Abstract

Introduction

Throughout the past two decades of civilizations, from Severe acute respiratory syndrome (SARS) in 2003, to Middle East Respiratory Syndrome (MERS) first identified in 2012, to the novel coronavirus (COVID-19) outbreak in 2020, humans are threatened by respiratory infectious disease almost every 8 years. Except for an active search for vaccine, centered around all scientific efforts is the estimation of the disease spread.

Accompanied by this outbreak, studies related to the spread of COVID-19 mounted: Scientists and scholars around the globe are actively gathering COVID-19 data from all countries, predicting the peak time, and evaluating the effective of different public health policies. Among these publications, there are numerous estimations for serial interval [1],[2] using data from China, studies showing positive association between environmental factors and mortality rate of COVID-19 [3] , and post-pandemic seasonality forecast [4]. However, since COVID-19 outbreak happened in China and other European countries two months earlier than it happened in the US, most of the publications up to today focused on disease spread outside of America, and publications focused on the US are mostly descriptive and state-level. In this paper, we shift perspective to US territory, and more specifically, we aim to predict county-level disease spread in the US, taking into account three public health policies (Social-Distancing, Mask Policy and testing) and seven demographic characteristics (population density, proportion of older population [9], education level, poverty percentage, ethnicity, winter temperature, and maximum humidity [8]).

Data Source

County-level daily cases were extracted from Johns Hopkins University GitHub repository https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/. State-level public health policies and effective dates were extracted from https://www.kaggle.com/lin0li/. State-level testing data was extracted from https://covidtracking.com/data, which is the most credible source before CDC released the number. Various county-level demographic characteristics were extracted from https://www.census.gov/. County-level environmental data was extracted from https://www.ncei.noaa.gov/news/noaa-offers-climate-data-counties and for county-level temperature data, thanks to Wu Xiao at Harvard T.H Chan school of public health, who did the zonal statistics to aggregate winter temperature and averaged across grid cells in each county https://github.com/wxwx1993/PM_COVID/tree/master/Data

Method

In epidemiology, the basic reproductive number R_0 is often computed to quantify the intensity of an outbreak. R_0 could be intuitively thought of the expected number of infectees directly generated by one infector in a population where all individuals are susceptible to be infected. However, in the context of COVID-19, we don't assume that everyone is susceptible due to cross-immunity [4], thus a daily effective reproductive number is computed for every county in the US [5].

$$R(t) = \sum_{s=t}^{t+stop} \frac{b(s)g(s-t)}{\sum_{a=0}^{stop} b(s-a)g(a)} \tag{1}$$

where b(s) is the incidence number on day s, g(a) is the value of serial interval distribution at day a. The serial interval for this novel coronaviruses has not been firmly established. In this analysis, we estimated serial interval distribution to be from a Weibull distribution with mean 7.5 days and standard deviation 3.4 days [10] and varied this assumption in supplementary analysis based on other estimations for the serial interval. We assumed that the maximum generation interval was the first day when 99% of the density had been obtained. In addition, the first few dates were truncated for each county to avoid biased estimates of R_e because confirmed cases were not well documented in the beginning, and last few dates were also truncated to eliminate the misunderstanding that disease has been eradicated in this county. Number of dates to truncate depends on the serial interval estimation adopted.

The daily effective reproductive number $R(t)_i$ on day t of i^{th} county can be expressed as a county-specific constant modified by the proportion of susceptible individuals at the beginning of day t and further modified by external fixed and time-varying covariates [6]

$$R(t)_i = constant_i \times S_i(t)^{z_i} \times \prod_{k=1}^{7} x_{ik}^{\beta_k} \times \prod_{j=1}^{3} c_{ij}(d)^{\alpha_j(d)} \tag{2}$$

where x_{ik} is the k^{th} fixed variables for i^{th} county, and $c_{ij}(d)$ is the j^{th} time-varying variables for i^{th} county at day d. $S_i(t)$ stands for the proportion of susceptible individuals up to today adjusting for cumulative infections, county-level fixed variables are population per squared-mile, proportion of population over 60 years old, proportion of population with more than high school degree, poverty percentage, proportion of African American, winter maximum temperature in Fahrenheit and winter maximum humidity. Time-varying variables are the state-level social-distancing order, which is assumed to be implemented county-wide, nation-wide mask-wearing order and state-level number of tests per 1000

people over time.

To account for the fact that the influence of COVID-19 testing on containing disease spread is not conspicuous immediately, time-varying effects $\alpha_3(d)$ for the number of tests per 1000 people over time is incorporated to account for the exposure-lag-response associations (i.e delayed response), where d is a collection of dates before the date of interest, variables social-distancing order and mask-wearing order are calculated as days after policy enacted otherwise zero if not yet enacted. Taking the log on both side, we have a regression.

$$\begin{aligned} log(R(t)_i) =& log(R_{0i}) + z_i log S_i(t) + \sum_{k=1}^{7} \beta_k log(x_{ik}) + \sum_{j=1}^{2} \alpha_j(t) log(c_{ij}(t)) + \alpha_3(d) log(c_{i3}(d)) \\ =& log(R_{0i}) + z_i log S_i(t) + \sum_{k=1}^{7} \beta_k log(x_{ik}) + \sum_{j=1}^{2} \alpha_j(t) log(c_{ij}(t)) + \sum_{d=t-b_3}^{t} \alpha_3(d) log(c_{i3}(d)) \end{aligned} \tag{3}$$

where b_3 is the number of days we want to look back from the day of interest for number of testings per 1000 people. To estimate each $\alpha_3(t-b_3)$,$\alpha_3(t-b_3+1)$... ...$\alpha_3(t-1), \alpha_3(t)$, we proposed restricted cubic spline approach.

(1) Firstly, a linear mixed effect model with proportion of susceptible individual at time t and 7 fixed variables (second and third term in equation (3), respectively) and random intercept representing different counties were constructed to be benchmark against with the later results when social-distancing order, mask-wearing policy and COVID-19 testings are incorporated.

(2) Secondly, adding the fourth and fifth terms in equation (3), three linear models with b_3= 5, 10 and 20 and the effects of social-distancing and mask-wearing being constant were fitted to determine the most reasonable lag days for the effect of COVID-19 testing.

(3) Thirdly, counties with observations fewer than lag days were removed. Social-distancing

and mask-wearing variables were also fitted using natural cubic spline with 4 knots to account for time-varying effects, and tests per 1000 people were again fitted with natural cubic spline, B Spline and cubic spline with 10 knots and pre-determined lag days from step two. Three linear mixed effect models with random intercepts were fitted and compared in term of mean squared error.

Results

- findings

Discussion

- whether policies are effective

Reference

[1] Du, Zhanwei, et al. "The Serial Interval of COVID-19 from Publicly Reported Confirmed Cases." 2020, doi:10.1101/2020.02.19.20025452.

[2] Nishiura, Hiroshi, et al. "Serial Interval of Novel Coronavirus (2019-NCoV) Infections." 2020, doi:10.1101/2020.02.03.20019497.

[3] Wu, Xiao, et al. "Exposure to Air Pollution and COVID-19 Mortality in the United States." July 2020, doi:10.1101/2020.04.05.20054502.

[4] Kissler, Stephen M, et al. "Projecting the Transmission Dynamics of SARS-CoV-2 through the Post-Pandemic Period." June 2020, doi:10.1101/2020.03.04.20031112.

[5] Wallinga, J, and M Lipsitch. "How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers." Proceedings of the Royal Society B: Biological Sciences, vol. 274, no. 1609, 2006, pp. 599–604., doi:10.1098/rspb.2006.3754.

[6] Beest, Dennis E. Te, et al. "Driving Factors of Influenza Transmission in the Netherlands." American Journal of Epidemiology, vol. 178, no. 9, Dec. 2013, pp. 1469–1477., doi:10.1093/aje/kwt132.

[7] Wang, Molin, et al. "Quantifying Risk over the Life Course - Latency, Age-Related Susceptibility, and Other Time-Varying Exposure Metrics." Statistics in Medicine, vol. 35, no. 13, Oct. 2016, pp. 2283–2295., doi:10.1002/sim.6864.

[8] Mecenas, P., Bastos, R., Vallinoto, A., and Normando, D. (2020). Effects of temperature and humidity on the spread of COVID-19: A systematic review. doi: 10.1101/2020.04.14.20064923

[9] Johnson, K. M. (2020). An Older Population Increases Estimated COVID-19 Death Rates in Rural America. doi: 10.34051/p/2020.384

[10] Li, Qun, et al. "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia: NEJM." New England Journal of Medicine, 26 Mar. 2020, www.nejm.org/doi/full/10.1056/NEJMoa2001316.

Supplementary

Model for the i^{th} county at time point t:

$$R(t)_i = R_{0i} \times S_i(t)^{z_i} \times \prod_{k=1}^{7} x_{ik}^{\beta_k} \times \prod_{j=1}^{3} c_{ij}(d)^{\alpha_j(d)} \tag{4}$$

Annotations:

- $R(t)_i$: effective reproductive number at time t for the i^{th} county
- R_{0i}: basic reproductive number at the beginning for the i^{th} county

- $S_i(t)$: proportion of susceptible at time t in the county i: $\frac{S_i(0)-h(t)_i}{S(0)_i}$
- $S_i(0)$: susceptible at the beginning in the county i (assuming all population)
- $h(t)_i$: number of cases up to time t-1
- $d_{i1}...d_{i6}$: 6 fixed variable/effects for the i^{th} county: % of essential workers, humidity, poverty, population density, education, and % over 60 years old
- $c_{i1}...c_{i3}$: 3 time-varying variables/effects for i^{th} county: # of tests per 1000 people, social distancing policy, and mask policy

Taking the log on both side

$$log(R(t)_i) = log(R_{0i}) + z_i log S_i(t) + \sum_{k=1}^{7} \beta_k log(d_{ik}) + \sum_{j=1}^{3} \{ \sum_{d=t-b_j}^{t} \alpha_j(d) log(c_{ij}(d)) \} \quad (5)$$

Interpretation:

- $log(R_{0i})$: county specific intercept
- b_j: # of days we will look back for variable c_j, this will define the time window when the variable could affect $R(t)_i$
- z_i: how quickly susceptibles deplete
- $c_{ij}(t)$: the exposure value for covariate j for county i at time t

Step 1: Estimate R_0 for each county in the US

"J. Wallinga, M. Lipsitch, How generation intervals shape the relationship between growth rates and reproductive numbers. Proc. R. Soc. B Biol. Sci. 274, 599–604 (2007)."

$$R(t) = R(today) = \sum_{s=today}^{today+stop} \frac{b(s)g(s-today)}{\sum_{a=0}^{stop} b(s-a)g(a)} \quad (6)$$

Annotation:

- "The reproductive number to the time t at could be find by summing over all possible times t"
- $\sum_{s=t}^{t+stop}$: look ahead starting today through t+stop day
- $R(t)$: the effective reproductive number today (t)
- $b(t)$: incidence number at time t
- $g(t)$: value for general interval distribution at time t (the generation interval distribution is the probability distribution function for the time from infection of an individual to the infection of a secondary case by that individual)

- stop: max generation interval as first day that captures > 99% of the density

Step 2: Estimate time-varying effect by ML using restricted cubic spline
"Wang, Molin, et al. "Quantifying Risk over the Life Course - Latency, Age-Related Susceptibility, and Other Time-Varying Exposure Metrics." Statistics in Medicine, vol. 35, no. 13, Oct. 2016, pp. 2283–2295., doi:10.1002/sim.6864."
We will estimate each $\alpha_j(t-b_j), \alpha_j(t-b_j+1)...\alpha_j(t)$, j = 1, 2, and 3 with a restricted cubic spline approach with Q knots

$$\sum_{d=t-b_j}^{t} \alpha_j(d) log(c_{ij}(d)) = \sum_{d=t-b_j}^{t} \{[\phi_1 B_{-1}(d) + ... + \phi_Q B_{Q-2}(d)] log(c_{ij}(d))\}$$

For example:

$$\alpha_\phi(t-b_j) = \phi_1 B_{-1}(t-b_j) + \phi_2 B_0(t-b_j) + ... + \phi_Q B_{Q-1}(t-b_j)$$

$$\alpha_\pi(t-b_j) = \pi_1 B_{-1}(t-b_j) + \pi_2 B_0(t-b_j) + ... + \pi_Q B_{Q-2}(t-b_j)$$

$$\alpha_\psi(t-b_j) = \psi_1 B_{-1}(t-b_j) + \psi_2 B_0(t-b_j) + ... + \psi_Q B_{Q-2}(t-b_j)$$