in ILI rate over the control to infer the burden of COVID-19. *div-IDEA* is based on an epidemiological model, the IDEA model [19], fitted to the observed 2019-2020 ILI (prior to the introduction of COVID-19 to the US), while *div-Vir* is based on the time-evolution of empirical observations of positive virological influenza test statistics. The third method, based on the *COVID Scaling* approach, leverages healthcare ILI visits and COVID-19 test statistics to directly infer the proportion of ILI due to COVID-19 in the full population. The fourth method, based on the *mortality MAP* (*mMAP*) approach, uses the time series of COVID-19-attributed deaths in combination with the observed epidemiological characteristics of COVID-19 in hospitalized individuals, to infer the latent disease onset time series, which is then scaled up to estimate case counts using the expected infection fatality ratio (IFR). The Methods section provides extensive details on the assumptions and data sources for each of these approaches.

2.1 Adjusted Assumptions Represent Most Likely Scenarios

Each of our methods has an adjusted version, which represents our best guess taking into account all information available to us, and an unadjusted version, which uses pre-COVID-19 baseline information. Specifically, the adjusted divergences (*div-IDEA* and *div-Vir*) and *COVID Scaling* methods incorporate an increased probability that an individual with ILI symptoms will seek medical attention after the start of the COVID-19 outbreak based on recent survey data [20, 21]. The adjusted *mMAP* method supplements the confirmed COVID-19 deaths with unusual increases in pneumonia-related deaths across the country that may represent untested COVID-19 cases. In most states, as seen in Fig. 1, the adjusted estimates from each method are more closely clustered than their unadjusted counterparts, increasing our confidence in the adjusted range estimates of COVID-19 prevalence.

2.2 Estimated Case Counts Far Surpass Reported Positive Cases

We produced estimates for the national and state levels using these four methods for the time period between March 1, 2020 and April 4, 2020. These methods estimate that there had been 2.7 to 8.3 million COVID-19 cases in the US; in comparison, around 311,000 positive cases had been officially recorded during that time period. Fig. 1 displays the COVID-19 case count estimates from our methods at the national and state levels (and New York City) compared with the reported case numbers. The results suggest that the estimated true numbers of infected cases are uniformly much higher than those reported.

For reference, if one only adjusts the number of reported cases by the (likely) percentage of asymptomatic cases (18% [6] to 50% [7, 8]) and symptomatic cases not seeking medical attention (up to 73% [22]), one would conclude that the actual number of cases is higher, and about four to eight times the number of reported cases; this ratio would also be constant across states. In contrast, our methods frequently estimate 10-fold to 100-fold more cases than those reported and show significant state-level variability. The median estimate for the ratio of actual cases to reported cases for the adjusted *div-IDEA* method is 23 (with a 90% interval from 6 to 114), for adjusted *div-Vir* is 25 (5, 88), for adjusted *COVID-Scaling* is 14 (3, 62), and for adjusted *mMAP* is 9 (4,14). This highlights that models using only confirmed test cases may significantly underestimate the actual COVID-19 prevalence in the United States, which is consistent with what previous studies have shown [9, 18].

These methods also provide separate cumulative case estimates for each week within the studied period (*mMAP* provides daily estimates, but these are aggregated by week for comparison). Fig. 3 highlights the rapid increase in estimated COVID-19 cases over the United States as well as in New York City, Washington, and Louisiana, three locations which experienced early outbreaks. These

methods suggest that states under-reported COVID-19 case counts even early in March, likely due to limited testing availability. In New York and Louisiana, the estimates were more similar across methods than in Washington. Since Washington had already experienced an outbreak by February 28 [23], testing shortages may have been more pronounced than in the other states. Our divergence analysis approach does not rely on any COVID-19 testing data and therefore may provide more accurate estimates in Washington.

2.3 State-level Comparisons

Using the adjusted versions of our methods, we estimate between 14 and 33 states (31 using the median adjusted estimate) have actual case counts above 10 times the reported counts, depending on the method (Figs. 1 and 2). Five locations have at least one adjusted estimate above 100 times the reported counts (Nebraska, Oregon, Missouri, Hawaii, and Puerto Rico). Furthermore, our methods suggest that places with low official case counts, such as Alaska, Wyoming, South Dakota, and North Dakota, are in fact experiencing significantly more COVID-19 cases than are being tested. Even places with high official case counts, such as Georgia, Pennsylvania, and Texas, appear to be significantly under-reporting. As expected, our methods produce consistent high estimates in New York and New Jersey, which have reported especially high numbers of confirmed cases; though, compared to other states, New York reports a relatively high percentage of its predicted cases across all methods, suggesting that under-reporting may be less of a problem there.

All four methods generally agree on the ordering of states by case count (Table 1). Furthermore, they show strong rank correlations (larger than 0.65 across states and methods) with the reported case counts. *mMAP* has an especially high 0.96 correlation with the reported case counts, which is likely because official COVID-19 deaths and positive COVID-19 cases represent the same pool of patients and are therefore subject to the same bias. The other methods, however, rely on aggregate data from ILINet, which may cover a different set of patients. While the rank-correlation across methods is high, *mMAP* generally yields lower estimates than the others (Fig. 2). One possible explanation is that many deaths caused by COVID-19 are not being officially counted as COVID-19 deaths because of a lack of testing (and that accounting for increased pneumonia deaths does not fully capture this) [24]; further evidence of this reasoning is that New York City started reporting plausible COVID-19 deaths (as in, not needing a test result) [25], and *mMAP*'s estimates are closer (and actually higher) than the other methods' estimates there.

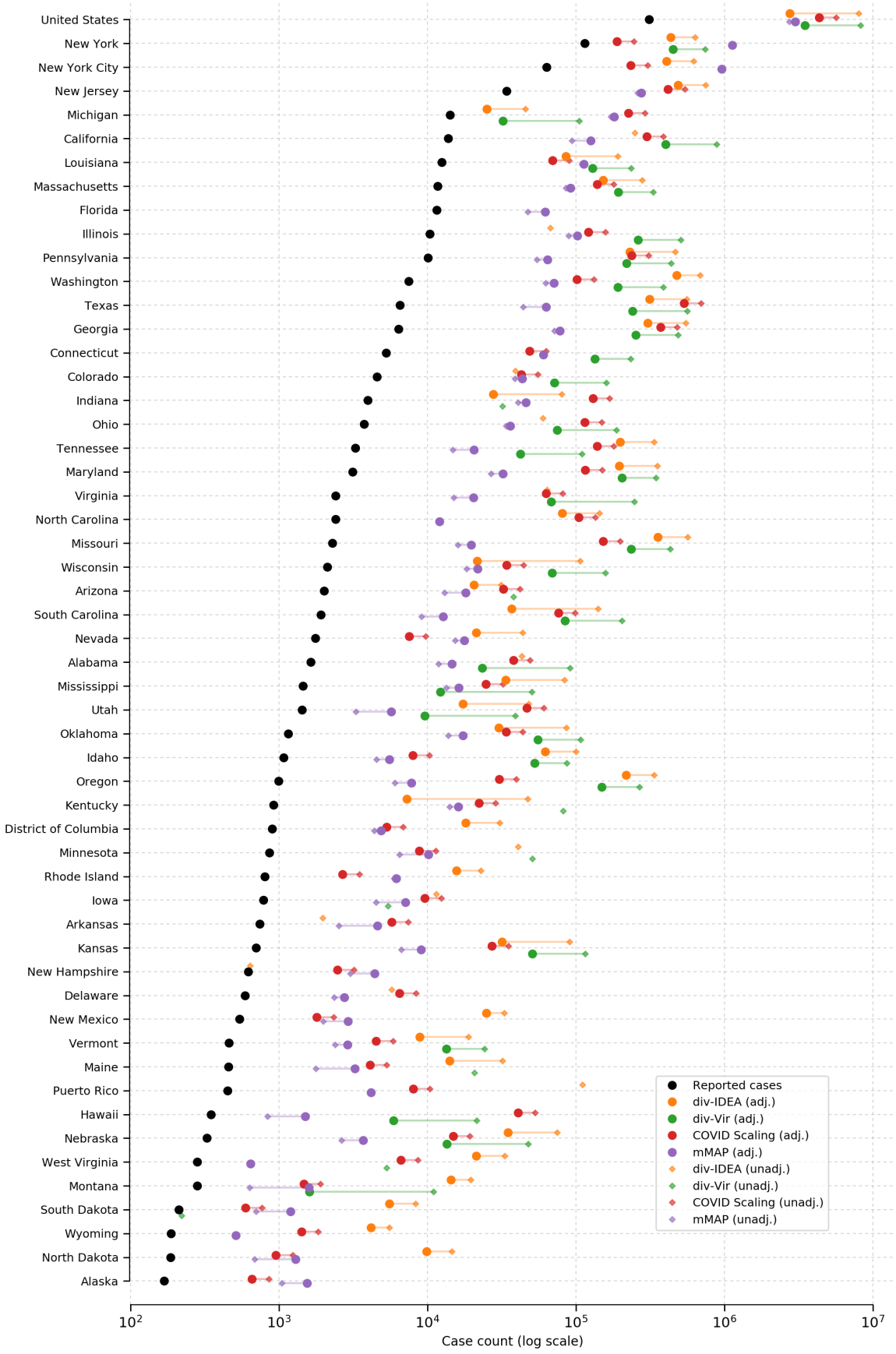


Figure 1: (Continued on the following page)

Figure 1: (On previous page) COVID-19 case count estimates compared with reported case counts at the national and state levels (and New York City) from March 1, 2020 to April 4, 2020. Cases are presented on a log scale. Adjusted methods take into account increased visit propensity (*div-IDEA*, *div-Vir*, *COVID Scaling*) and pneumonia-recorded deaths (*mMAP*). In places where the ILI-based methods show no divergence in observed and predicted ILI visits, the estimates of COVID-19 cannot be calculated and are not shown. Note that Florida does not provide ILI data, so only *mMAP* could be estimated there.

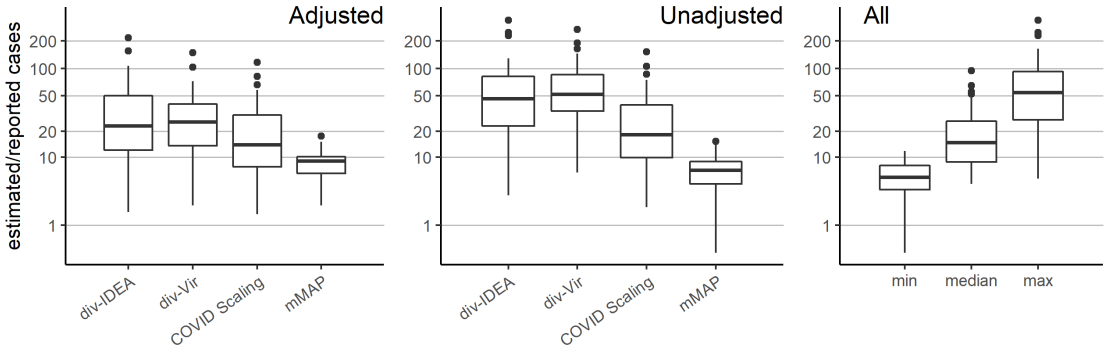


Figure 2: Distribution of the state-level ratios of estimated to reported case counts from March 1, 2020 to April 4. The right-hand plot shows the results of using all methods together: taking the *min*, *median*, *max* of the state-level estimates across methods.

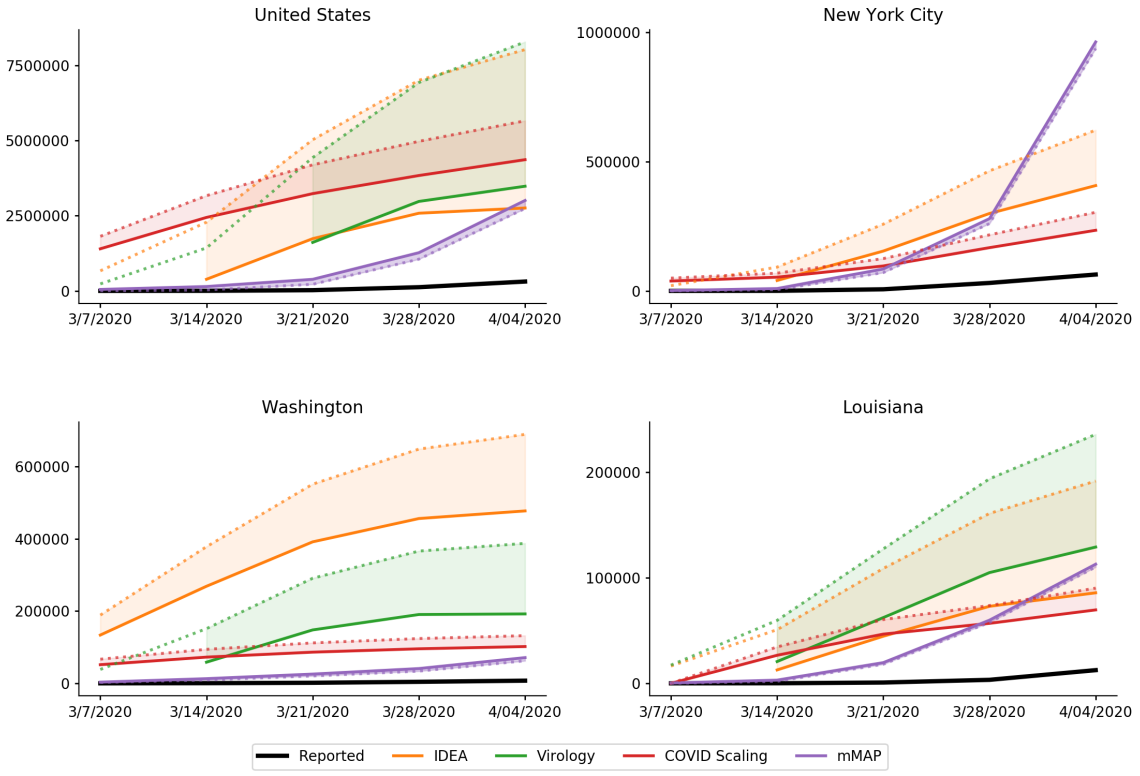


Figure 3: Cumulative weekly case counts since March 1, 2020 for the United States, New York City, Washington, and Louisiana, as estimated by each method and the reported cases. The estimate for each week indicates total cases up to the denoted date. Solid and dotted lines indicate the adjusted and unadjusted estimates, respectively. Refer to the Supplementary Materials for results over all locations.

	Reported	div-IDEA	div-Virology	COVID Scaling	mMAP
Reported	1.00	0.66	0.70	0.90	0.98
div-IDEA	–	1.00	0.75	0.72	0.65
div-Virology	–	–	1.00	0.70	0.70
COVID Scaling	–	–	–	1.00	0.88
mMAP	–	–	–	–	1.00

Table 1: Pairwise Spearman correlations between adjusted methods and reported case counts across the state level.

3 Discussion

We present four methods based on three distinct approaches to estimate the COVID-19 prevalence across the United States. The methods are complementary, in that they rely on different assumptions and use diverse datasets. Despite their clear differences, these methods estimate that the likely COVID-19 prevalence varies from 10 to 100 times higher, at the state level, than what has been reported so far in the U.S. As of April 16th, 2020, about 650,000 COVID-19 cumulative cases have been reported in the US. Assuming our (national) multiplicative factors to be good approximations for what took place from April 4th to April 16th, the current cumulative number of COVID-19 cases nationally could be anywhere from 6 to 16 million (9-fold to 25-fold higher than confirmed cases).

By design and due to the utilized data sources, our estimates using data from ILINet and confirmed cases (*Divergence* method and *COVID-Scaling*) likely better capture the number of COVID-19 cases as they would be detected at the time of hospitalization; thus, they may be inherently lagged by roughly 12 days after initial infection [26]. Considering an empirical (and likely naive) doubling time of a week, this means the prevalence of COVID-19, at any point in time (and stage of infection), could be 2-4 times higher than the estimates presented here. Taking this lag into consideration would suggest that it is plausible that up to 32 million cases individuals may be infected as of April 16th (50-fold higher than confirmed cases).

By providing ranges of estimates, both within and across models, this suite of methods offers a robust picture of the uncertainty in state-level COVID-19 case counts. When making public health decisions to respond to COVID-19, it is important to account for the uncertainty in estimates of case prevalence; the multiple estimates presented here provide a better picture of this than single point estimates.

Our approaches could be expanded to include other data sources and methods to estimate

prevalence, such as Google searches [27, 28, 29], electronic health record data [30], clinician's searches [31], and/or mobile health data [32]. Accurate and appropriately-sampled serological testing would provide the most accurate estimate of prevalence and would be useful for public health measures, especially when attempting to relax current shelter-in-place recommendations. In addition, appropriately-designed studies based on serologiucal testing could be used to evaluate the reliability of the methods presented in this study. This could inform prevalence estimation methods for COVID-19 in other countries as well as for future pandemics. The ILI-based methods presented in this study demonstrate the potential of existing and well-established ILI surveillance systems to monitor future pandemics that, like COVID-19, present similar symptoms to ILI. This is especially promising given the WHO initiative launched in 2019 to expand influenza surveillance globally [33]. Incorporating estimates from influenza and COVID-19 forecasting and participatory surveillance systems may prove useful in future studies as well [34, 35, 36, 37, 38, 39].

Limitations. The uncertainty and bias of each individual method should be considered carefully. The *Divergence* methods suffer from the same challenges faced when attempting to scale CDC-measured ILI activity to the entire population [40]. In particular, scaling to case counts in a population requires estimates for $p(\text{visit})$, the probability that a person seeks medical attention for any reason, and $p(\text{visit} \mid \text{ILI})$ which captures health care seeking behavior given that a person is experiencing ILI; these estimates are likely to change over time, especially during the course of a pandemic. Moreover, the weekly prevalence estimates from this method decrease towards the end of March, perhaps caused by a change in health care seeking behavior after the declaration of a national emergency on March 13, 2020 and the widespread implementation of shelter-in-place mitigation strategies. *COVID Scaling* relies on the assumption that COVID-19 positive test proportions uniformly represent the pool of all ILI patients and that shortages in testing do not bias the positive proportion upward or downward. Finally, *mMAP* is limited by assumptions of the Infection Fatality Rate, the distribution of time from case onset to death, and accurate reporting of COVID-19 deaths (or in the case of adjusted *mMAP*, that excess pneumonia deaths capture all unreported COVID-19 deaths). A high-level summary of the three methods, their estimation strategy, and their assumptions are provided in Table 2.

4 Conclusions

We have presented three complementary approaches for estimating the true COVID-19 prevalence in the United States from March 1 to April 4, 2020 at the national, state, and city (New York City) levels. The approaches rely on different datasets and modeling assumptions in order to balance the inherent biases of each individual method. While the case count estimates from these methods vary, there is general agreement among them that the actual state-level case counts are likely 10

to 100 times greater than what is currently being reported.

A more accurate picture of the burden of COVID-19 is actionable knowledge that will help guide and focus public health responses. Inevitably, as social distancing measures are relaxed, there will be a resurgence in cases. Yet, if the true case counts are near the upper bound of our estimated counts, then a substantial proportion (up to 10%) of the US population may have already been infected. In such a scenario, the US population may be closer to herd immunity than previously anticipated, and we may expect that subsequent waves of infection will eventually decrease in magnitude, until COVID-19 becomes a relatively controllable seasonal affliction like influenza [41].

5 Data and Methods

CDC ILI and Virology: The CDC US Outpatient Influenza-like Illness Surveillance Network (ILINet) monitors the level of ILI circulating in the US at any given time by gathering information from physicians' reports about patients seeking medical attention for ILI symptoms. ILI is defined as having a fever (temperature of 37.8+ Celsius) and a cough or a sore throat. ILINet provides public health officials with an estimate of ILI activity in the population but has a known availability delay of 7 to 14 days. National level ILI activity is obtained by combining state-specific data weighted by state population [12]. Additionally, the CDC reports information from the WHO and the National Respiratory and Enteric Virus Surveillance System (NREVSS) on laboratory test results for influenza types A and B. The data is available from the CDC FluView dashboard [11]. We omit Florida from our analysis as ILINet data is not available for Florida.

COVID-19 Case and Death Counts: The US case and death counts are taken from the New York Times repository, which compiles daily reports of counts at the state and county levels across the US [42]. For the *mMAP* validation in the supplementary materials, the case and death counts from other countries are taken from the John's Hopkins University COVID-19 dashboard [43]. Counts are taken up until April 14, 2020.

COVID-19 Testing Counts: In addition, daily time series containing positive and negative COVID-19 test results within each state were obtained from the COVID Tracking Project [44].

US Demographic Data: The age-stratified, state-level population numbers are taken from 2018 estimates from the US census [45].

5.1 Approach 1: Divergence

Viewing COVID-19 as an intervention, this approach aims to construct control time series representing the counterfactual 2019-2020 influenza season without the effect of COVID-19. While inspired by the synthetic control literature [46, 47], we are forced to construct our own controls since COVID-19 has had an effect in every state. We formulate a 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).

We construct two such controls, one model-based and one data-driven.

5.1.1 Method 1: Incidence Decay and Exponential Adjustment Model

The Incidence Decay and Exponential Adjustment (IDEA) model [19] is a single equation epidemiological model that estimates disease prevalence over time early in an outbreak while accounting for control activities and behaviours. The model is as follows:

$$I(t) = \left(\frac{R_0}{(1+d)^t} \right)^t$$

where $I(t)$ is the incident case count at serial interval time step t. R_0 is the basic reproduction number, and d is a discount factor modeling reductions in the effective reproduction number with time due to public health interventions, changes in public behavior, or environmental factors. The IDEA model has been shown to be identical to Farr's law for epidemic forecasting and can be expressed in terms of a susceptible-infectious-removed (SIR) compartmental model with improving control [48].

We fit the IDEA model to ILI case counts from the start of the 2019-2020 influenza season to the last week of February 2020. The start of the 2019-2020 influenza season is defined in a location specific manner as the first occurrence of two consecutive weeks with ILI activity above 2%. Model fitting is done using non-linear least squares with the Trust Region Reflective algorithm as the optimizer. Next, the model is used to predict what ILI would have been had the COVID-19 pandemic not occurred. In other words, we use the IDEA model ILI estimates as the counterfactual when assessing the impact of the COVID-19 intervention. When fitting the IDEA model, we use a serial interval of half a week, consistent with the serial interval estimates from [49] for influenza. We note that serial interval estimates from [50] for COVID-19 as well as from [51] for SARS-CoV-1 are longer than that of influenza, but that is not an issue as we use IDEA to model ILI.

5.1.2 Method 2: Virology

As an alternative control to the IDEA model, we also present an estimator of ILI activity using influenza virology results. As suggested by [17], there has been a divergence in March between CDC measured ILI activity and the fraction of ILI specimens that are influenza positive. Clinical virology time series were obtained from the CDC virologic surveillance system consisting of over 300 laboratories participating in virologic surveillance for influenza through either the US WHO Collaborating Laboratories System or NREVSS [12]. Total number of tests, total influenza positive tests, and percent positive tests are our variables of interest.

None of the three time series satisfy both properties of a valid control, as defined in 5.1. Total specimens and percent positive do not satisfy property 2 since total specimens is directly susceptible to increase when ILI caused by COVID-19 is added. Total positive flu tests satisfies property 2, but also depends on ILI through the quantity of tests administered.

We propose a modification that satisfies the properties. Let F_t^+, N_t, I_t denote positive flu tests, total specimens, and ILI visit counts respectively. In addition, let F_t be the true underlying ILI counts. For any week t we assume the following relation:

$$F_t = \frac{F_t^+ \cdot I_t}{N_t}$$

There are two interpretations of this quantity: 1) It extrapolates the positive test percentage (F_t^+/N_t) to all ILI patients (I_t), a quantity known in the mechanistic modeling literature as ILI+ [52]. 2) Test frequency is a confounder in the relationship between the number of positive tests (F^+) and total flu (F_t). Adjusting for test frequency closes the indirect pathway between F_t and F_t^+ [53]. In the Supplementary Material, we demonstrate over a series of examples that this estimator behaves as desired. Each estimate of F_t is then scaled to population ILI cases using least squares regression over pre-COVID-19 ILI counts.

5.1.3 ILI Case Count Estimation

In order to fit the IDEA and virology models, we estimate the ILI case count in the population from the CDC's reported percent ILI activity, which measures the fraction of medical visits that were ILI related.

In a similar fashion to the approach of [40], we can use Bayes' rule to map percent ILI activity to an estimate of the actual population-wide ILI case count. Let $p(\text{ILI})$ be the probability of any person having an influenza-like illness during a given week, $p(\text{ILI} \mid \text{visit})$ be the probability that a person seeking medical attention has an influenza-like illness, $p(\text{visit})$ be the probability that a

person seeks medical attention for any reason, and $p(\text{visit} \mid \text{ILI})$ the probability that a person with an influenza-like illness seeks medical attention. Bayes' rule gives us

$$p(\text{ILI}) = \frac{p(\text{visit})}{p(\text{visit} \mid \text{ILI})} \cdot p(\text{ILI} \mid \text{visit})$$

$p(\text{ILI} \mid \text{visit})$ is the CDC's reported percent ILI activity, for $p(\text{visit})$ we use the estimate from [40] of a weekly doctor visitation rate of 7.8% of the US population, and for $p(\text{visit} \mid \text{ILI})$ we use a base estimate of 27%, consistent with the findings from [22]. Once $p(\text{ILI})$ is calculated, we multiply $p(\text{ILI})$ by the population size to get a case count estimate within the population.

5.1.4 Visit Propensity Adjustment

We note that health care seeking behavior varies by region of the United States as shown in [22]. To better model these regional behavior differences, we adjust $p(\text{visit} \mid \text{ILI})$, the probability that a person with an influenza-like illness seeks medical attention, using regional baselines for the 2019-2020 influenza season [12].

Additionally, because our method estimates the increase in ILI visits due to the impact of COVID-19, we must distinguish an increase due to COVID-19 cases from an underlying increase in medical visit propensity in people with ILI symptoms. Due to the widespread alarm over the spread of COVID-19, it would not be unreasonable to expect a potential increase in ILI medical visits even in the hypothetical absence of true COVID-19 cases.

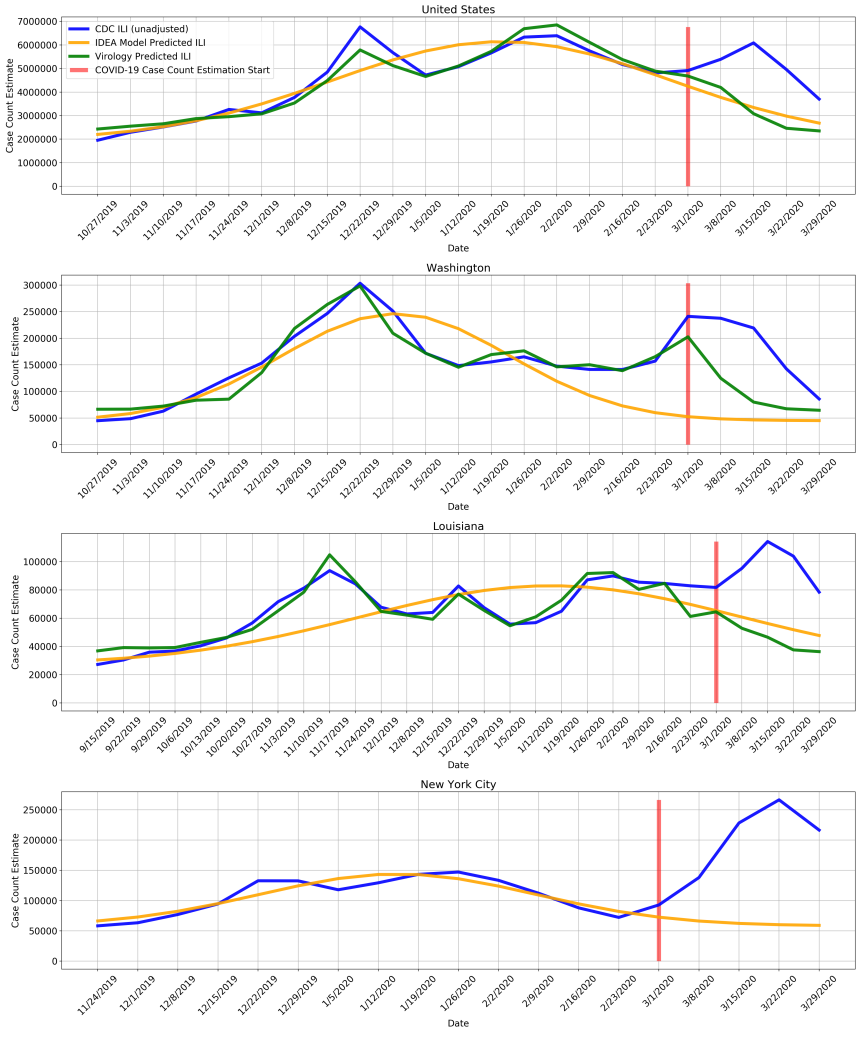
For this reason, we also explore increasing $p(\text{visit} \mid \text{ILI})$ from 27% to 35% to measure the possible effect of a change in health care seeking behavior due to COVID-19 media attention and panic. The increase of $p(\text{visit} \mid \text{ILI})$ to 35% is consistent with health care seeking behavior surveys done after the start of COVID-19 [20, 21]. The *Divergence* and *COVID Scaling* methods have *adjusted* versions which incorporate this shift as well as *unadjusted* versions that keep the baseline 27% propensity.

5.1.5 Estimating COVID-19 Case Counts

The ultimate goal is to estimate the true burden of COVID-19. The IDEA and virology predicted ILI case counts can be used to estimate CDC ILI had COVID-19 not occurred. In other words, the IDEA and virology predicted ILI can be used as counterfactuals when measuring the impact of COVID-19 on CDC measured ILI. The difference between the observed CDC measured ILI and the counterfactual (IDEA predicted ILI or virology predicted ILI) for a given week is then the estimate of COVID-19 case counts for that week. Fig. 4 shows example observed CDC measured ILI, IDEA model predicted ILI, and virology predicted ILI. The supplementary materials contain similar plots to Fig. 4 for all locations. For this method as well as the following two, we start

estimating COVID-19 case counts the week starting on March 1, 2020. We note that while the IDEA and virology ILI predictions tend to track CDC ILI well earlier in the flu season, after COVID-19 started to impact the United States there is a clear divergence between predictions and observed CDC ILI, with CDC ILI increasing while the counterfactual estimates decrease.

Figure 4: COVID-19 is treated as an intervention, and we measure COVID-19 impact on observed CDC ILI, using IDEA model predicted ILI and virology predicted ILI as counterfactuals. The difference between the higher observed CDC ILI and the lower IDEA model predicted ILI and virology predicted ILI is the measured impact of COVID-19. The impact directly maps to an estimate of COVID-19 case counts. Virology predicted ILI is omitted when virology data is not available.



5.2 Approach 2: COVID Scaling

This approach infers the COVID-19 fraction of the total ILI by extrapolating testing results obtained from the COVID Tracking Project [44], following the same reasoning as the Virology Di-

vergence method. That is,

$$C_t = \frac{C_t^+ \cdot I_t}{N_t^c}$$

where C_t^+, N_t^c, I_t denote positive COVID-19 tests, total COVID-19 specimens, and ILI visit counts respectively.

State-level testing results were aggregated to the weekly level and positive test percentages were computed using the positive and negative counts, disregarding pending tests. Positive test counts were adjusted for potential false negatives. There are varying estimates for the false negative rate for the RT-PCR used in COVID-19 tests, with some reports suggesting rates as high as 25-30% [54, 55]. We apply a 15% false negative rate in our analysis; repeating our analysis using a range of values from 5% to 25% yielded little difference in our estimates. On the other hand, COVID-19 testing is highly specific, so we assume no false positives. Then, the number of false negatives (FN) can be computed from the recorded (true) positives (TP) and the false negative rate (fnr) as

$$FN = TP \cdot \frac{fnr}{1 - fnr}$$

Because COVID-19 testing is sparse in many states, there are issues with zero or low sample sizes, as well as testing backlogs. Rather than taking the empirical positive test percentage (C_t^+/N_t^c), we first smoothed the percentages over time by taking convex combinations with the probabilities from the previous weeks, weighted by relative specimen count. This has a Bayesian posterior interpretation and is mathematically equivalent to computing probabilities using cumulative positive and total counts instead of in-week counts (for convenience, C_t^+ and N_t^c henceforth refer to these respective quantities). This helped but did not address all issues with case backlog, so we further smoothed the estimates using a Bayesian spatial model:

Denote p_{jt} as the probability that a given ILI patient in state j and week t has COVID-19. Under the condition that testing is applied uniformly, the COVID-19 status of patient i from state j in week t is

$$X_{jt}^{(i)} \sim Bernoulli(p_{jt})$$

Assuming COVID-19 status is independent in each ILI patient, the state testing results follow a Binomial distribution. We apply a spatial prior based on first-order conditional dependence:

$$p_{jt} \sim Beta(\alpha_{jt} N_{0t}, (1 - \alpha_{jt}) N_{0t})$$
$$\alpha_{jt} = \frac{1}{|\mathcal{N}_j|} \sum_{k \in \mathcal{N}_j} p_{kt}$$

where $\mathcal{N}_j$ are the neighbors of state j. The strength of the prior was specified by setting N_{0t} to be the number of total tests at the 5th quantile among all states in each week. Finally, we compute α_{jt}

by replacing each p_{kt} by their empirical estimates. Using the Beta-Binomial conjugacy we derive closed-form posterior mean estimates for p_{jt}:

$$\hat{p}_{jt} = \frac{C_{jt}^{+} + \alpha_{jt} N_{0t}}{N_{jt}^{c} + N_{0t}}$$

As previously explained, the weekly, state-level reported percent ILI were then multiplied by $\hat{p}_{jt}$ to get an estimate of the percent of medical visits that could be attributed to COVID-19. These values were subsequently scaled to the whole population using the same Bayes' rule method as described in *ILI Case Count Estimation* (4.2.3).

5.3 Approach 3: Mapping Mortality to COVID-19 Cases

Other studies have introduced methods to infer COVID-19 cases from COVID-19 deaths using (semi-)mechanistic disease models [15] or statistical curve-fitting based on assumptions of epidemic progression [16], but, to the best of our knowledge, no methods have been proposed to directly infer cases without either of these assumptions.

Mortality Map (*mMAP*) uses, under a Bayesian framework, reported deaths to predict previous true case counts. *mMAP* accounts for right-censoring (i.e. COVID-19 cases that are not resolved yet) by adapting previously used methods [13]. A study of clinical cases in Wuhan found that the time from hospitalization to death roughly follows a log-normal distribution with mean 13 and standard deviation 12.7 [26]. Using this distribution, a time series of reported deaths, D, and the age-adjusted IFR, we estimate the distribution of cases C, defined at the usual time of hospitalization, using an iterative Bayesian approach. We use Bayes' rule to define the probability that there was a case on day t given a death on day τ

$$p(\text{case on } t \mid \text{death on } \tau) = \frac{p(\text{death on } \tau \mid \text{case on } t) \cdot p(\text{case on } t)}{p(\text{death on } \tau)} \tag{1}$$

Let C_{d^*} denote the predicted distribution of when D are classified as cases (i.e. are hospitalized), C_d denote the predicted distribution of when D and future deaths are classified as cases (so adjusted for right-censoring), and t_{max} denote the most recent date with deaths reported. Let $p(\text{death on } \tau \mid \text{case on } t) = p(T = (\tau - t))$ denote the log-normal probability. *mMAP* performs the following steps:

1. Initialize the prior probability of a case on day t, $p_0(\text{case on } t)$, as uniform.

2. Repeat the following for each iteration i:

- Calculate $C_{d^*}^{(i)}$.

$$\begin{aligned} C_{d^*}^{(i)}(t) &= \sum_{\tau=t+1}^{t_{max}} D(\tau) \cdot p_{i-1}(\text{case on } t \mid \text{death on } \tau) \\ &= \sum_{\tau=t+1}^{t_{max}} D(\tau) \cdot \frac{p(T=(\tau-t)) \cdot p_{i-1}(\text{case on } t)}{\sum_{s=1}^{\tau-1} p(T=(\tau-s)) \cdot p_{i-1}(\text{case on } s)} \end{aligned} \tag{2}$$

  where the denominator is equivalent to $p(\text{death on } \tau)$ in (1).

- We estimate that the proportion $p(T \leq (t_{max}-t))$ of $C_d^{(i)}(t)$ have died by t_{max} and use this to adjust for right censoring.

$$C_d^{(i)}(t) = \frac{C_{d\cdot}^{(i)}(t)}{p(T \leq (t_{max}-t))} \tag{3}$$

- Update prior probabilities

$$p_i(\text{case on } t) = \frac{C_d^{(i)}(t)}{\sum C_d^{(i)}(t)} \tag{4}$$

- Repeat until $\frac{||C_d^{(i)} - C_d^{(i-1)}||}{||C_d^{(i)}||} \leq \epsilon$, where ϵ is a pre-specified tolerance level.

3. $C_d(t)$ represents the number of cases on day t that will lead to death. We scale this to estimate the number of all cases by divide by the IFR.

$$C(t) = \frac{C_d(t)}{IFR} \tag{5}$$

mMAP has connections to expectation-maximization, though further theoretical work is needed to establish the connection. We found that *mMAP* performs significantly better than using independent maximum-likelihood based estimates [i.e. solving $C_{d^*}(t) = \sum_{\tau=t+1}^{t_{max}} D(\tau) \cdot p(T=(\tau-t))$]. Supplementary section 4.2 demonstrates that *mMAP* successfully predicts cases in simulated and true scenarios using data from six countries. As well, supplementary section 4.1 demonstrates that if *mMAP* converges, which it does for every US state, C_d fully explains deaths under the assumed probability distribution (6), and that this satisfies the calculation of the fatality rate as presented in [13].

$$D(t) = \sum_{\tau=1}^{t-1} p(T=t-\tau) \cdot C_d(\tau) \ , \ \forall \ t \ \in 1..t_{max} \tag{6}$$

The IFR for each state is calculated using the age-stratified fatality rate [56], which estimates IFR of all cases - symptomatic and asymptomatic, and the population age structure provided by the US census [45].

5.3.1 Accounting for Unreported COVID-19 Deaths

While *mMAP* assumes all COVID-19 deaths are reported, some deaths will be unreported because of a limited testing and false negative results [57]. Previous research on the H1N1 epidemic estimated that the ratio of lab-confirmed deaths to actual deaths caused by the disease was 1:7 nationally [58] and 1:15 globally [59]. While the actual rate of under-reporting is unknown, we include an adjustment, mMAP^{adj}, that uses an estimate of unreported COVID-19 deaths based on reports of excess pneumonia deaths (note that this is similar to the divergence methods, except that those measure excess ILI visits). mMAP^{adj} assumes that excess pneumonia deaths in March 2020 were due to COVID-19.

The CDC reports weekly pneumonia deaths, $D_P(w)$, expected weekly pneumonia deaths based on a model of historical trends, $\mathbb{E}[D_P(w)]$, and deaths that are classified as pneumonia and COVID-19, $D_{P \cap COV}(w)$ [60, 61]. We estimate that the number of un-classified COVID-19 deaths each week, $D_U(w)$, is the following:

$$D_U(w) = D_P(w) - \mathbb{E}[D_P(w)] - D_{P \cap COV(w)} \tag{7}$$

This results in 355, 438, 605, and 540 nationwide excess deaths for the four weeks from March 1 to March 28, which is the most recent data at the time of writing this paper. To account for missing data in recent weeks, We assume that the weekly number of excess deaths remains constant after March 21, i.e. that there were 540 excess deaths the weeks of March 29 - April 4 and April 5 - April 11. Since the expected pneumonia deaths are not available at the state level, excess deaths are calculated nationally and then attributed to each state s in proportion to the number of pneumonia deaths in that state. Then, the weekly excess deaths are evenly distributed across each day of the week.

$$\begin{aligned} D_U^s(w) &= D_U(w) \cdot \frac{D_P^s(w)}{\sum_{s \in S} D_P^s(w)} \\ D^{s,adj}(t) &= D^s(t) + \frac{1}{7} D_U^s(w), \text{ where } t \in w \end{aligned} \tag{8}$$

5.4 Aggregation of Estimates

The divergence-based methods predict national COVID-19 prevalence directly using national ILI data. *mMAP* predicts national prevalence using national death data, while *COVID Scaling* estimates national prevalence by aggregating the case estimates from each state.

The *Divergence* and *COVID Scaling* methods provide separate case estimates for each week within the studied period, which are summed to the total cumulative case estimates. *mMAP* provides daily estimates which are further aggregated by week.

Approach	*Divergence*	*COVID Scaling*	*mMAP*
Brief Description	Treat COVID-19 case count estimation as a causal inference problem. COVID's impact on ILI activity is measured using as controls an Incidence Decay with Exponential Adjustment model as well as influenza testing statistics.	Extrapolate state-level positive test percentages for COVID-19 to the weekly ILI data to estimate COVID-19 proportion in medical visits, then scale to the whole population.	Using reported COVID-19 deaths, the IFR, and a distribution of time from cases to deaths, predict the latent case distribution.
Data Input	ILI activity and influenza test results.	ILI activity and COVID-19 test results.	COVID-19 deaths.
Model Assumptions	1. The divergence between predicted ILI activity for the 2019-2020 season and measured ILI activity after the start of the COVID-19 pandemic can be attributed to COVID-19. 2. Scaling from ILI to population is reliable.	1. COVID-19 test reports accurately represent the pool of weekly ILI visits. 2. Delayed test reporting does not significantly affect positive test proportions after applying smoothing 3. Scaling from ILI to population is reliable.	1. All COVID-19 deaths are reported (mMAP) or explained by excess pneumonia deaths (mMAP^{adj}). 2. The distribution of time from cases to death is log-normal. 3. The age-stratified IFR is the same as reported in [56].
Expected Bias	This method can be sensitive to model fit and changes in healthcare seeking behavior among symptomatic individuals.	ILI visits and COVID-19 tests may capture different segments of the sick population.	May underestimate cases as many COVID-19 related deaths may go unreported or untested.

Table 2: Comparing the three approaches to estimate COVID-19 cases in the US.

References

[1] World Health Organization. Report of the who-china joint mission on coronavirus disease 2019.

[2] Centers for Disease Control and Prevention. Locations with confirmed covid-19 cases.

[3] Marc Lipsitch and Mauricio Santillana. Enhancing situational awareness to prevent infectious disease outbreaks from becoming catastrophic. *Global Catastrophic Biological Risks*, pages 59–74, 2019.

[4] Sheila Kaplan Sheri Fink Katie Thomas Michael D. Shear, Abby Goodnough and Noah Weiland. The lost month: How a failure to test blinded the u.s. to covid-19.

[5] Arjun K. Manrai and Kenneth D. Mandl. Covid-19 testing: overcoming challenges in the next phase of the epidemic.

[6] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. *Eurosurveillance*, 25(10), 2020.

[7] Tara John. Iceland lab's testing suggests 50% of coronavirus cases have no symptoms. *CNN*, Apr 2020.

[8] Michael Day. Covid-19: identifying and isolating asymptomatic people helped eliminate virus in italian village, 2020.

[9] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, 2020.

[10] Justin Kaashoek and Mauricio Santillana. Covid-19 positive cases, evidence on the time evolution of the epidemic or an indicator of local testing capabilities? a case study in the united states. *Available at SSRN: https://ssrn.com/abstract=3574849, April*, 2020.

[11] Centers for Disease Control and Prevention. Fluview.

[12] Centers for Disease Control and Prevention. U.s. influenza surveillance system: Purpose and methods.

[13] 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.

[14] TW Russell, J Hellewell, S Abbott, CI Jarvis, K van Zandvoort, et al. Using a delay-adjusted case fatality ratio to estimate under-reporting. *Centre for Mathematical Modeling of Infectious Diseases Repository*, 2020.

[15] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Helen Coupland, et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries. 2020.

[16] IHME COVID, Christopher JL Murray, et al. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv*, 2020.

[17] Graham C. Gibson Estee Cramer Caitlin M. Rivers Nicholas G. Reich, Evan L. Ray. Looking for evidence of a high burden of covid-19 in the united states from influenza-like illness data.

[18] Justin D Silverman and Alex D Washburne. Using ili surveillance to estimate state-specific case detection rates and forecast sars-cov-2 spread in the united states. *medRxiv*, 2020.

[19] David N Fisman, Tanya S Hauck, Ashleigh R Tuite, and Amy L Greer. An idea for short term outbreak projection: nearcasting using the basic reproduction number. *PloS one*, 8(12), 2013.

[20] Jonathan Rothwell. Estimating covid-19 prevalence in symptomatic americans.

[21] Pascal Geldsetzer. Knowledge and perceptions of covid-19 among the general public in the united states and the united kingdom: A cross-sectional online survey.

[22] Kristin Baltrusaitis, Alessandro Vespignani, Roni Rosenfeld, Josh Gray, Dorrie Raymond, and Mauricio Santillana. Differences in regional patterns of influenza activity across surveillance systems in the united states: Comparative evaluation. *JMIR Public Health and Surveillance*, 5(4):e13403, 2019.

[23] Temet M McMichael. Covid-19 in a long-term care facility—king county, washington, february 27–march 9, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69, 2020.

[24] Sarah Kliff and Julie Bosman. Official counts understate the u.s. coronavirus death toll. *The New York Times*, Apr 2020.

[25] Gwynne Hogan. Death count expected to soar as nyc says it will begin reporting probable covid deaths in addition to confirmed ones. *Gothamist*, Apr 2020.

[26] Natalie M Linton, Tetsuro Kobayashi, Yichi Yang, Katsuma Hayashi, Andrei R Akhmetzhanov, Sung-mok Jung, Baoyin Yuan, Ryo Kinoshita, and Hiroshi Nishiura. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right

truncation: a statistical analysis of publicly available case data. *Journal of clinical medicine*, 9(2):538, 2020.

[27] Shihao Yang, Mauricio Santillana, and Samuel C Kou. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.

[28] Fred S Lu, Mohammad W Hattab, Cesar Leonardo Clemente, Matthew Biggerstaff, and Mauricio Santillana. Improved state-level influenza nowcasting in the united states leveraging internet-based data and network approaches. *Nature communications*, 10(1):1–10, 2019.

[29] Shihao Yang, Samuel C Kou, Fred Lu, John S Brownstein, Nicholas Brooke, and Mauricio Santillana. Advances in using internet searches to track dengue. *PLoS computational biology*, 13(7), 2017.

[30] Mauricio Santillana, AT Nguyen, Tamara Louie, Anna Zink, Josh Gray, Iyue Sung, and John S Brownstein. Cloud-based electronic health records for real-time, region-specific influenza surveillance. *Scientific reports*, 6:25732, 2016.

[31] Mauricio Santillana, Elaine O Nsoesie, Sumiko R Mekaru, David Scales, and John S Brownstein. Using clinicians' search query data to monitor influenza epidemics. *Clinical Infectious Diseases*, 59(10):1446, 2014.

[32] Aaron C Miller, Inder Singh, Erin Koehler, and Philip M Polgreen. A smartphone-driven thermometer application for real-time population-and individual-level influenza surveillance. *Clinical Infectious Diseases*, 67(3):388–397, 2018.

[33] World Health Organization. Global influenza strategy 2019–2030.

[34] Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, and Kaiyuan Sun. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 2020.

[35] Shengjie Lai, Nick W Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R Floyd, Amy Wesolowski, Chi Zhang, Xiangjun Du, Hongjie Yu, et al. Effect of non-pharmaceutical interventions for containing the covid-19 outbreak: an observational and modelling study. *medRxiv*, 2020.

[36] Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O Nsoesie, and John S Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10), 2015.

[37] Mark S Smolinski, Adam W Crawley, Kristin Baltrusaitis, Rumi Chunara, Jennifer M Olsen, Oktawia Wójcik, Mauricio Santillana, Andre Nguyen, and John S Brownstein. Flu near you: crowdsourced symptom reporting spanning 2 influenza seasons. *American journal of public health*, 105(10):2124–2130, 2015.

[38] John S Brownstein, Shuyu Chu, Achla Marathe, Madhav V Marathe, Andre T Nguyen, Daniela Paolotti, Nicola Perra, Daniela Perrotta, Mauricio Santillana, Samarth Swarup, et al. Combining participatory influenza surveillance with modeling and forecasting: Three alternative approaches. *JMIR public health and surveillance*, 3(4):e83, 2017.

[39] Dianbo Liu, Leonardo Clemente, Canelle Poirier, Xiyu Ding, Matteo Chinazzi, Jessica T Davis, Alessandro Vespignani, and Mauricio Santillana. A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019*, 2020.

[40] Wan Yang, Marc Lipsitch, and Jeffrey Shaman. Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*, 112(9):2723–2728, 2015.

[41] Stephen Kissler, Christine Tedijanto, Edward Goldstein, Yonatan Grad, and Marc Lipsitch. Projecting the transmission dynamics of sars-cov-2 through the post-pandemic period. 2020.

[42] New York Times. Data from the new york times, based on reports from state and local health agencies, 2020. `https://github.com/nytimes/covid-19-data`.

[43] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 2020. Data obtained from `https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`.

[44] The covid tracking project. `https://covidtracking.com/`.

[45] U.S. Census Bureau. Annual estimates of the civilian population by single year of age and sex for the united states and states: April 1, 2010 to july 1, 2018, 2019. `https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html`.

[46] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274, 2015.

[47] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[48] Mauricio Santillana, Ashleigh Tuite, Tahmina Nasserie, Paul Fine, David Champredon, Leonid Chindelevitch, Jonathan Dushoff, and David Fisman. Relatedness of the incidence decay with exponential adjustment (idea) model,"farr's law" and sir compartmental difference equation models. *Infectious disease modelling*, 3:1–12, 2018.

[49] Margaretha Annelie Vink, Martinus Christoffel Jozef Bootsma, and Jacco Wallinga. Serial Intervals of Respiratory Infectious Diseases: A Systematic Review and Analysis. *American Journal of Epidemiology*, 180(9):865–875, 10 2014.

[50] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020.

[51] Marc Lipsitch, Ted Cohen, Ben Cooper, James M Robins, Stefan Ma, Lyn James, Gowri Gopalakrishna, Suok Kai Chew, Chorh Chuan Tan, Matthew H Samore, et al. Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300(5627):1966–1970, 2003.

[52] Sen Pei, Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Forecasting the spatial transmission of influenza in the united states. *Proceedings of the National Academy of Sciences*, 115(11):2752–2757, 2018.

[53] Andrew Forney Carlos Cinelli and Judea Pearl. A crash course in good and bad control.

[54] Chunxia Qin, Fang Liu, Tzu-Chen Yen, and Xiaoli Lan. 18 f-fdg pet/ct findings of covid-19: a series of four highly suspected cases. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–6, 2020.

[55] Yang Yang, Minghui Yang, Chenguang Shen, Fuxiang Wang, Jing Yuan, Jinxiu Li, Mingxia Zhang, Zhaoqin Wang, Li Xing, Jinli Wei, et al. Laboratory diagnosis and monitoring the viral shedding of 2019-ncov infections. *medRxiv*, 2020.

[56] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, Gina Cuomo-Dannenburg, et al. Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. 2020.

[57] Emma Brown, Beth Reinhard, and Aaron Davis. Coronavirus death toll: Americans are almost certainly dying of covid-19 but being left out of the official count. *the Washington Post*, Apr 2020.

[58] Vivek Charu, Lone Simonsen, Roger Lustig, Claudia Steiner, and Cécile Viboud. Mortality burden of the 2009-10 influenza pandemic in the united states: improving the timeliness of

influenza severity estimates using inpatient mortality records. *Influenza and other respiratory viruses*, 7(5):863–871, 2013.

[59] Fatimah S Dawood, A Danielle Iuliano, Carrie Reed, Martin I Meltzer, David K Shay, Po-Yung Cheng, Don Bandaranayake, Robert F Breiman, W Abdullah Brooks, Philippe Buchy, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza a h1n1 virus circulation: a modelling study. *The Lancet infectious diseases*, 12(9):687–695, 2012.

[60] Centers for Disease Control and Prevention. Pneumonia and influenza mortality surveillance from the national center for health statistics mortality surveillance system. `https://gis.cdc.gov/grasp/fluview/mortality.html` accessed April 12, 2020.

[61] Centers for Disease Control and Prevention. Provisional death counts for coronavirus disease (covid-19). `https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm` accessed April 12, 2020.

Supplementary Materials

April 18, 2020

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 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.

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

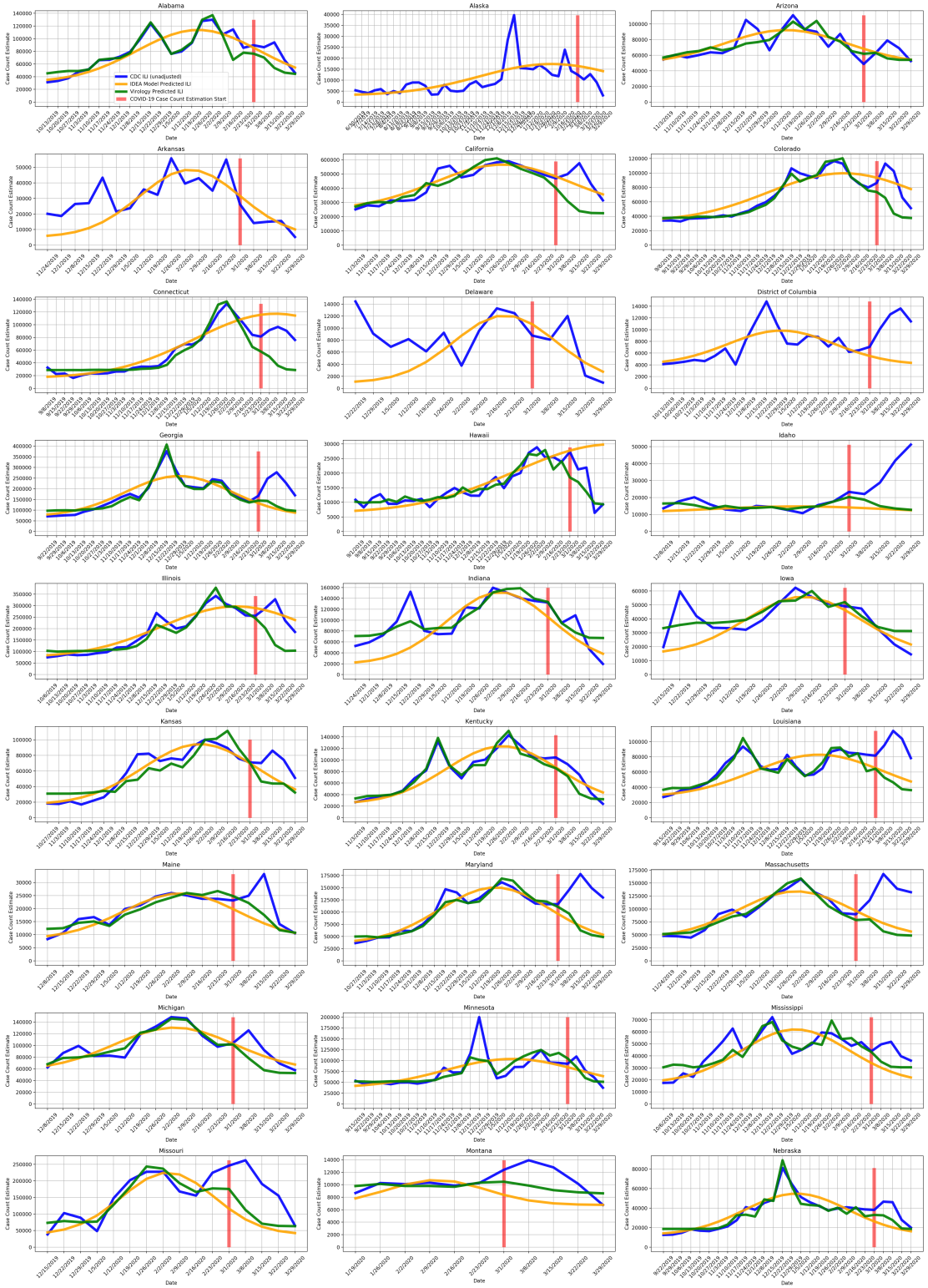
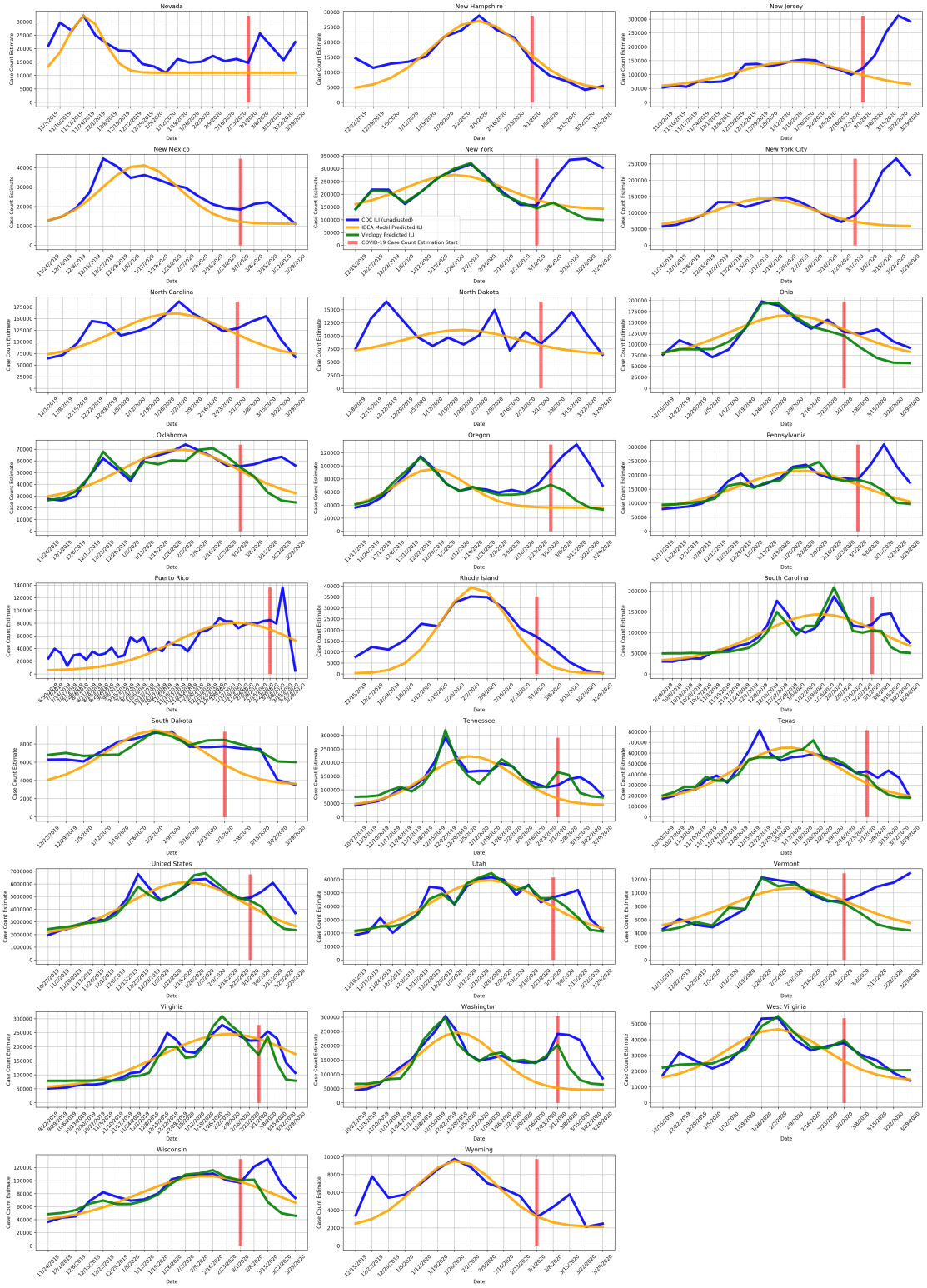


Figure 2: Divergence model fits for second half of locations.



2 Time Series Plots for All Methods

Figures 3 and 4 show the cumulative estimated counts for each week over our study period, compared with cumulative reported counts, in each location in the United States. The solid and dotted lines indicate adjusted and unadjusted methods, respectively.

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

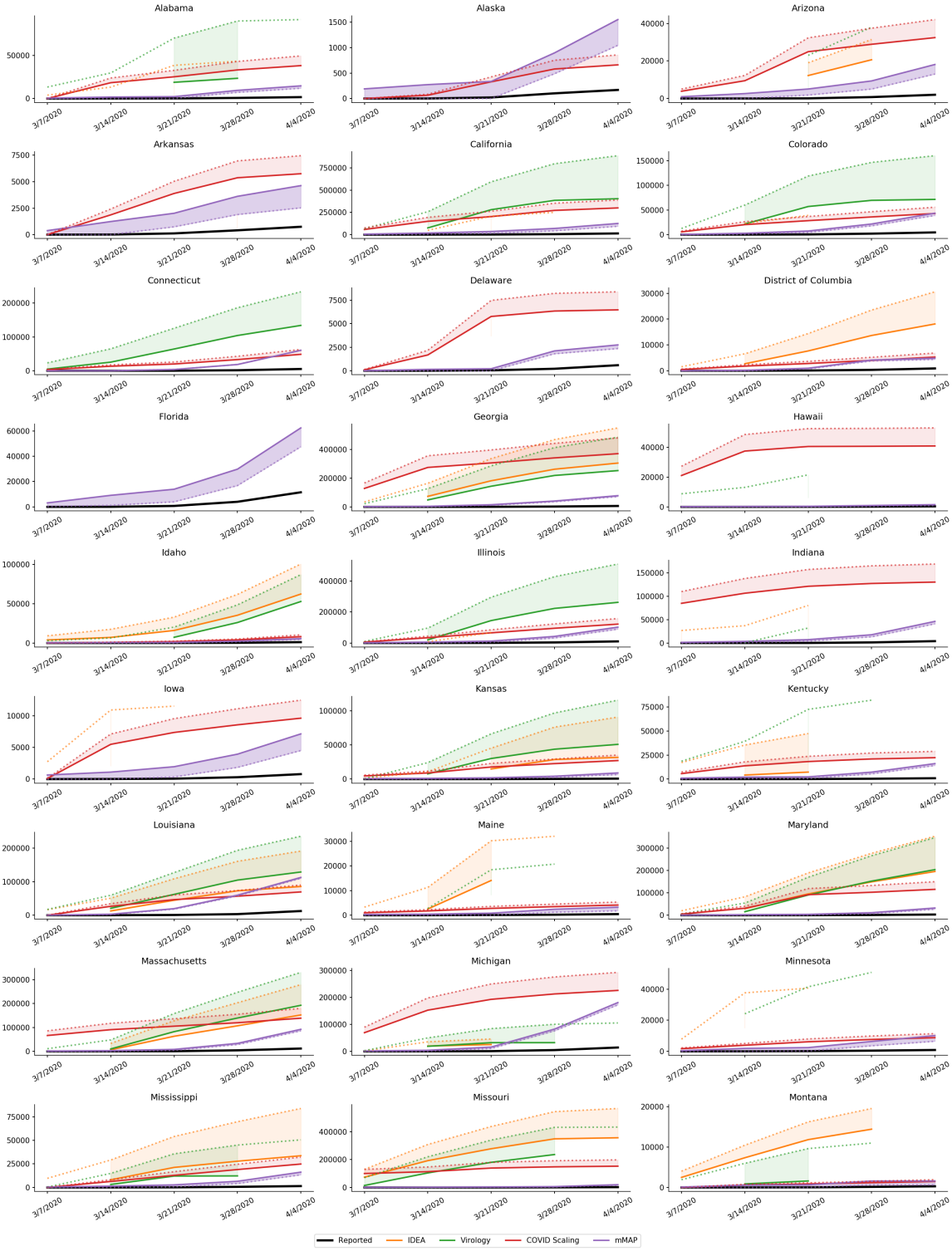
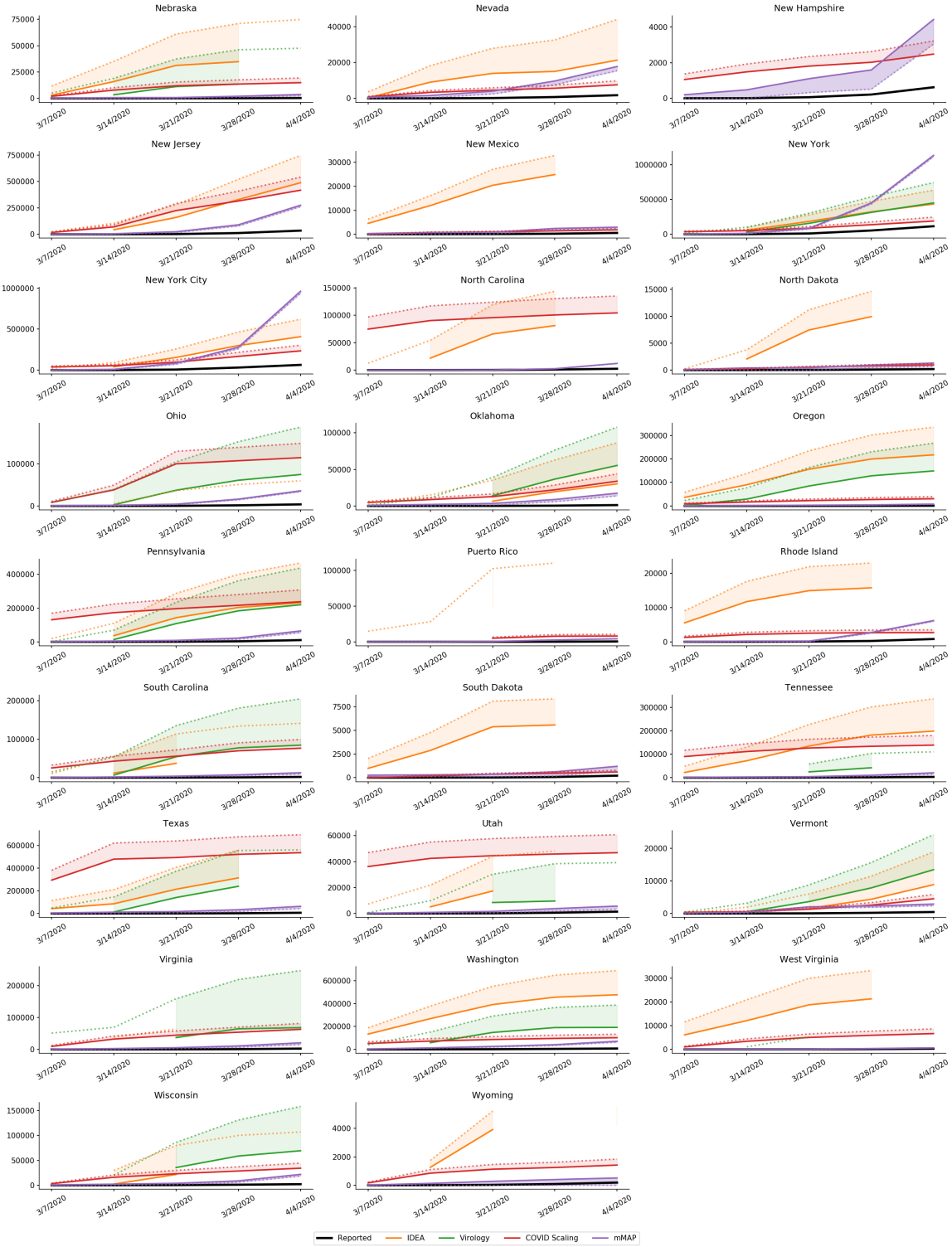


Figure 4: Cumulative case time series for second half of locations.



3 Virology-based Estimation

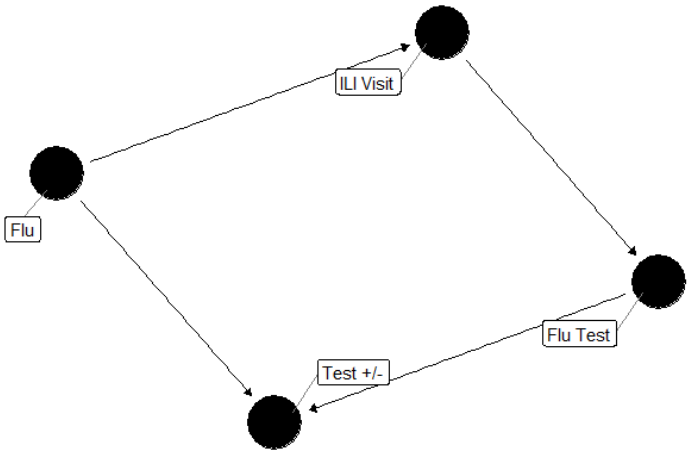


Figure 5: Causal DAG affecting flu positive results.

Both the virology-based *Divergence* model and the *COVID Scaling* method rely on the extrapolation of positive testing data to the actual prevalence 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 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.

4.1.1 Satisfying Infected Fatality Ratio Calculation

The authors of [1] propose an unbiased estimator of the IFR as the following. In the paper, they calculated the case fatality ratio, defined as the proportion of deaths per reported case, and defined

C as reported cases, but here we are defining C as the total infections, so the IFR should be used instead of CFR

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

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 IFR by showing that from our estimates of C, the RHS above equals the LHS. Note that in our formulation of C, $C_d = IFR \cdot C$, since C_d is the time series of cases that end up in death, and C is the time series of all cases.

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

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 IFR, and distribution of time from case onset to death, this method can accurately predict the unobserved case time series.

4.2 Simulated and Empirical Validation

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

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

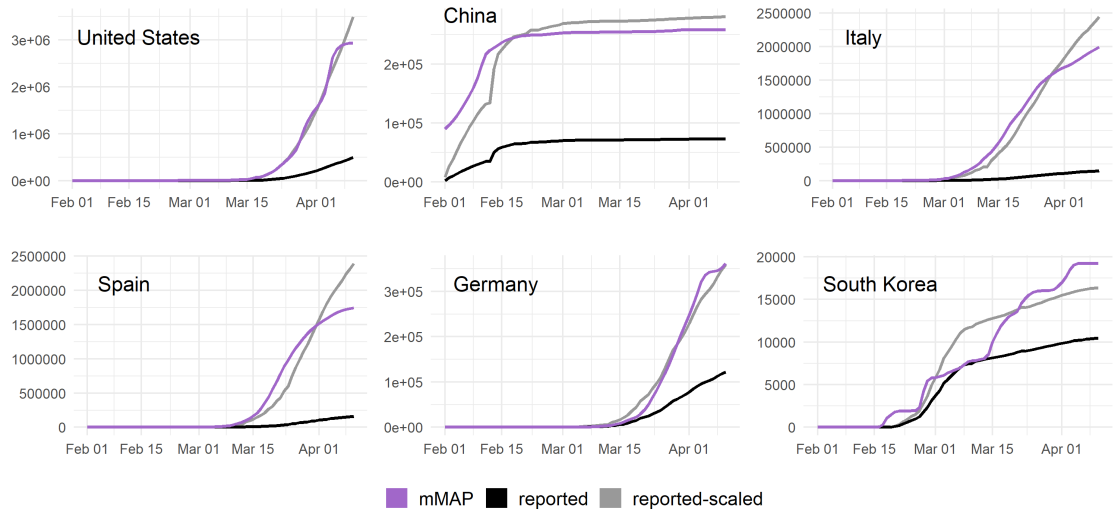
While the scales differ, the trends of predicted cases generally follow the trends of reported cases; this is especially the case where there is linear to exponential growth as in Germany, Italy, Spain,

and the United States. For these countries, at the beginning of April, the relative slope for reported cases is higher than for *mMAP* predictions; this could be a result of increasing case detection at the start of April. In South Korea, *mMAP* does not match the trend as well, likely because of the sharp change in case growth after the first week in March.

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 case onset to death and an IFR of 0.013. 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 IFR.

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.