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

Justin D. Silverman[1,2,3,7], Nathaniel Hupert[4,5], and Alex D. Washburne[6,7]

[1] *College of Information Science and Technology, Penn State University*
[2] *Department of Medicine, Penn State University*
[3] *Medical Scientist Training Program, Duke University*
[4] *Weill Cornell Medicine, Cornell University*
[5] *New York-Presbyterian Hospital*
[6] *Department of Microbiology and Immunology, Montana State University*
[7] *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 syndromic case detection rate of SARS-CoV-2 in the US to be approximately 1 our of 100. This corresponds to at least 28 million presumed symptomatic SARS-CoV-2 patients across the US during the three week period from March 8 to March 28. 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 3.5 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 use syndromic time series for early detection and understanding of emerging infectious diseases.

1 Introduction

The ongoing SARS-CoV-2 pandemic continues to cause substantial 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 epidemic 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.

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 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 the syndromic case detection rate is approximately 1%, corresponding to at least 28 million new ILI cases due to SARS-CoV-2. Our analysis provides empirical corroboration of previous hypotheses of substantial undocumented cases yet places the estimated undocumented case rate an order of magnitude higher than prior reports [6]. The SARS-CoV-2 prevalence estimates obtained from the ILI surge are consistent an epidemic doubling time of less than 3.5 days. A 3.5 day doubling time is substantially faster than many prior reports [7, 8] yet is consistent with the 3-day doubling time of observed deaths due to COVID within the US. Our findings support 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. Finally, we find that the ILI surge peaks the week starting March 15, and we discuss the potential explanations for this phenomenon.

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 $\rho > 0.35$ and $p < 0.01$ for the last three weeks; Figure S1A).

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 28.5 million excess individuals in the 3 week period between March 8 and March 28, 2020 compared to the same period in 2019 (95% credible interval of 25.9-30.5 million). Notably, we find that the ILI surge appears to peak during the week starting on March 15 and subsequently decreases in numerous states the following week; the notable exceptions being New York and New Jersey, two of the states that have been the hardest hit by the epidemic.

2.2 Investigating ILI Admission Rates

Our prevalence estimates could be falsely elevated if patient behavior has recently changed leading to increasing detection of mild ILI. If the ILI surge reflected higher rates of detection of typically mild ILI, we would expect emergency department ILI rates would increase yet the proportion of those ILI cases admitted to the hospital would decrease. We were able to obtain data to evaluate this hypothesis from New York City's influenza surveillance network [9]. In the month of March, the daily number of ILI visits to emergency departments across New York City increased while the proportion of those who went on to be admitted also increased by as much as 3-fold compared to the baseline rate prior to March (Figure S1). This suggests that patients are presenting less often for mild ILI, and such decrease in care-seeking behavior, if similar across the US, could be deflating the size of the ILI surge in later weeks of March.

2.3 Syndromic 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 syndromic 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 week ending March 14, COVID cases in the states with the highest estimated syndromic case detection rate (Washington, Nevada, and Michigan) are only capturing approximately 1% of ILI surges in those states (Figure 2). In the last week ending on March 28, we estimate the detection rate across the US increased to be 1.5% (95% credible interval 1.2%-1.9%).

2.4 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 define bounds and understand the 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 [10], 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 doubling time is longer than 3.5-days, as such slow growth scenarios fail to produce enough infected individuals to match the observed excess ILI.

Across the entire US, the doubling rate for deaths due to COVID is 3.01 days ($\pm$0.001, p-value of test that doubling rate is less than 3.5 days approximately 0). Under a 1-day lag from the onset of infectiousness to reporting as ILI, the doubling time of deaths in the US imply an expected 98.6% clinical rate (the proportion of patients who have symptoms for which they would present to a health care provider) if the entirety of the first week of ILI surge is comprised of COVID patients (Figure 3A). Adjusting the ILI surge to account for decreased care-seeking does not produce congruence between the epidemic curve and the ILI surge, suggesting additional factors can be affecting the ILI surge, such as successful interventions, even faster decreases in care-seeking than observed in New York, or epidemic growth rates faster than 3 days.

Faster growth rates, however, require lower clinical rates to explain the ILI surge. If the US epidemic prior to March 14 grew at the rate of deaths in Italy, doubling every 2.65 days, it could better match the curvature of the ILI surge and would imply a clinical rate of 16.5% (Figure 3B). For a four-day lag between the onset of infectiousness and presentation with ILI, the doubling time of US deaths produces, on average, too few COVID cases to explain the excess ILI on March 14. However, 29.8% of the stochastic simulations with a growth rate similar to that of US deaths produced enough COVID cases to explain the ILI surge and thus suggest either secondary introductions, super-spreading, or rapid transmission events in early transmission chains to exceed the ILI surge [11]. On the other hand, the doubling time of deaths in Italy could capture the US excess ILI with a 38.8% clinical rate. If researchers produce estimates of growth rates for the US epidemic, the ILI surge can be used to estimate bounds and ranges of possible clinical rates (Figure 3C). If the entirety of the ILI surge is attributable to COVID, it suggests a slowest-possible doubling time of 3.5 days for the US epidemic starting on January 15.

3 Discussion

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. The slowest epidemic doubling time that could explain the ILI surge would be 3.5 days and this rate could only be achieved unusually fast early transmission or super-spreading events and a clinical rate near 100%. We measure the doubling time of deaths due to COVID with in the US to be 3.01 days and note that this is consistent with the bound imposed by the ILI surge. 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 distribution of clinical symptoms than estimated from prior studies of SARS-CoV-2+ inpatients.

Excess ILI appears to have peaked during the week starting on March 15th, leading the observed ILI dynamics to diverge from the overall epidemic dynamics implied by the growth rate of COVID deaths in the US. If the ILI dynamics were proportional to the epidemic curve then the two could be related with a constant subclinical rate. However, the changing ratio between the ILI surge and the epidemic curves parameterized by the growth rate of US deaths suggests additional mechanisms may be behind the ILI slowdown. First, a slowdown in ILI outpatient arrivals could be due to decrease in care-seeking where patients with mild ILI are less likely to present to the hospital as evident in emergency departments across New York City. Adjusting our ILI prevalence estimates based on the effect observed in New York City aligns ILI estimates more closely with predicted dynamics, yet the discrepancy remains. The remaining deviation could reflect more extreme changes in patient behaviors than those seen in New York City or successful interventions leading to lower transmission rates.

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 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 [12], 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 syndromic 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 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 syndromic 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}) [13]. 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 [14].

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 [15]. 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. The dataset of deaths in Italy was downloaeded from https://github.com/pcm-dpc/COVID-19 on April 6, 2020.

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 S3.

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* [16]; 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 weeks after March 7th 2020 as only these later weeks had high correlation to the COVID confirmed case rate (Figure S1).

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.

4.6 Estimating syndromic case detection rates

Assuming that the majority of SARS-CoV-2 testing within the US has been directed by patient symptoms[17], 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 Growth Rate estimation

Growth rates were estimated for the US and Italy by poisson generalized linear models predicting new deaths with date. Data on COVID deaths in the US were obtained from https://github.com/nytimes/covid-19-data on April 6, 2020 and all deaths from March 5, 2020 to April 1, 2020, were summed by date. Initially, April 2-5 were included but were found to have extremely high leverage and were hence excluded from our analysis. Data on COVID deaths in Italy were obtained from https://github.com/pcm-dpc/COVID-19. The same procedure was applied, focusing on deaths from February 24 until March 12. The slope from poisson regression was used as the estimated exponential growth rate, yielding a US growth rate $r_{US} = 0.23$ or a 3.01 day doubling time and $r_{IT} = 0.26$ or a 2.65 day doubling time.

4.8 Epidemic simulations and clinical rates

SEIR models,

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

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

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

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

were 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 2,000 simulations were run for each of the two growth rates (US and Italy) analyzed. Growth rates were drawn at random with r_{US} $N(r_{US}, 0.1)$ and r_{IT} $N(r_{IT}, 0.1)$. To illustrate the mutual dependence between estimates of growth rate, clinical rate, and the lag between the onset of infectiousness to presentation to a doctor with ILI, 2,000 simulations with uniform growth rates in the interval [0.173,0.365] corresponding to a range of doubling times between 1.9 days and 4 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 initial Gillespie algorithm was implemented on the assumption that a large amount of variation in the epidemic trajectory stems from uncertainty in trajectory of early transmission chains. 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 $Y_t = \sum_i y_{it}^{\dagger}$ as the national excess ILI, the clinical rate implied by a given simulation was estimated as

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

for a given time delay t_d it takes from the onset of infectiousness to a patient reporting to the doctor with ILI.

4.9 Code Availability

All code and data required to reproduce our results is publicly available at https://github.com/jsilve24/ili_surge.

5 Acknowledgements

We thank Rachel Silverman, Raina Plowright, and Dan Rosenheck for their manuscript comments. JDS was supported in part 340 by the Duke University Medical Scientist Training Program (GM007171).

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] "New york city department of heath and mental hygeiene, influenza surveilance report week ending march 21, 2020 (week 12)," 2020. [Online]. Available: https://www1.nyc.gov/assets/doh/downloads/pdf/hcp/weekly-surveillance03212020.pdf

[10] 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.

[11] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, "Superspreading and the effect of individual variation on disease emergence," *Nature*, vol. 438, no. 7066, pp. 355–359, 2005.

[12] 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.

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

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

[15] "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

[16] 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.

[17] "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

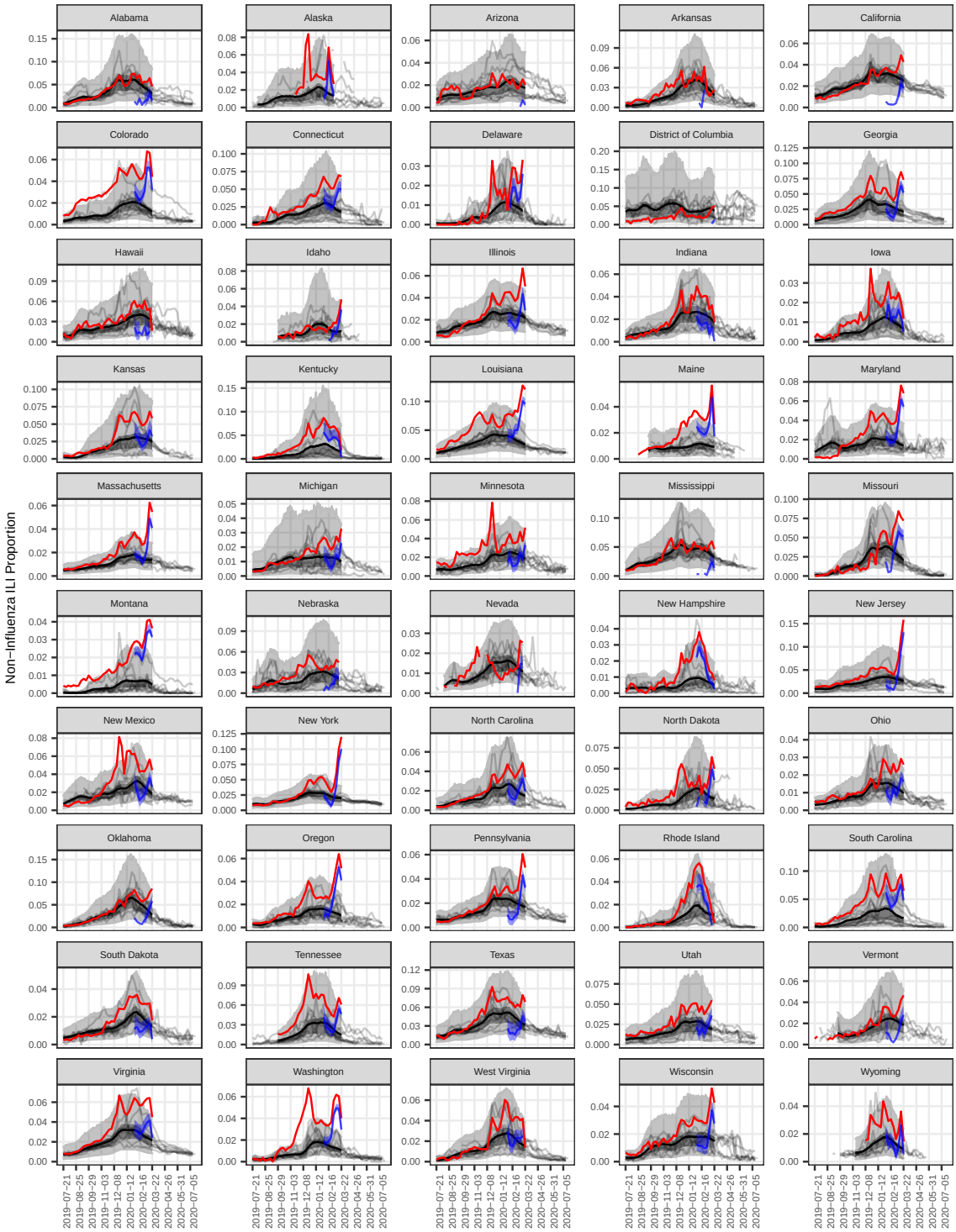

Figure 1: The excess non-influenza ILI is extracted from all non-influenza ILI by identifying the amount of non-influenza ILI in excess of seasonal norms (blue point and error bars represent the posterior median and 95% credible set for ILI not explained by non-COVID endemic respiratory pathogens). A 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 beyond seasonal norms are shown as a blue line (posterior median) and blue ribbon (posterior 95% credible set). A number of regions are not represented due to insufficient laboratory influenza data to complete our analysis (see *Methods* for full details).

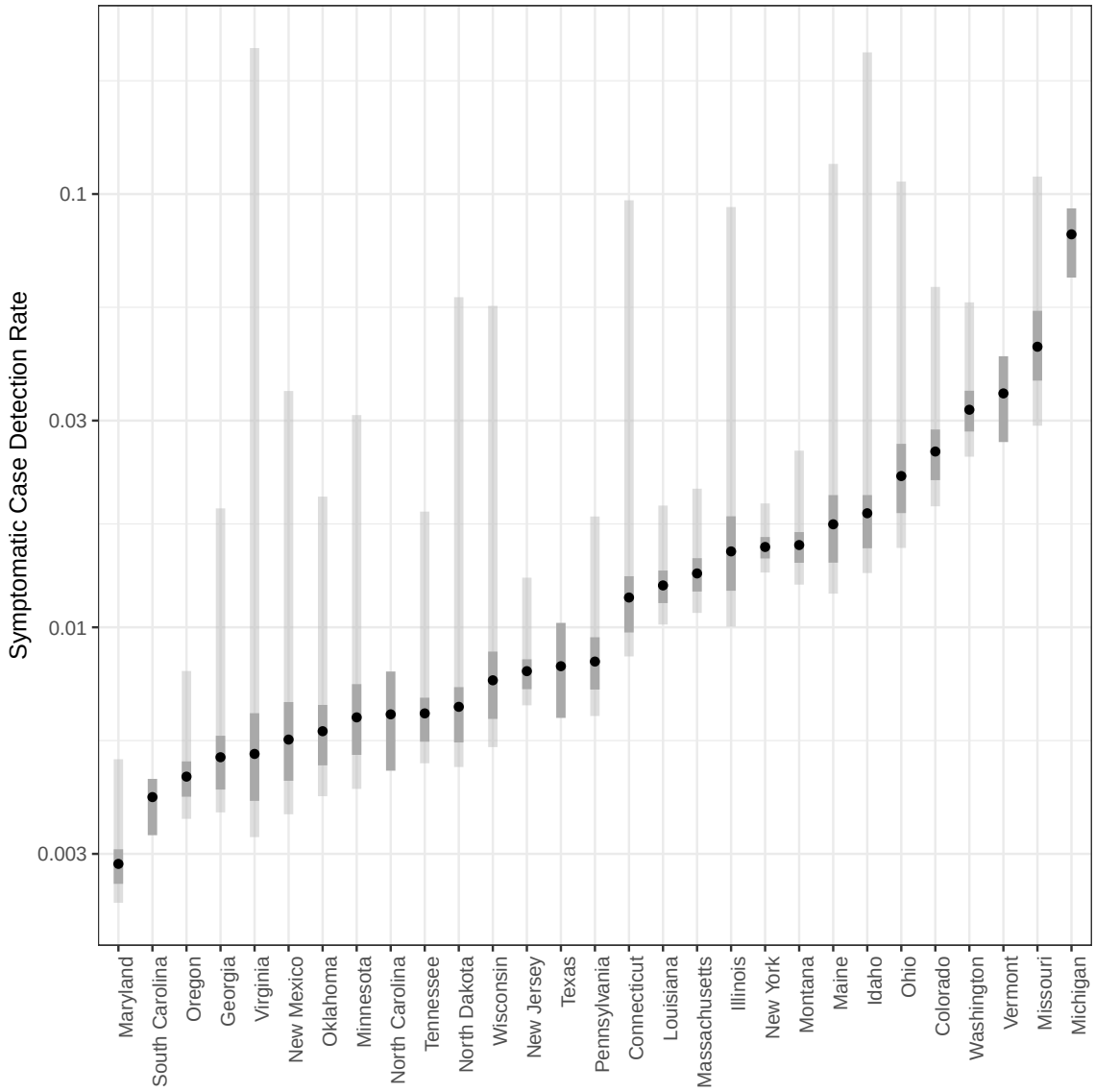

Figure 2: Assuming the non-influenza excess ILI for the week starting March 22 consists entirely of patients with COVID, the probability that a symptomatic COVID+ patient will be detected varies by state but even the highest syndromic case detection rates are likely below 10%. Across the US the syndromic case detection rate is 1.5% (95% credible set 1.2% to 1.9%). In Figure S2 we show how the syndromic case detection rate varies over time across states.

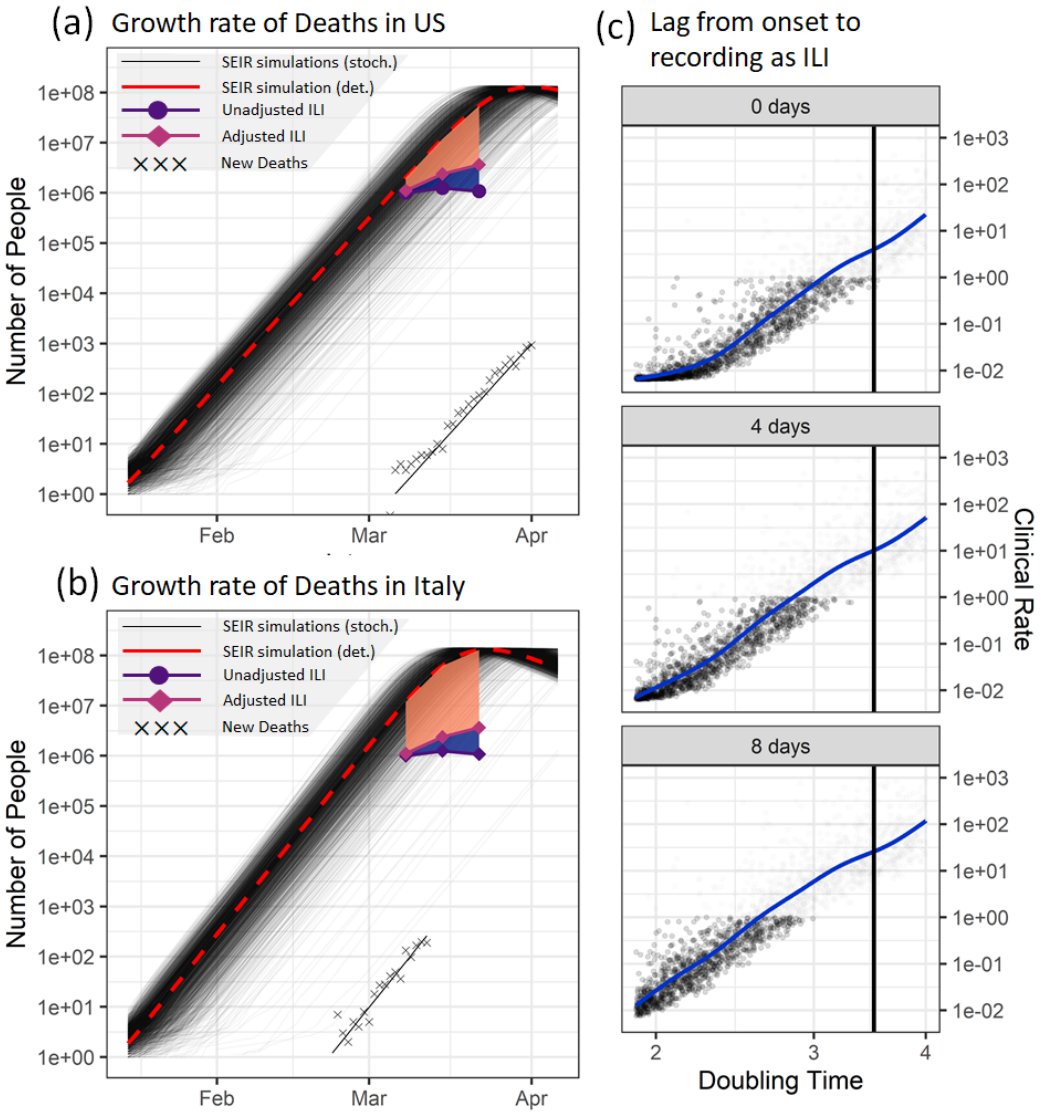

Figure 3: (a) The excess ILI estimated falls within the range of what one could expect from a US epidemic growing at the exponential growth rate defined by the growth of new deaths. Adjusting the ILI surge based on decreased care seeking in New York does not reconcile the difference between the ILI surge and the epidemic curve from US growth rates, suggesting additional forces are at play. The cause of the apparent deceleration in the ILI surge is hypothesized to be due to some combination of successful interventions, faster decreases in care-seeking behavior changing than measured in New York, and/or other possibilities including faster growth and higher subclinical rates. (b) If the growth rate in the US is faster than US deaths suggest, such as a growth rate observed in Italy prior to the Italian lockdown (2.645 day doubling time), it could provide alternative explanations of the curvature of the excess ILI through a larger subclinical rate and epidemic curves near their peak at the time of the peak of the ILI surge. Serology or other measures of prevalence are needed to reconcile these alternative hypotheses. (c) More generally, the ILI surge forces a dependence between growth rate (doubling time), the clinical rate, and the lag between the onset of infectiousness and presentation to the doctor with ILI, where faster growth implies a slower clinical rate.

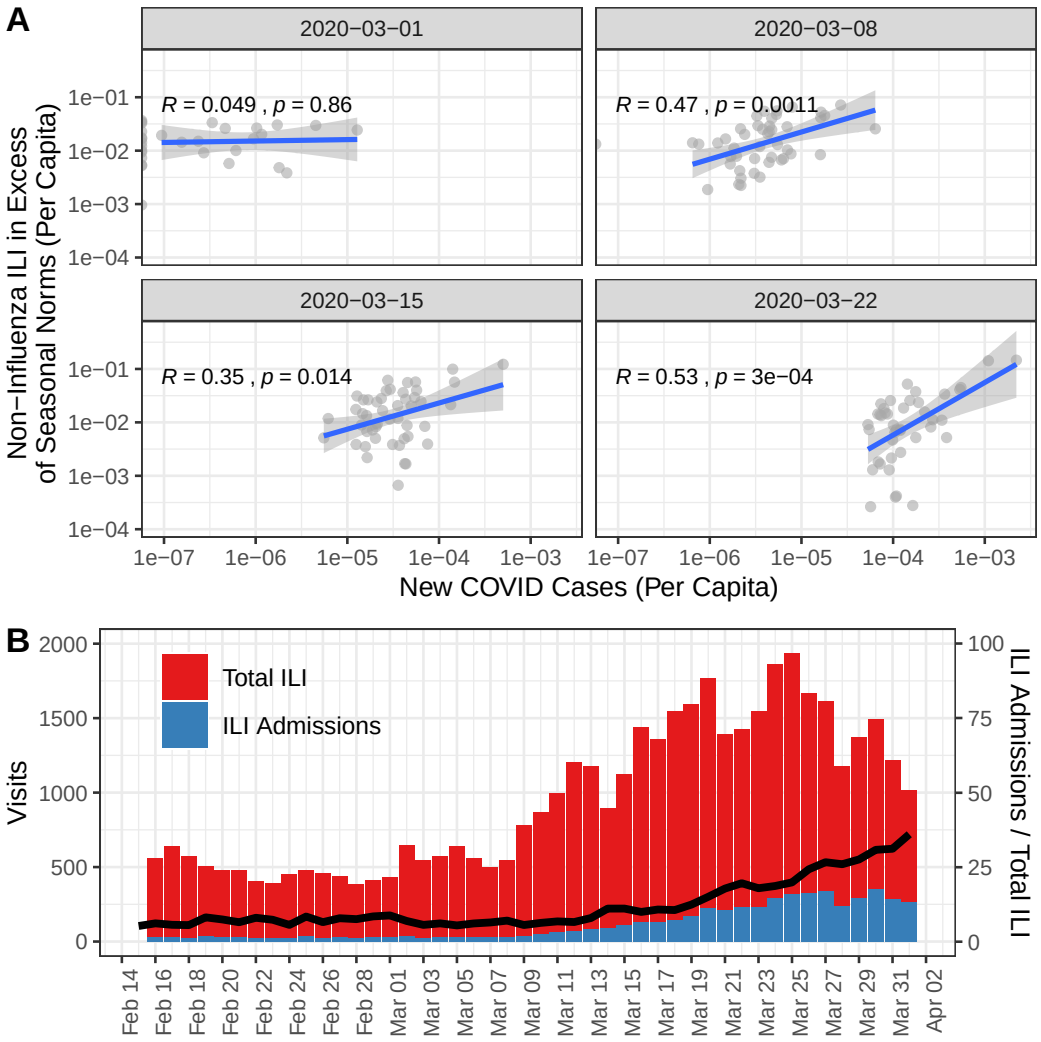

Figure S1: (A) Excess ILI correlates strongly with patterns of newly confirmed COVID cases. This correlation is strongest for the last three weeks of data, when other seasonal respiratory pathogens are at their lowest. (B) Of all ED visits for ILI in New York City, the proportion (black line) of those severe enough to warrant admission to the hospital has increased in the past month.

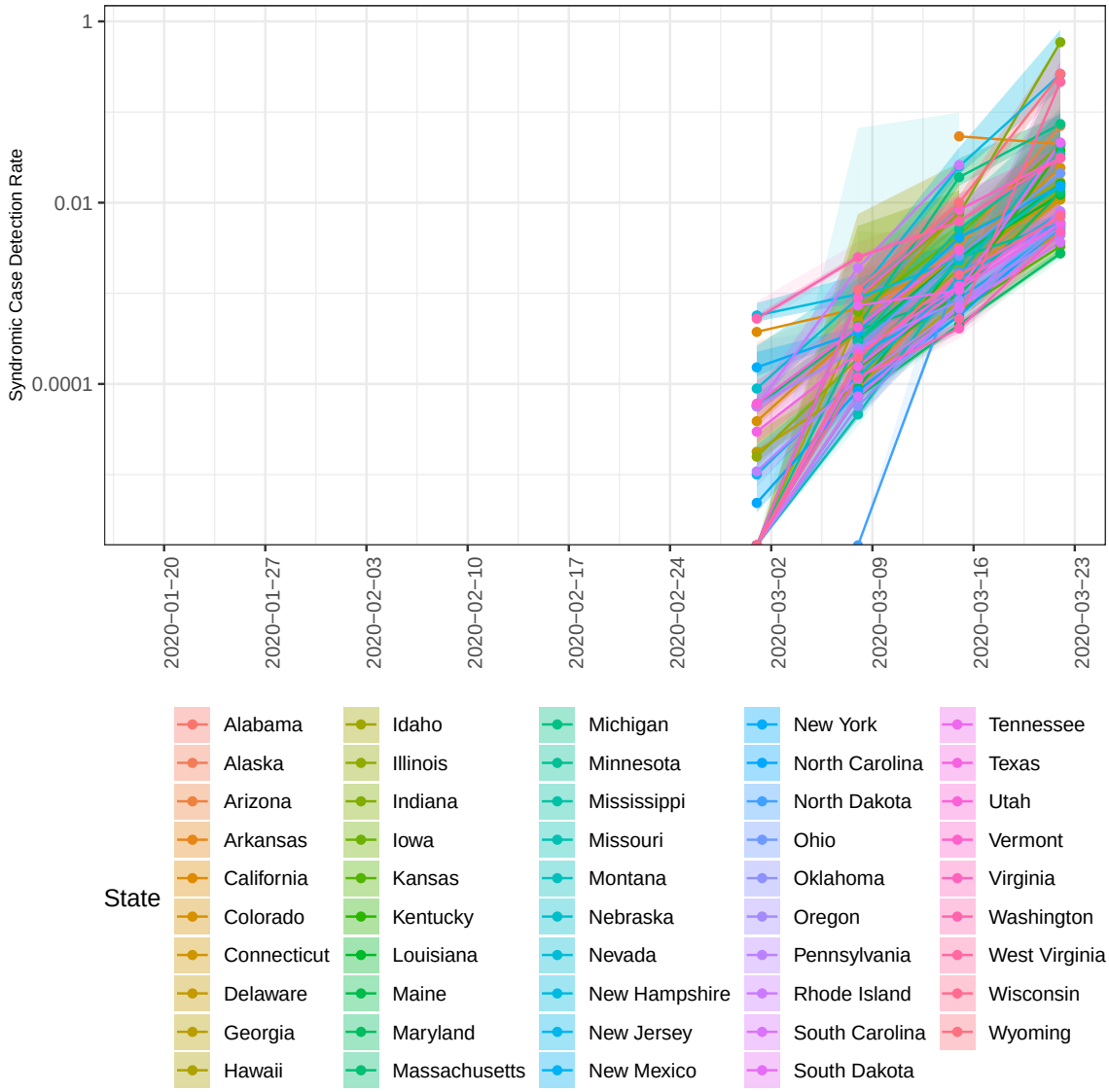

Figure S2: Since March 1, 2020, the case-detection of symptomatic COVID patients has increased by a factor of ≈ 100. This likely represents increased awareness of community transmission within the US combined with increased availability of testing. Still, the syndromic case detection rate remains below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

Figure S3: 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.