medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Using ILI surveillance to estimate state-specific case detection rates and forecast SARS-CoV-2 spread in the United States

Justin D. Silverman[1,2,4] and Alex D. Washburne[3,4]

[1] *College of Information Science and Technology, Penn State University*
[2] *Department of Medicine, Penn State University*
[3] *Department of Microbiology and Immunology, Montana State University*
[4] *Both authors contributed equally to this manuscript*

Abstract

Detection of SARS-CoV-2 infections to date has relied on RT-PCR testing. However, a failure to identify early cases imported to a country, bottlenecks in RT-PCR testing, and the existence of infections which are asymptomatic, sub-clinical, or with an alternative presentation than the standard cough and fever have resulted in an under-counting of the true prevalence of SARS-CoV-2. Here, we show how publicly available CDC influenza-like illness (ILI) outpatient surveillance data can be repurposed to estimate the detection rate of symptomatic SARS-CoV-2 infections. We find a surge of non-influenza ILI above the seasonal average and show that this surge is correlated with COVID case counts across states. By quantifying the number of excess ILI patients in March relative to previous years and comparing excess ILI to confirmed COVID case counts, we estimate the symptomatic case detection rate of SARS-CoV-2 in the US to be 1/100 to 1/1000. This corresponds to approximately 10 million presumed symptomatic SARS-CoV-2 patients across the US during the week starting on March 15, 2020. Combining excess ILI counts with the date of onset of community transmission in the US, we also show that the early epidemic in the US was unlikely to be doubling slower than every three days. Together these results suggest a conceptual model for the COVID epidemic in the US in which rapid spread across the US are combined with a large population of infected patients with presumably mild-to-moderate clinical symptoms. We emphasize the importance of testing these findings with seroprevalence data, and discuss the broader potential to repurpose outpatient time series for early detection and understanding of emerging infectious diseases.

1 Introduction

The ongoing SARS-CoV-2 pandemic continues to cause tremendous morbidity and mortality around the world [1, 2]. Regional preparation for the pandemic requires forecasting the growth rate of the epidemic, the timing of the peak, the demand for hospital resources, and the degree to which current policies may curtail the epidemic, all of which benefit from accurate estimates of the true prevalence of the virus within a population [3]. Confirmed cases are thought to be underestimates of true prevalence due to some unknown combination of patients not reporting for testing, testing not being conducted, and false-negative test results. Estimating the true prevalence informs the scale of upcoming hospital, ICU and ventilator surges, the proportion of individuals who are susceptible to contracting the disease, and estimates of key epidemiological parameters such as the epidemic growth rate and the fraction of infections which are sub-clinical.

The current literature suggests that the predominant symptoms associated with COVID are fever, cough and sore-throat; that is, patients often present with an influenza-like illness (ILI) yet test negative for influenza [4, 5]. With many COVID patients having a similar presentation as patients with influenza, existing surveillance networks in place for tracking influenza could be used to help track COVID.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Here, we quantify background levels of non-influenza ILI over the past 10 years and identify a recent surge of non-influenza ILI starting the first week of March, 2020. This surge of excess

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

ILI correlates with known patterns of SARS-CoV-2 spread across states within the US, suggesting the surge is unlikely to be due to other endemic respiratory pathogens, yet is orders of magnitude larger than the number of confirmed COVID cases reported. Together this suggests that the true prevalence of SARS-CoV-2 within the US is much larger than currently appreciated and that even the highest symptomatic case detection rates are likely lower than 3% corresponding to approximately 9 million new ILI cases due to SARS-CoV-2. Our analysis provides empirical corroboration of previous hypotheses of substantial undocumented cases [6] yet places the estimated undocumented case rate an order of magnitude higher than prior reports [6]. Moreover, these updated prevalence estimates predict that epidemic doubling times greater than 3.5 days [7, 8] would be unable to account for the magnitude of the ILI surge. We test our hypothesis of sub 3-day doubling times in the US by analyzing both state and national COVID surveillance data, finding a broad agreement of doubling time less than 3.5 days in both confirmed COVID case counts and documented COVID deaths. Our findings highlight probable trajectories of the US epidemic and provide evidence for a conceptual model for COVID spread in the US in which more rapid spread than previously reported is coupled with a larger undiagnosed population to give rise to currently observed trends.

2 Results

2.1 Influenza like illness surge

We identified excess ILI cases by first subtracting cases due to influenza and then subtracting the seasonal signal of non-influenza ILI (Figure 1). Many states, including Washington, New York, Oregon, Pennsylvania, Maryland, Colorado, New Jersey, and Louisiana, have had a recent surge in number of non-influenza ILI cases far in excess of seasonal norms. For example, in the second week of March, 2020, Oregon saw 50% higher non-influenza ILI than it had ever seen since the inception of the ILINet surveillance system within the US. We find that with 95% probability approximately 4% of all outpatient visits in Oregon during this time were for ILI that could not be explained by either influenza or the normal seasonal variation of respiratory pathogens. We find that as the seasonal surge of endemic non-influenza respiratory pathogens declines, this excess ILI correlates more strongly with state-level patterns of newly confirmed COVID cases suggesting that this surge is a reflection of ILI due to SARS-CoV-2 (Pearson $\rho = 0.8$ and $p < 10^{-10}$ for the last two weeks; Figure S1). To equate this surge to state-wide or national case counts, we assume that the average number of patients seen per week by sentinel providers is representative of their respective states that week. Using this assumption, the total excess non-influenza ILI across the US was approximately 9.2 million excess individuals in the week starting March 15th, 2020 compared to the same week in 2019 (95% credible interval of 8.0-10.1 million).

2.2 Symptomatic Case Detection Rate

The rate at which SARS-CoV-2+ patients with ILI symptoms are identified as having COVID varies by state and over time (Figure S2). Our estimated symptomatic case detection rates have been increasing over the month of March, which can be expected given increases in testing capacity across the US since the February 28 detection of community transmission in Washington State. For the latest week ending March 14, COVID cases in the states with the highest estimated symptomatic case detection rate (Washington, Nevada, and Michigan) are only capturing approximately 1% of ILI surges in those states (Figure 2). Across the entire US we find the average symptomatic case detection rate to be 0.75% (95% credible interval 0.59%-1.0%).

2.3 Epidemic Growth Rates and Clinical Rates

The true prevalence of SARS-CoV-2 is unknown. However, if we assume the excess non-influenza ILI is almost entirely due to SARS-CoV-2, an assumption that becomes more valid as the virus becomes more prevalent, we can use the excess non-influenza ILI to understand the constraints and mutual dependence of exponential growth rates, the rate of subclinical infections, and the time between the onset of infectiousness and a patient reporting as ILI Figure 3. With a January 15 start date of the US epidemic [9], allowing early stochasticity from start-time to the onset of regular exponential growth, we find that it's impossible to explain the ILI surge with an epidemic whose

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

doubling time is longer than 3.5-days, as such slow growth scenarios fail to reach the observed excess ILI even when accounting for possible asymptomatic cases.

We tested our hypothesized constraint that doubling times across the US would be shorter than 3.5 days by analyzing the doubling time confirmed COVID cases and deaths due to COVID. We find that both confirmed case data and COVID fatality data support our estimated doubling-time bound derived from ILI data. Across the entire US, the doubling rate for deaths due to COVID is 2.520 days (± 0.001, p-value of test that doubling rate is less than 3.5 days approximately 0) and the doubling time for confirmed cases is 2.455 days (± 0.001, p-value of test that doubling rate is less than 3.5 days approximately 0). Under a 4-day lag from the onset of infectiousness to reporting as ILI, these doubling times capture the ILI surge with clinical rates of 4.4% and 3.3%, respectively.

3 Discussion

Here we use outpatient ILI surveillance data from around the US to estimate the prevalence of SARS-CoV-2+. We find a clear, anomalous surge in ILI outpatients during the COVID epidemic that correlates with the progression of the epidemic across the US. The surge of non-influenza ILI outpatients is much larger than the number of confirmed case in each state, providing evidence of large numbers of symptomatic probable COVID cases that remain undetected. Moreover, this excess ILI surge allows us to develop a bound for the slowest epidemic doubling time, and the lowest clinical rate that would still be consistent with the observed surge. We test whether or not observed epidemic doubling times are shorter than our upper-bound and find that the observed doubling time for both confirmed cases and deaths is faster than 3 days within the US. Together, the surge in ILI and analysis of doubling times suggest that SARS-CoV-2 has spread rapidly throughout the US since it's January 15th start date and is likely accompanied by a large undiagnosed population of potential COVID outpatients with presumably milder clinical symptoms than would be thought based on prior studies of SARS-CoV-2+ inpatients.

Our study has several limitations. First, the observed ILI surge may represent more than just SARS-CoV-2 infected patients. A second epidemic of a non-seasonal pathogen that presents with ILI or changing patient behaviors causing higher rates of presentation for typical seasonal ILI could confound our estimates of ILI due to SARS-CoV-2. Alternatively, it is also possible that our use of ILI data has underestimated the prevalence of SARS-CoV-2 within the US. While early clinical reports focused on cough and fever as the dominant features of COVID [5], other reports have documented digestive symptoms as the complaint affecting up to half of patients with laboratory-confirmed COVID [11], and alternative presentations, including asymptomatic or unnoticeable infections, could result in ILI surges underestimating SARS-CoV-2 prevalence.

Additionally, our models have several limitations. First we assume that ILI prevalence within states can be scaled to case counts at the state level. This is based on the assumption that the average number of cases seen by sentinel providers in a given week is representative of the average number of patients seen by all providers within that state in a given week. Errors in this assumptions would cause proportional errors in our estimated case counts and symptomatic case detection rate. Second, our epidemic models are crude, US-wide SEIR models varying by growth rate alone and as such do not capture regional variation or intervention-induced changes in transmission. Our models were used to estimate growth rates from ILI for testing with COVID data and to estimate the mutual dependency of growth rate, the lag between the onset of infection and presentation to a doctor, and clinical rates; these models were not intended to be fine-grained forecasts for municipality hospital burden and other common goals for COVID models. Finer models with regional demographic, and case-severity compartments are needed to translate our range of estimated prevalence, growth rate, and clinical rates into actionable models for public health managers.

While an ILI surge tightly correlated with COVID case counts across the US strongly suggests that SARS-CoV-2 has potentially infected millions in the US, laboratory confirmation of our hypotheses are needed to test our findings and guide public health decisions. Our conceptual model for the epidemic with the US makes clear and testable predictions. Our model would suggest relatively high rates of community seropositivity in states that have already seen an ILI surge. A study of ILI patients from mid-march who were never diagnosed with COVID could test our model's predictions about the number and regional prevalence of undetected COVID cases presenting with ILI during that time. If seroprevalence estimates are consistent with our estimated prevalence from

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

these ILI analyses, it would strongly suggest lower case severity rates for COVID and indicate the value of ILI and other public time-series of outpatient illness in facilitating early estimates of crucial epidemiological parameters for rapidly unfolding, novel pandemic diseases. Since not all novel pandemic diseases are expected to present with influenza-like symptoms, surveillance of other common presenting illnesses in the outpatient setting could provide a vital tool for rapidly understanding and responding to novel infectious diseases.

4 Methods

In what follows, let i index state i and let t index week t (with $t = 0$ referring to October 3, 2010; the start of ILINet surveilance). Our analysis centers around decomposing the probability of testing positive for COVID, δ into the product of the symptomatic case detection rate for ILI patients, δ^s, and the probability that a COVID patients presents to the clinic with ILI, δ^c.

4.1 Data Sources

Since 2010 the CDC has maintained ILINet for weekly influenza surveillance. Each week approximately 2,600 enrolled providers distributed throughout all 50 states as well as Puerto Rico, the District of Columbia and the US Virgin Islands, report the total number of patient encounters n_{it} and the total number of which met criteria for influenza-like illness (ILI – defined as a temperature 100F [37.8C] or greater, and a cough or sore-throat without a known cause of than influenza; y_{it}) [12]. Let d_{it} denote the number of reporting providers in state i in week t. For scale, in the 2018-2019 season ILINet reported approximately 60 million outpatient visits. Coupled to these data are weekly state-level reports from clinical and public health labs detailing the number of patient samples tested for influenza n_{it}^{flu} as well as the number of these samples which are positive for influenza y_{it}^{flu}. Therefore ILINet data can be thought of as a weekly state-level time-series representing the superimposed prevalence of various viruses which can cause ILI. ILINet data was obtained through the CDC FluView Interactive portal [13].

In addition to ILINet data, US State population data for the 2020 year was downloaded from https://worldpopulationreview.com/states/. The number of primary care providers in each state per 100,000 residents b was obtained from Becker's Hospital Review [14]. COVID confirmed case counts were obtained from The New York Times' database maintained at https://github.com/nytimes/covid-19-data. This dataset contains the daily cumulative confirmed case count for COVID for each state z_{il} for day l.

4.2 Data Processing

Within the ILINet dataset, New York City and New York were summed into a combined New York variable representing both New York city and the surrounding state. Due to incomplete data in one or more of the data-sources described above the Virgin Islands, Puerto Rico, The Commonwealth of the Northern Mariana Islands, and Florida were excluded from subsequent analysis. In addition, daily cumulative confirmed COVID cases were converted to weekly counts of new cases by

$$\tilde{z}_{it} = \sum_{l \in t} z_{it} - z_{i(t-1)}.$$

4.3 Extracting non-influenza ILI signal

To subtract influenza signal from y_{it} we assume that the population of patients with ILI within a state are the same population that are potentially tested for influenza. This assumption allows us to calculate the number of non-influenza ILI cases as

$$\tilde{y}_{it} = \left(1 - \frac{y_{it}^{flu}}{n_{it}^{flu}}\right) y_{it}.$$

The resulting time-series $\tilde{y}_{it}$ are shown in Figure S4.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

4.4 Identifying ILI Surges

We identified ILI surges in $\tilde{y}_{it}$ by training a model on $\tilde{y}_{it}$ for all data prior to July 21, 2019. We then used this model to predict the prevalence of non-influenza ILI ($\hat{\pi}_{it}$) for dates after and including July 21, 2019. We calculated the ILI surge as the difference between the observed proportion of non-influenza ili $\tilde{y}_{it}/n_{it}$ and $\hat{\pi}_{it}$.

More specifically, to account for variation in the number of reporting providers, we trained the following binomial logistic-normal model

$$\tilde{y}_{it} \sim \text{Binomial}(\pi_{it}, n_{it}) \tag{1}$$

$$\pi_{it} = \frac{\exp(\eta_{it})}{1+\exp(\eta_{it})} \tag{2}$$

$$\eta_{it} \sim N(\lambda_i(t), \sigma^2) \tag{3}$$

$$\lambda_i(t) \sim \mathcal{GP}(\theta, \sigma^2\Gamma) \tag{4}$$

$$\sigma^2 \sim \text{InverseGamma}(\upsilon, \xi) \tag{5}$$

$$\theta(t) = \theta \tag{6}$$

$$\Gamma(t, t+s) = \alpha \exp\left(\frac{-s^2}{2\rho^2}\right) \tag{7}$$

We made the following prior specifications: We set the bandwidth parameter for the squared exponential kernel as $\rho = 3$ representing a strong local correlation in time that died off sharply beyond 3 weeks, $\alpha = 1$ representing a signal to noise ratio of of approximately 1, $\upsilon = 1$ and $\xi = 1$ representing weak prior knowledge regarding the overall scale of variation in the the latent space. Finally, we set $\theta = -2.197$ representing an off-season prevalance of 0.1% non-influenza ILI. Samples from the posterior predictive density $p(\pi_{it}|y_{i1}, \ldots, \tilde{y}_{iT}, n_{i1}, \ldots, n_{iT})$ were collected using the function *basset* from the R package *stray* [15]; a total of 4000 such samples were collected in this analysis. We define the prevalence of non-influenza ILI in excess of normal seasonal variation as $y^*_{it} = \tilde{y}_{it}/n_{it} - \hat{\pi}_{it}$.

To exclude variation attributable to unseasonably high rates of other ILI causing viruses (such as the outbreak of RSV in Washington state in November-December 2019) we only investigate y^*_{it} for sets of 1 or more weeks $(t, t+1, \ldots, t+K)$ such that the following criteria are met:

1. the 2nd order moving average of $y^*_{i(t+k)} < y^*_{i(t+k+1)}$ must remain positive for all K
2. at least one week in $(t, t+1, \ldots, t+K)$ corresponds to a week after February 23, 2020

Together these criteria for attributing excess non-influenza ILI to COVID reflect our assumption that we expect COVID must be increasing within communities since Febuary 23, 2020.

4.5 Calculating scaling factors to relate ILINet data to COVID case numbers

As COVID new case counts $\tilde{z}_{it}$ represent the number of confirmed cases in an entire state and ILINet data represents the number of cases seen by a select number of enrolled providers, we must estimate scaling factors w_i to enable comparison of ILINet data to confirmed case counts at the who state level. Let π^*_{it} denote the probability that a patient with ILI in state i has COVID as estimated from ILINet data. Let p_i denote the population of state i and let b_i denote the number of primary care providers per 100,000 people in state i. We simulated the number of COVID cases (excess ILI meeting criteria above) as

$$\lambda_{it} \sim \text{Poisson}\left(\frac{n_{it}}{d_{it}}\right) \tag{8}$$

$$y^\dagger_{it} = \frac{b_i p_i}{10^5} \lambda_{it} \pi^*_{it} \tag{9}$$

That is we translate the inferred proportion of individuals with ILI due to COVID to the state level by considering the average number of patients seen by each provider in the study ($\frac{n_{it}}{d_{it}}$) and the number of primary care providers in state i ($\frac{b_i p_i}{10^5}$). Notably to account for potential errors in these scaling factors, we add propogate uncertainty into our calculation by using Monte-Carlo simulation of the average number of patients seen by each provider in the study.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

4.6 Estimating symptomatic case detection rates

Assuming that the majority of SARS-CoV-2 testing within the US has been directed by patient symptoms[16], the pool of newly diagnosed SARS-CoV-2+ patients is a subset of the pool of SARS-CoV-2+ patients who are identified as having ILI. Therefore, we calculate the probability that a symptomatic SARS-CoV-2+ patient will be identified as having SARS-CoV-2 as $\delta^s_{it} = \tilde{z}_{ij}/\pi^*_{it}$ (Figure S2).

4.7 Accounting for asymptomatic patients

To account for the contribution of assymptomatic SARS-CoV-2+ patients we use a recent analysis of cohort surveillance from the Diamond Princess [10]. Monte-Carlo simulations were used to propagate error from our uncertainty regarding potential asymptomatic infections affecting the clinical rate δ^c into our calculation of posteriors for epidemic trajectories. To match posterior estimates from Mizumoto et al.[10], we use quantile matching to parameterize $\delta^c \sim \text{Beta}(\alpha, \beta)$ to achieve a mean of .179 and a 95% probability set of $(.155, .202)$. We include this contribution to the overall detection rate δ_{it} as $\delta_{it} = \delta^s \delta^c_{it}$.

4.8 Epidemic growth rates and clinical rates

For a given epidemic growth rate, the clinical rate was estimated by comparing the excess ILI across the US $y^\dagger_t = \sum_i y^\dagger_{it}$ to the number of infections one might expect under simulated epidemic parameterized by the growth rate. An SEIR model,

$$\dot{S} = \zeta - \beta SI - \omega_b S \tag{10}$$

$$\dot{E} = \beta SI - \gamma E - \omega_b E \tag{11}$$

$$\dot{I} = \gamma E - \nu I - \omega_i I \tag{12}$$

$$\dot{R} = \nu I - \omega_b R \tag{13}$$

was parameterized for the US to a timescale of units days by setting $\zeta = 3.23 \times 10^{-5}$ corresponding to a crude birth rate of 11.8 per 1000 per year, a baseline mortality rate $\omega_b = 2.38 \times 10^{-7}$ corresponding to 8.685 per 1000 per year, an infectious mortality rate $\omega_i = 2.62 \times 10^{-7}$, incubation period γ^{-1} of 3 days, infectious period ν^{-1} of 10 days, and β parameterized to ensure $I(t)$ grew with a specified exponential growth rate early in the epidemic. A total of 14,000 simulations were run, each having a unique exponential growth rates, r, drawn uniformly from $r \in [log(2)/7, log(2)/1.5]$, reflecting doubling times in the range of $1.5 \leq t_2 \leq 7$ days.

Each simulation was initialized with $(S, E, I, R, t) = (3.27 \times 10^8, 0, 1, 0, 0)$ where time 0 was January 15. To simulate the stochastic time it took from the first case to the onset of regular exponential growth, a Gillespie algorithm was used from the initial conditions until either $t = 50$ (March 5, 2020) or $E(t) + I(t) = 100$. The output from Gillespie simulations was input as an initial value into the system of differential equations and integrated until the August 5, 2020. The number of infected individuals on a given day was the last observed $I(t)$ for that day, and a weekly pool of infected patients was computed by a moving sum over the number of infected individuals every day for the past week, $I_w(t) = \sum_{k=0}^{k=6} I_{t-k}$.

Defining $Y_t = \sum_i y^\dagger_{it}$ as the national excess ILI, the clinical rate was

$$\delta^s(t_d) = \frac{Y_t}{I_{w(t-t_d)}} \tag{14}$$

for a given time delay t_d it takes from the onset of infectiousness to a patient reporting to the doctor with ILI. An estimate of clinical rate as a function of exponential growth rate was obtained by a generalized additive model $\log(\delta^c) = s(\log(r))) + \epsilon$ where $s(.)$ denotes a piece-wise cubic spline.

Simulated epidemic trajectories were matched to posterior samples of Y_t through kernel density estimation. For t corresponding to the weeks starting on March 8 and 15, 2020, the density each simulated trajectory $I(t)$ was evaluated against the density

$$v_t \sim \text{log-normal}(0, 0.1) \tag{15}$$

$$p(I(t)|Y_t + v_t) \tag{16}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

where noise v_t was added to account for additional sources of error in our calculations of scaling factors, and p is a density estimated via kernel density estimation. The two weeks in March were chosen as the ILI surge had the largest correlation with covid case counts during these two weeks.

4.9 Estimating Doubling Times

Doubling times for confirmed cases and deaths due to COVID were estimated using Poisson regression with a log-link with an intercept and a covariate for the day of the epidemic. Poisson regression parameters r were converted to doubling times d using $d = log(2)/r$.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

References

[1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu *et al.*, "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, 2020.

[2] W. H. Organization *et al.*, "Coronavirus disease 2019 (covid-19): situation report, 59," 2020.

[3] J. Lourenco, R. Paton, M. Ghafari, M. Kraemer, C. Thompson, P. Simmonds, P. Klenerman, and S. Gupta, "Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the sars-cov-2 epidemic," *medRxiv*, 2020.

[4] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study," *The Lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[5] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china," *Jama*, 2020.

[6] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2)," *Science*, 2020. [Online]. Available: https://science.sciencemag.org/content/early/2020/03/24/science.abb3221

[7] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[8] N. Imai, A. Cori, I. Dorigatti, M. Baguelin, C. A. Donnelly, S. Riley, and N. M. Ferguson, "Report 3: transmissibility of 2019-ncov," 2020.

[9] M. L. Holshue, C. DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural *et al.*, "First case of 2019 novel coronavirus in the united states," *New England Journal of Medicine*, 2020.

[10] K. Mizumoto, K. Kagaya, A. Zarebski, and G. Chowell, "Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020," *Eurosurveillance*, vol. 25, no. 10, 2020.

[11] L. Pan, M. Mu, H. Ren *et al.*, "Clinical characteristics of covid-19 patients with digestive symptoms in hubei, china: a descriptive, cross-sectional, multicenter study," *Am J Gastroenterol*, 2020.

[12] "U.s. influenza surveillance system: Purpose and methods," 2020. [Online]. Available: https://www.cdc.gov/flu/weekly/overview.htm

[13] "Fluview interactive," 2020. [Online]. Available: https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

[14] "Primary care physician supply in all 50 states, ranked," 2020. [Online]. Available: https://www.beckershospitalreview.com/rankings-and-ratings/primary-care-physician-supply-in-all-50-states-ranked.html

[15] J. D. Silverman, K. Roche, Z. C. Holmes, L. A. David, and S. Mukherjee, "Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes," *arXiv e-prints*, p. arXiv:1903.11695, Mar 2019.

[16] "Coronavirus test: What you need to know," 2020. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-test-what-you-need-to-know

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

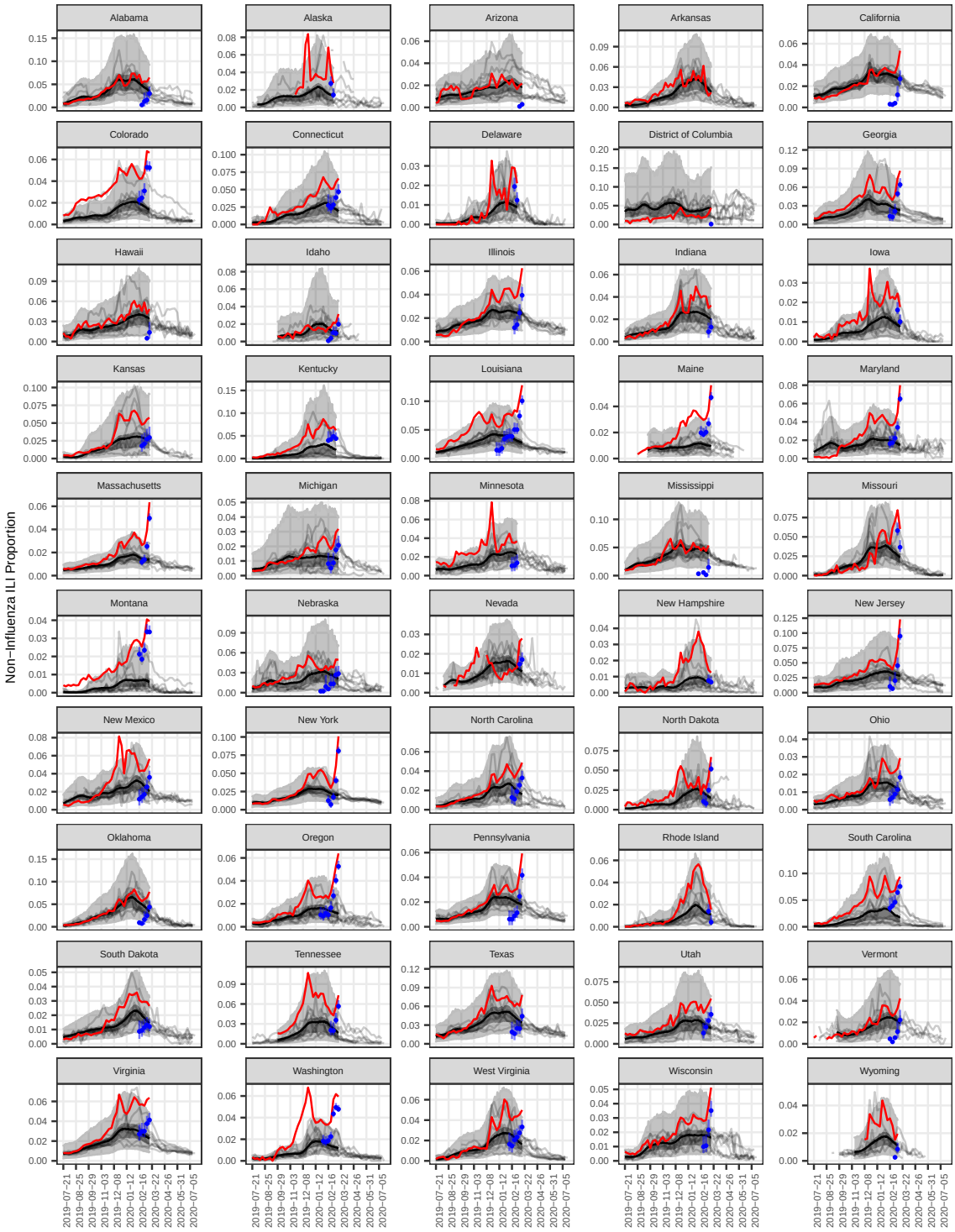

Figure 1: The excess non-influenza ILI is extracted from all non-influenza ILI by identifying the amount of non-influenza ILI in excess of seasonal norms (blue point and error bars represent the posterior median and 95% credible set for ILI not explained by non-COVID endemic respiratory pathogens). A binomial logistic-normal non-linear regression model was fit to non-influenza ILI data from 2010-2018 (grey lines). The model predicted the expected amount of non-influenza ILI in the 2019-2020 season (grey ribbons represent the 95% and 50% credible sets; the black line represents the posterior median). Observed non-influenza ILI outside of the 95% credible set was attributed to COVID based on the observed correlation across states with newly confirmed COVID case rates (Figure S1). A number of regions are not represented due to insufficient laboratory influenza data to complete our analysis (see *Methods* for full details).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

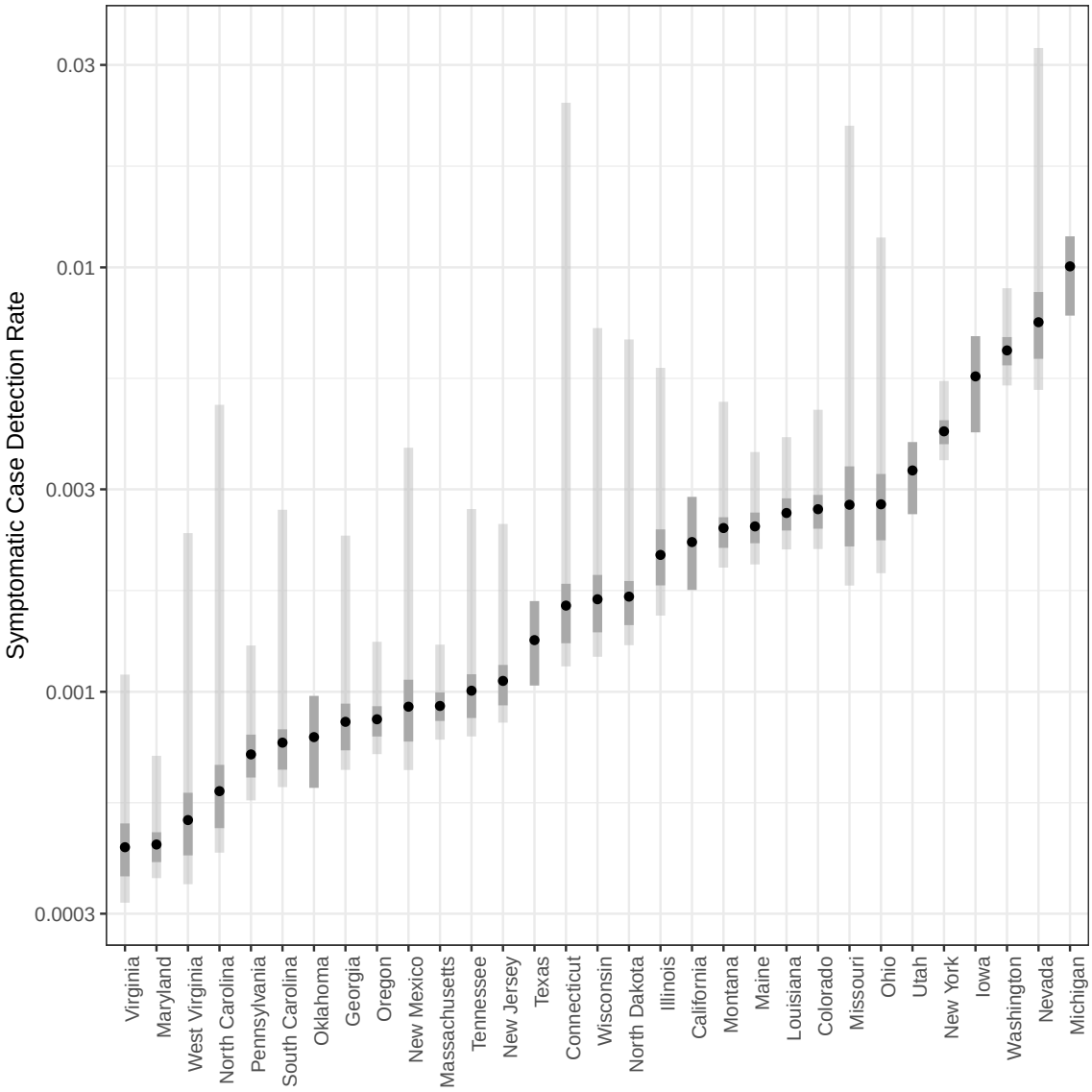

Figure 2: Assuming the non-influenza excess ILI for the week starting March 15 consists entirely of patients with COVID, the probability that a symptomatic COVID+ patient will be detected varies by state but even the highest symptomatic case detection rates are likely below 3%.In Figure S2 we show how the symptomatic case detection rate varies over time across states.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

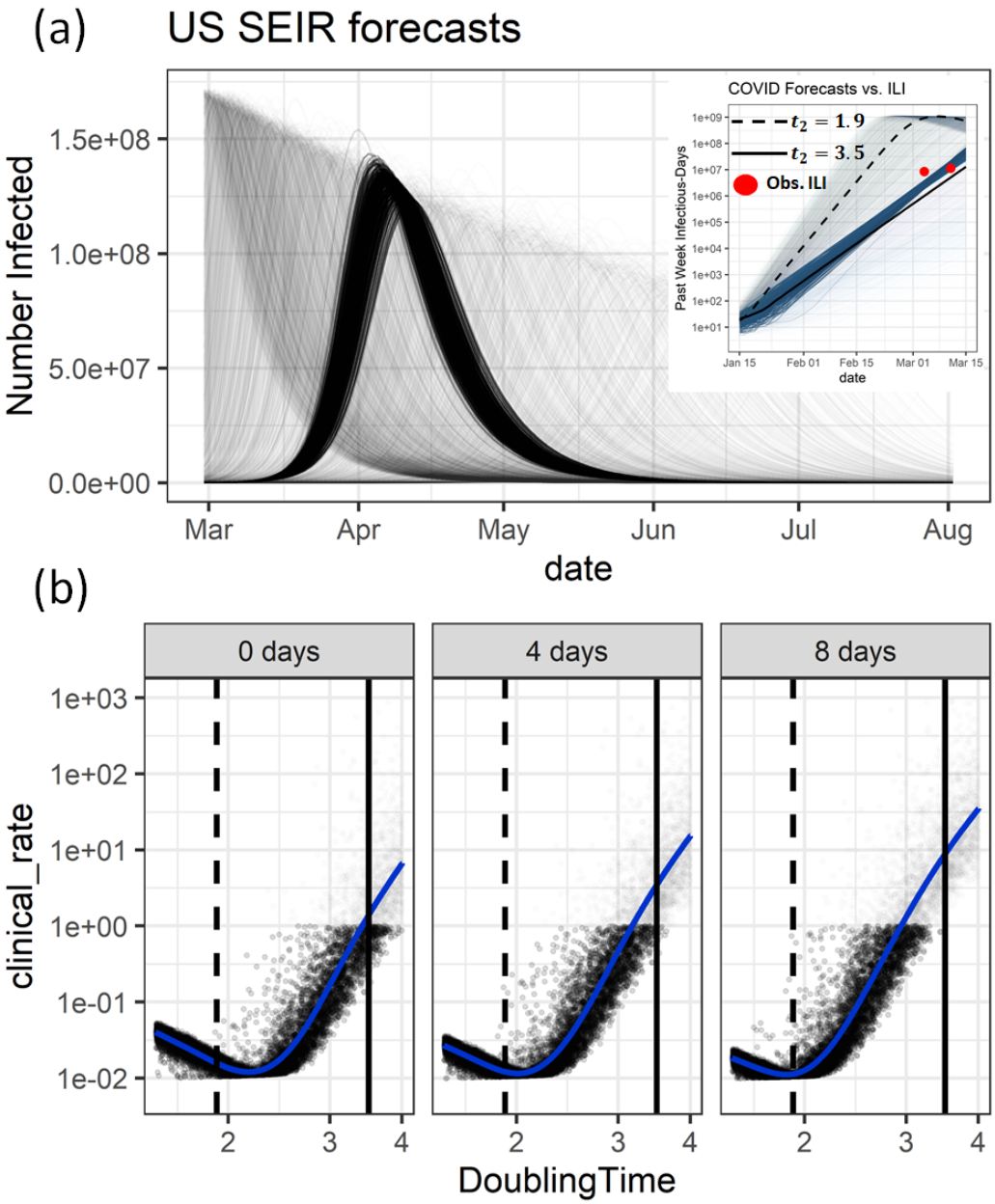

Figure 3: (a) 14,000 random SEIR models predict a range of current number of COVID cases in the US; less transparant trajectories are better agreements with the ILI surge under an assumption that 17% of infections are subclinical [10]. The inset plot shows COVID forecasts comparable to weekly ILI counts (red dots) by summing I(t) over the prior week. A $t_2 = 3.5$ day doubling time (solid black lines) is the slowest possible growth rate which can account for the observed data with a 4-day lag between the onset of infectiousness and reporting as ILI. Faster doubling times, such as $t_2 = 1.6$ (dashed black lines) can explain observed ILI data but require a larger rate of subclinical infections to do so. (b) Different growth rates imply different clinical rates in order to account for the excess ILI surge, and the relationship between growth rate and clinical rate varies with the unknown time it takes from the onset of infectiousness to presentation an ILI outpatient. The estimated 2.4-2.5 day doubling times across the US, for example, produce an estimated 3-4% clinical rate.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

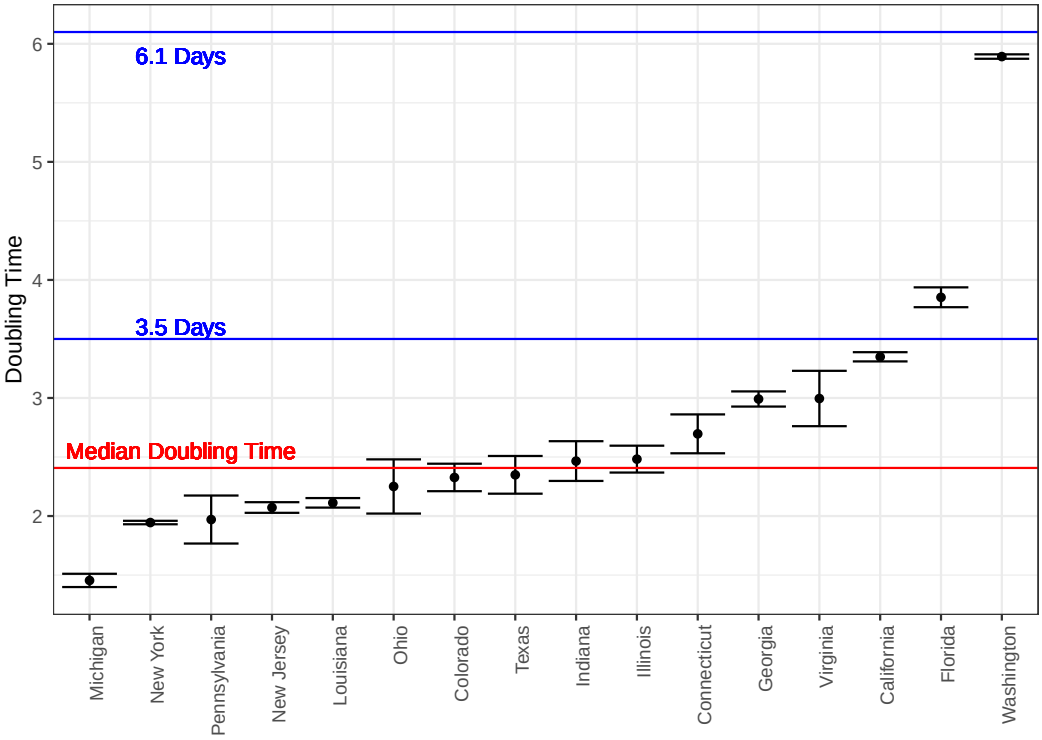

Figure 4: Doubling times for deaths due to COVID vary by states. Estimated doubling time and standard error are plotted. The median doubling for states is shown in red and falls below our bound of 3.5 days. Doubling times of confirmed cases is shown in Figure S3.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

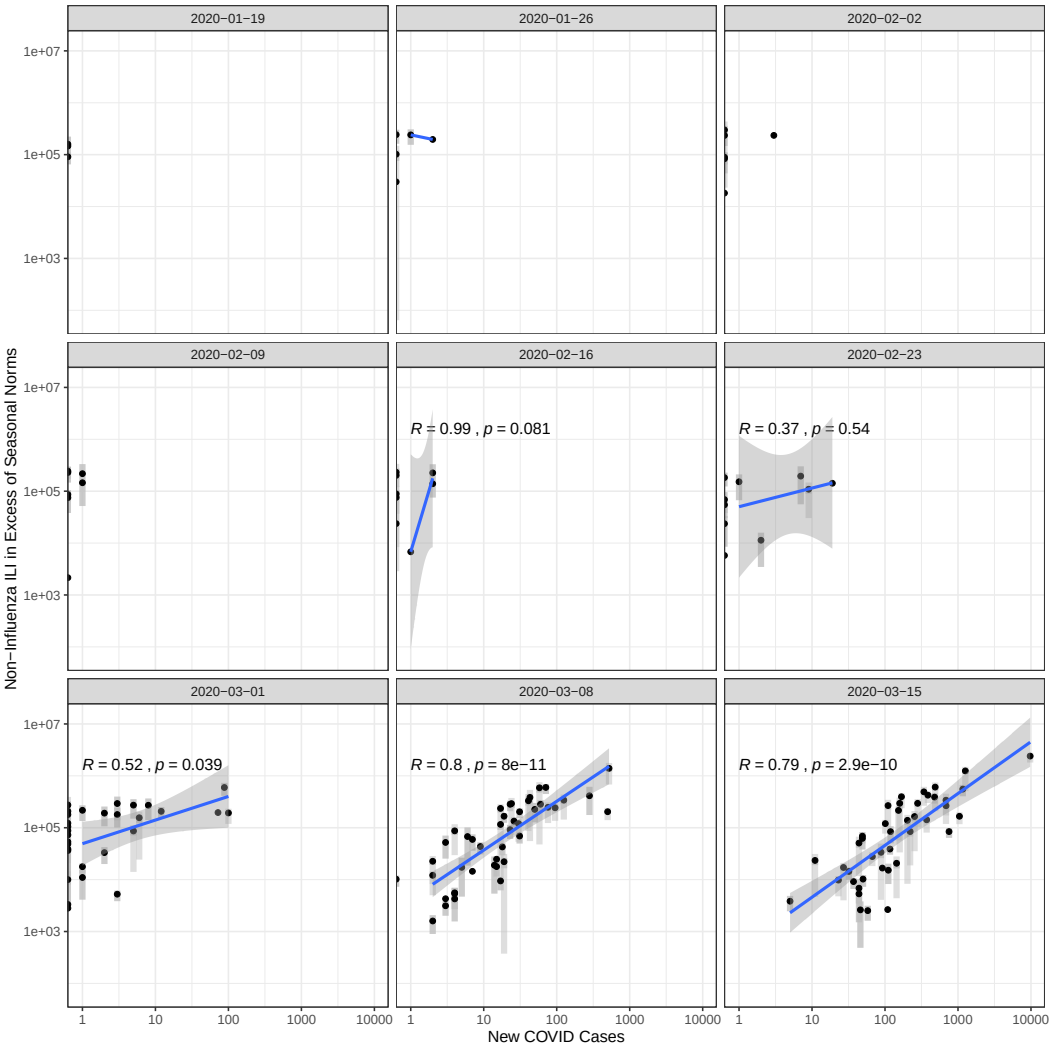

Figure S1: Excess ILI correlates strongly with patterns of newly confirmed COVID cases. This correlation is strongest for the last two weeks of data, when other seasonal respiratory pathogens are at their lowest.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

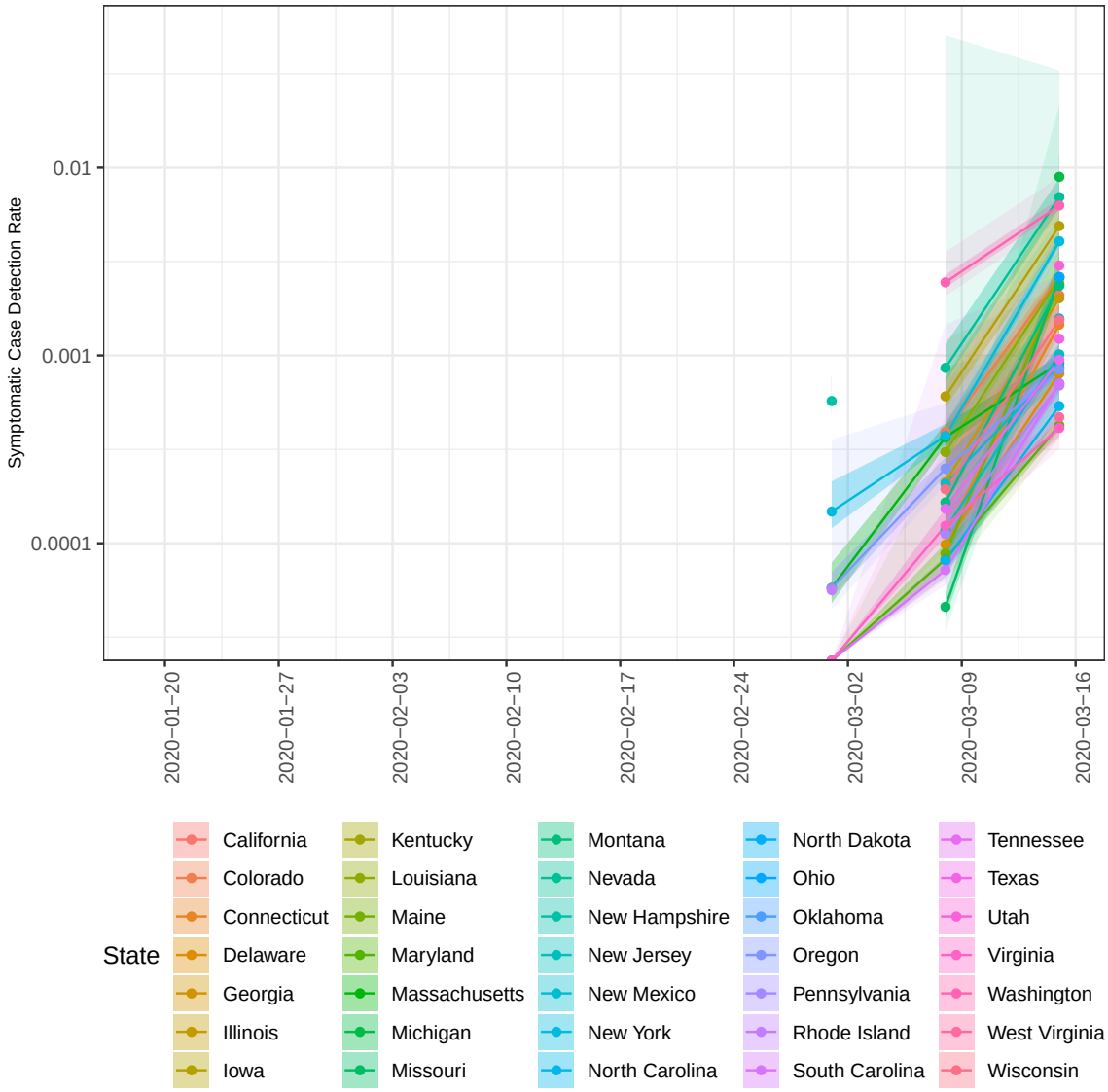

Figure S2: Since March 1, 2020, the case-detection of symptomatic COVID patients has increased by a factor of ≈ 100. This likely represents increased awareness of community transmission within the US combined with increased availability of testing. Still, the symptomatic case detection rate remains below 1% for most states with many states with detection rates closer to 0.1%.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

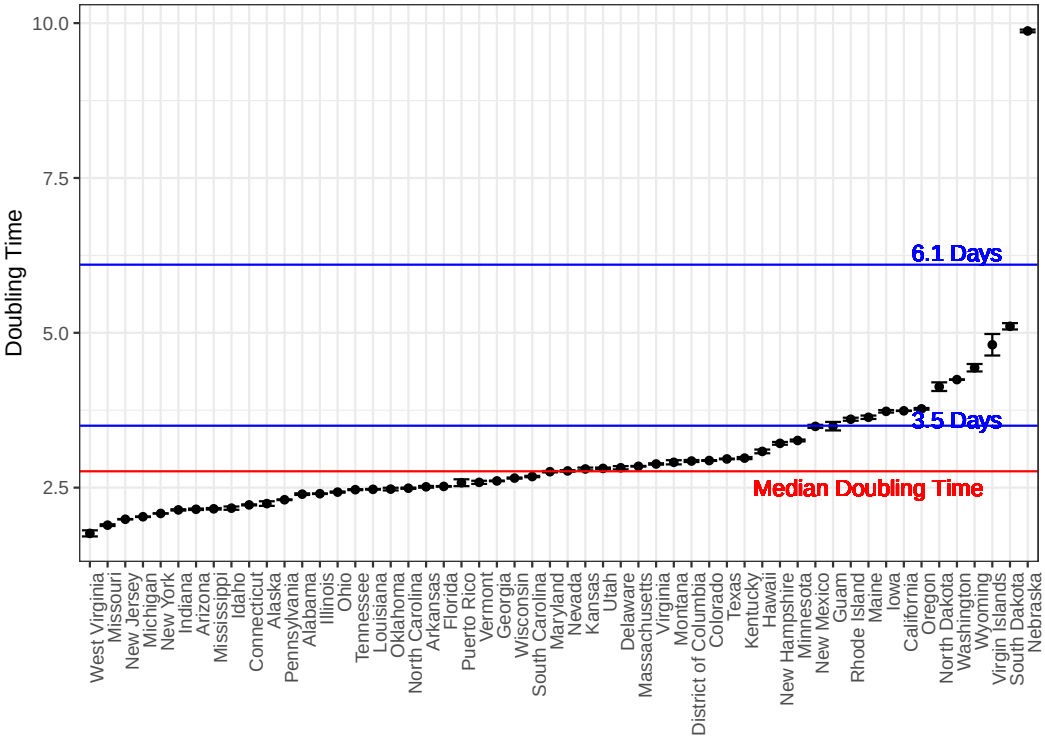

Figure S3: Doubling times of Confirmed Cases vary by states.... Estimated doubling time and standard error are plotted. The median doubling for states is shown in red and falls below our bound of 3.5 days.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.01.20050542; this version posted April 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

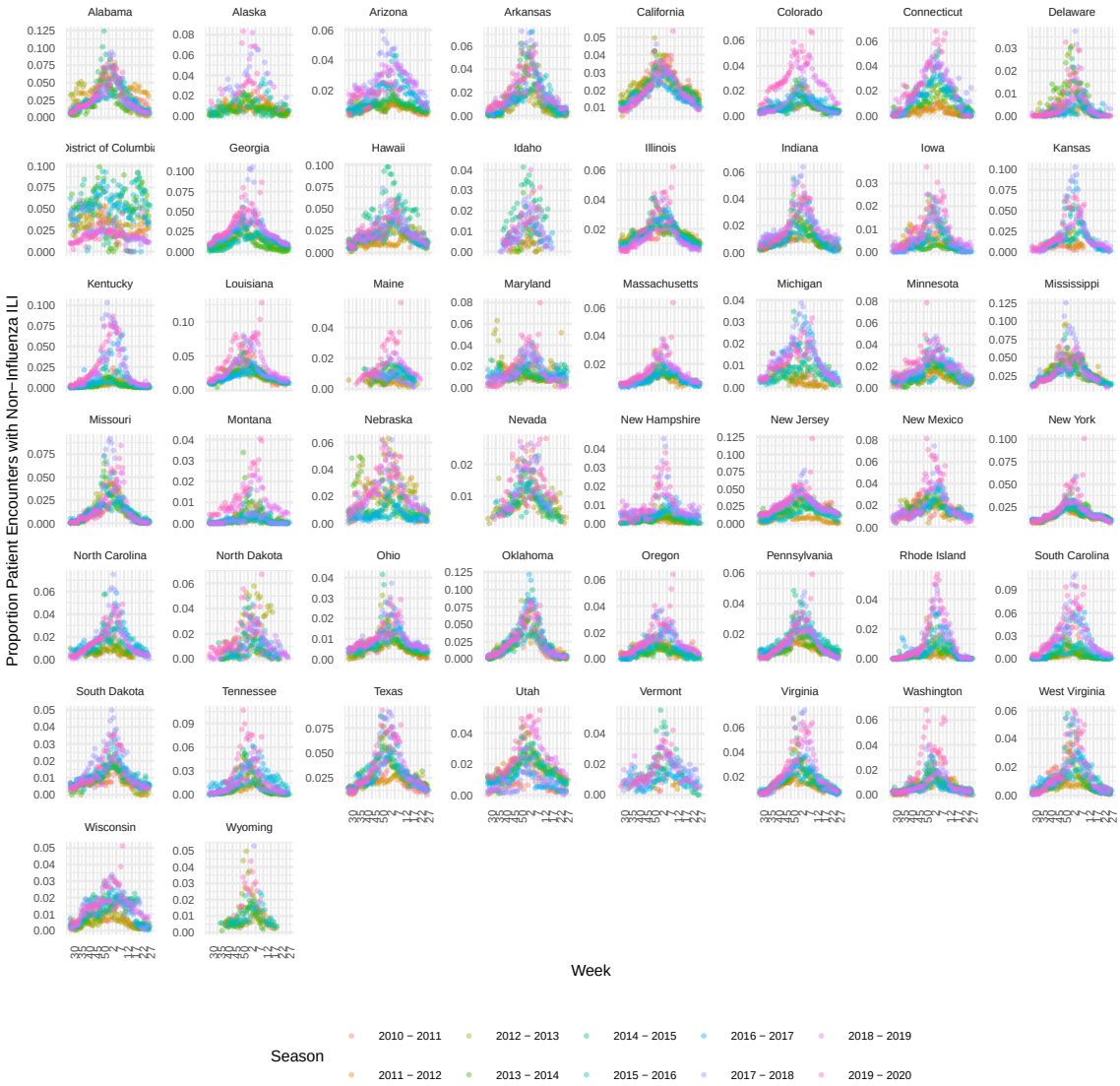

Figure S4: Once the signal attributable to influenza is extracted, the proportion of Patient encounters in which patient had non-influenza ILI ($\tilde{y}_{it}/n_{it}$) displays strong seasonal trends. The most notable deviations from these trends occur around Febuary to March of the 2019-2020 flu season and align with the onset of the COVID epididemic within the US.