medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license .

Statistics and Its Interface Volume 0 (2020) 1–10

The timing and effectiveness of implementing mild interventions of COVID-19 in large industrial regions via a synthetic control method

Ting Tian* , Wenxiang Luo* , Jianbin Tan* , Yukang Jiang* , Minqiong Chen, Wenliang Pan, Songpan Yang, Jiashu Zhao, Xueqin Wang, and Heping Zhang†

The outbreak of novel coronavirus disease (COVID-19) has spread around the world since it was detected in December 2019. The Chinese government executed a series of interventions to curb the pandemic. The "battle" against COVID-19 in Shenzhen, China is valuable because populated industrial cities are the epic centres of COVID-19 in many regions. We made use of synthetic control methods to create a reference population matching specific characteristics of Shenzhen. With both the synthetic and observed data, we introduced an epidemic compartmental model to compare the spread of COVID-19 between Shenzhen and its counterpart regions in the United States that didn't implement interventions for policy evaluation. Once the effects of policy interventions adopted in Shenzhen were estimated, the delay effects of those interventions were referred to provide the further control degree of interventions. Thus, the hypothetical epidemic situations in Shenzhen were inferred by using time-varying reproduction numbers in the proposed SIHR (Susceptible, Infectious, Hospitalized, Removed) model and considering if the interventions were delayed by 0 day to 5 days. The expected cumulative confirmed cases would be 1546, which is 5.75 times of the observed cumulative confirmed cases of 269 in Shenzhen on February 3, 2020, based on the data from the counterpart counties (mainly from Broward, New York, Santa Clara, Pinellas, and Westchester) in the United States. If the interventions were delayed by 5 days from the day when the interventions started, the expected cumulative confirmed cases of COVID-19 in Shenzhen on February 3, 2020 would be 676 with 95% credible interval (303,1959). Early implementation of mild interventions can subdue the epidemic of COVID-19. The later the interventions were implemented, the more severe the epidemic was in the hard-hit areas. Mild interventions are less damaging to the society but can be effective when implemented early.

*Dr. T. Tian, W. Luo, J. Tan and Y. Jiang contributed equally to this article.
†Corresponding author.

AMS 2000 subject classifications: Primary 00K00, 00K01; secondary 00K02.
Keywords and phrases: Mild interventions, Causal effects, Data assimilation, Synthetic control methods, Delay effects, Time-varying reproduction numbers.

1. INTRODUCTION

The confirmed cases of novel coronavirus (SARS-Cov-2) disease (COVID-19) had been detected since December 2019[30], and there are over 200 countries with reported confirmed cases in the world[35]. A potent characteristic of SARS-Cov-2 is that it can be contagious during the incubation period, where asymptomatic individuals could become a disseminating source[28]. Therefore, finding and implementing effective interventions is vital to disease control[17, 32].

On January 23, 2020, non-pharmaceutical interventions including all public transportation closures started in Wuhan, and a similar response was triggered in cities geographically near Wuhan cities, such as the cities within Hubei province and Wenzhou in Zhejiang province on February 1, 2020[20]. Guangdong province, where there was a large population of migrants from Hubei province[8], activated the first-level public health emergency response on January 23, 2020[34]. Shenzhen, with confirmed cases of COVID-19 initially reported in Guangdong province, implemented the interventions by closely following and isolating the close contacts of confirmed cases of COVID-19. By far, the strategic polices for direct protection (such as wearing face masks) and for transmission reduction (for example 14 days isolation for oversea travellers, cancellation of public gatherings, delayed reopening of schools) have been maintained in Shenzhen. The intervention and control strategies of Shenzhen were adopted as a classical example in the report from World Health Organization (WHO)[5].

As an industrial city, where there was a large population of inbound and outbound travellers[43], Shenzhen implemented a relatively mild intervention strategy compared with that of Wuhan and mounted an early response to the epidemic of COVID-19 relative to other cities such as Wenzhou. Therefore, there is a great interest to examine how

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

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

effective such mild but early interventions were in curbing the epidemic of COVID-19 in Shenzhen. We evaluated the treatment effects of intervention policies in Shenzhen by comparing the epidemic data in Shenzhen with its counterpart regions in the United States counties, and simulate the potential outcomes of different hypothetical starting dates of interventions to project the influence of the delay effects on the policy effectiveness.

To compare the effects of the control of epidemic in Shenzhen, the synthetic control method[2] as a classical methodology of causal inference was modified. It estimated the effects of short-term interventions between the affected unit and a combination of unaffected units[1]. To seek a series of unaffected units, i.e. the regions where interventions similar to those adopted in Shenzhen were not issued at the early stage of epidemic, we employed the infected counties in the United States as the quantitative reference for comparison.

Alternatively, in order to evaluate the delay effects of policy interventions appropriately, it's crucial to consider the pre-symptomatic transmission of COVID-19[16] to reflect the nature of COVID-19. Traditional epidemic models, such as SIR (Susceptible, Infectious, and Recovered) mode[14, 11] and SEIR (Susceptible, Exposed, Infectious, and Removed) model[29, 26], may not be suitable for modelling the spread of COVID-19[42] due to the lack of consideration for the incubation period and the pre-symptomatic transmission pattern during the incubation period. The literature suggests that the mean length of the incubation period ranges from 5 to 7 days, and considering the incubation period is important to the understanding of disease risk[28, 30, 7, 31].

In this paper, we proposed the SIHR (Susceptible, Infectious, Hospitalized, Removed) model, which considers pre-symptomatic transmission of COVID-19 by an appropriate compartmental assumption with the well-defined time-varying reproduction number. By modifying the estimated time-varying reproduction number[24, 36] to reflect delay effects of interventions, we simulated different outcomes of COVID-19 for comparison.

The rest of the paper is organized as follows. Section 2 describes the synthetic control method with a modified selection of control variables and the proposed SIHR model. The results of treatment effects and delay effects of mild interventions are reported in Section 3. We end this paper with some concluding remarks.

2. METHODS

2.1 Data Collection

Epidemic data were collected for daily cumulative confirmed cases of COVID-19 since January 19, 2020 in Shenzhen when the first confirmed case was reported[44]. Its corresponding population, latitude and areas were collected in the statistical yearbook. Also, the daily cumulative confirmed cases of COVID-19 for each county of each state in the United States were downloaded and available online[18] since March 1, 2020. The corresponding populations, areas, and latitudes were collected from the United States Census Bureau (USCB)[12, 13].

2.2 Effect of mild interventions

To examine the treatment effects of interventions implemented in Shenzhen, we envisioned that cumulative numbers of confirmed cases of COVID-19 per 100,000 persons in the population could be observed at the time t whether the interventions were implemented (treated) or not (not treated), which are denoted by $\{C_t^{(0)}(W_t)\}$.

$$C_t^{(0)}(W_t) = \frac{\text{Number of cumulative confirmed cases}}{\text{Resident population of infected region}} \times 100,000.$$

Here t is the time and W_t is the group indicator. "$W_t = 1$" corresponds to "exposed to interventions", and "$W_t = 0$" otherwise. Thus, the potential outcomes at the time t are $\{C_t^{(0)}(1), C_t^{(0)}(0)\}$ and the corresponding treatment effect of interventions is $C_t^{(0)}(1) - C_t^{(0)}(0)$. However, only one of the potential outcomes could be observed, i.e. the observed outcome at time t is defined as:

$$C_t^{(0)} = W_t C_t^{(0)}(1) + (1 - W_t) C_t^{(0)}(0). \tag{1}$$

Here the untreated outcomes $C_t^{(0)}(0)$ of "absence of interventions" were missing, which are the so-called counterfactual outcomes[38]. To formalize $C_t^{(0)}(0)$, we modified the synthetic control method (SCM)[2, 1] to construct the counterfactual outcomes by combining a number of untreated units instead of a single untreated unit.

Those potential untreated units were selected from the United States counties, where the interventions were not adopted at the early stage of epidemic. The epidemic data were reported on March 1, 2020 in the most infected counties of the United States[18]. However, the White House issued a "call to action" for coronavirus guidelines including cancelling gatherings over 10 people and staying at home starting on March 16, 2020[22]. There were no broad interventions implemented in the early stage of the epidemic in the United States, which allowed us to define a pre-intervention period from March 1 to March 16. Therefore, there were 16 days of duration before the interventions implemented in the United States. These 16-day epidemic data in the United States were used as the benchmark in our comparison with the 16-day epidemic data in Shenzhen from January 19 to February 3, 2020.

COVID-19 was detected and initially reported on January 19, 2020 in Shenzhen[44], and Guangdong province that includes Shenzhen as one of its cities activated "public responses" on January 23, 2020[34], so there were 4 days before Shenzhen implemented the interventions after the first case was reported. We use $T_0 = 4$ to denote this pre-intervention period, which is 4. We compared the data between Shenzhen from January 19 to February 3, 2020 and

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

counties in the United States from March 1 to March 16, 2020, by fixing the total duration at 16 days starting from the day of the first reported case in the locations of interest, i.e. "$t = 1$" means January 19 for Shenzhen, and March 1 for the counties of the United States. The effect of the post-intervention period at time t is $C_t^{(0)}(1) - C_t^{(0)}(0)$ for $t > T_0$. It can be estimated by:

$$C_t^{(0)}(1) - \hat{C}_t^{(0)}(0) := C_t^{(0)}(1) - \sum_{k=1}^{K} d_k C_t^{(k)}(0), \tag{2}$$

where $C_t^{(k)}(0)$ was the observed cumulative numbers of COVID-19 confirmed cases per 100,000 at the time t in k-th county of the United States in the absence of interventions. Thus, the counterfactual outcomes in the absence of the interventions of Shenzhen during the post-intervention period were calculated from a weighted linear combination of K untreated counties. The weights, d_k's, were non-negative and summed to 1[1]. Reliable estimates, $\hat{C}_t^{(0)}(0)$, can be obtained by constructing a synthetic region whose characteristics was similar to that in Shenzhen in the pre-intervention period[3].

The strategy is to find the important characteristics between Shenzhen and its "highly resembling" regions relating to the transmission of COVID-19 but not affected by the interventions. The literature suggests that the risk of transmission is higher in denser populations, and more urbanized regions may have more infected individuals[25, 15]. There is a large number of infected counties in the United States, and the population density of each county varies dramatically from near 0 to 18,721 persons per square kilometre. We selected the counties with the sizes of population densities (P) close to Shenzhen, i.e. metropolis areas in the United States.

We also considered latitude (L) of the counties because the latitude was found to be associated with the transmission of the COVID-19[39, 25, 40], before applying the data assimilation[37] procedure in the construction of the synthetic region.

Let $C_t := (C_t^{(0)}(0), C_t^{(1)}(0), ..., C_t^{(K)}(0))^T$ $(t \leq T_0)$. In order to estimate d_k appropriately, we assumed that P and L share a common least square structure related to SCM[1, 2]:

$$\begin{pmatrix} C_1^{(0)}(0) \\ \vdots \\ C_{T_0}^{(0)}(0) \\ P_s \\ L_s \end{pmatrix} \approx \begin{pmatrix} C_1^{(1)}(0) & ... & C_1^{(K)}(0) \\ & \vdots & \\ C_{T_0}^{(1)}(0) & ... & C_{T_0}^{(K)}(0) \\ P^{(1)} & ... & P^{(K)} \\ L^{(1)} & ... & L^{(K)} \end{pmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_K \end{pmatrix}$$

where P_s and L_s are the population density and the latitude of Shenzhen, respectively. Similarly, $P^{(k)}$ and $L^{(k)}$ are the population density and the latitude of the k-th county of the United States, respectively.

The literature[2, 1] recommended that not all the C_t, $t \leq T_0$ should be included in the estimation procedure, although the user can choose to include as many as M $(M \leq T_0)$ linearly independent combinations of pre-intervention outcomes [2]. Let

$$\mathcal{C} := (C_t)_{t=1:T_0}, \tag{3}$$

which can be viewed as the spatial-temporal data for our data. We performed principal component analysis (PCA)[19, 23] for the columns of $\mathcal{C}$ and evaluated the value of M and the linearly independent combinations of pre-intervention outcomes. The PCA reduced the redundant information of time signal and the impact of cumulative confirmed cases within the estimation procedure.

Let $(\text{Comp}_q^{(0)}, \text{Comp}_q^{(1)}, ..., \text{Comp}_q^{(K)})^T$ be the q-th component of the columns of $\mathcal{C}$ in the PCA. Also, let $Y := (P_s, L_s, \text{Comp}_1^{(0)}, ..., \text{Comp}_Q^{(0)})^T$ be a $(J \times 1)$ vector, where $J = Q + 2$, and Q is the chosen number of components. In other words, $Q := M$. We can choose Q to be the smallest number of the ranked principal components that explain at least 95% of the variance in $\mathcal{C}$. X is a $(J \times K)$ matrix whose k-th column is $(P^{(k)}, L^{(k)}, \text{Comp}_1^{(k)}, ..., \text{Comp}_Q^{(k)})^T$. The vector $D := (d_1, ..., d_K)^T$ is obtained by minimizing the distance $\| Y - XD \|_{\mathcal{H}}$:

$$\| Y - XD \|_{\mathcal{H}} = \sqrt{(Y - XD)^T \mathcal{H} (Y - XD)}, \tag{4}$$

where $\mathcal{H}$ is a $J \times J$ positive semidefinite matrix. The estimated weight $\hat{D}$ was obtained by Nelder-Mead methods[33, 27].

In addition to the population density and latitude, there are other factors related to the transmission of COVID-19. However, there were only 4 days in the pre-interventions period, the inclusion of more predictors in the linear construction of "synthetic Shenzhen" would result in collinearity during this short pre-intervention period; for example, temperature and latitude are correlated with each other. Alternatively, the Q components of pre-intervention outcomes included in the construction of "synthetic Shenzhen" can reflect the unobserved, shared, and time varying factors[3]. The gaps in the health care system might contribute to the variation in the number of confirmed cases in infected regions. However, the intervention and treatment of the COVID-19 depend on national and local authorities and resources. In the construction of "synthetic Shenzhen", we focus on the natural characteristics that are not easily affected by human decisions.

After the synthetic region of Shenzhen was constructed, an inference for the treatment effect of interventions was performed by a placebo test[3]. A "placebo effect"[4] was obtained by assigning an arbitrary sample of K counties in the United States as a "treated unit", Shenzhen and the remaining $(K - 1)$ counties synthesized to a combined untreated unit, the difference (gap) of COVID-19 cases per

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

100,000 between any of the "treated unit" and corresponding combined untreated unit during post-intervention period formed to a permutation distribution. If the value of this gap from Shenzhen was extreme compared to the remaining gaps in the permutation distribution, the treatment effect of interventions implemented in Shenzhen was considered statistically significant[2].

2.3 The delay effects of mild interventions

Besides the determination of treatment effect of intervention in Shenzhen, the timeliness of policy implementation may also affect the effectiveness of intervention. We considered the likely outcomes of the interventions if they were delayed by different days by using a SIHR model[42], which divided the total population into four classes: S (Susceptible to infectious disease), I (Infected and infectious without isolation), H (Hospitalized in well-isolation), and R (Removed from H because of the recovery or death). Let $Z_t := (S_t, I_t, H_t, R_t)$ be the random numbers of compartments S, I, H, R at time t, s.t. $S_t + I_t + H_t + R_t = N$, where N was the total number of population. We assumed $(Z_t)_{t\in 1:T}$ followed the hidden Markov structure:

$$Z_{t+1} = \mathcal{T}_{t,\Theta}(Z_t) + \epsilon_t,$$

where $\mathcal{T}_{t,\Theta}$ was the dynamic operator indexed by the time t and parameters Θ. ϵ_t were some noise variables, which were independent at different times. We treated $I_{1:T}$ as the latent variables.

For the noise, we followed the literature in infectious disease[9] by assuming

$$Z_{t+1} \sim \mathrm{Mult}(N, \mathcal{T}_{t,\Theta}(Z_t)/N),$$

where $\mathrm{Mult}(n, p)$ was the density of the multinomial distribution with the total number n and the incident rate vector p. Based on the mechanism of the spread of epidemic, we assumed the following dynamical operator $\mathcal{T}_{t,\Theta}$[41]:

$$\mathcal{T}_{t,\Theta}(Z_t) = Z_t + \begin{pmatrix} -\int_t^{t+1} \beta(u)\frac{S(u)}{N} I(u)\ \mathrm{d}u \\ \int_t^{t+1} \beta(u)\frac{S(u)}{N} I(u) - \frac{I(u)}{L_{in}}\ \mathrm{d}u \\ \int_t^{t+1} \frac{I(u)}{L_{in}} - \gamma H(u)\ \mathrm{d}u \\ \int_t^{t+1} \gamma H(u)\ \mathrm{d}u \end{pmatrix},$$

where $(S(t), I(t), H(t), R(t)) = (S_t, I_t, H_t, R_t)$, and $(S(u), I(u), H(u), R(u))$ were determined by the dynamic system (A1) in the Appendix with the initial value $(S(t), I(t), H(t), R(t))$ $(t < u \le t+1)$. The functional parameter β was time-varying contact rate controlled by additional parameters β_0, m_1, m_2. L_{in} was the mean length of incubation period.

L_{in} and γ can be assumed to be time-varying in SIHR model. However, since $I_{1:T}$ is unobserved, it may be crucial to fix L_{in} to ensure identifiability, which was assumed to be 5 days for COVID-19[7, 21, 30]. Parameter γ corresponding to "the removed compartment" was considered in compartment H which did not affect the estimation of $\beta(t)$. Because we focused on the SIHR model, we assumed γ to be time-invariant for simplicity.

Let $\Theta := (\beta_0, m_1, m_2, \gamma)$ are the parameters needed to be estimated.

To reflect the effect of interventions, we used the logistic function[41] to simulate the decreasing trend of $\beta(t)$:

$$\beta(t) = \frac{\beta_0}{1 + \exp(\lambda_{m_2}(t - m_1 - m_2/2))}, \tag{5}$$

where λ_{m_2} was chosen as $\frac{2\log((1-\epsilon)/\epsilon)}{m_2}$ and ϵ was fixed to be 0.01. The smaller the values of m_1 and m_2, the earlier effectiveness and the stronger intensity of interventions were implemented, respectively[41]. The trend of $\beta(t)$ was shown in Figure 1.

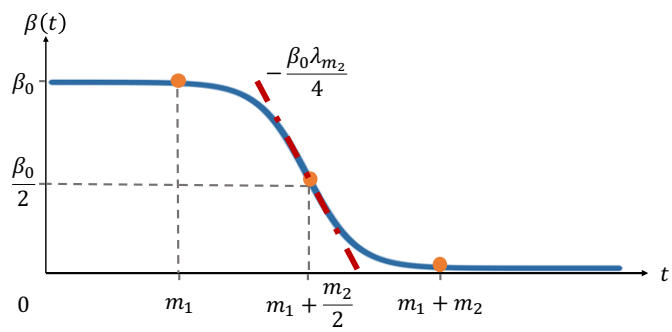

Figure 1. The trend of $\beta(t)$.

The slope of $\beta(t)$ at point $\frac{m_1+m_2}{2}$ is $-\frac{\beta_0\lambda_{m_2}}{4}$. A larger value of λ_{m_2} results in a larger slope, and the intensity of interventions would be stronger accordingly.

The time-varying reproduction number $\mathcal{R}_t$[24, 36] of SIHR model was calculated from:

$$\mathcal{R}_t = \frac{\beta_0 L_{in}}{1 + \exp(\lambda_{m_2}(t - m_1 - m_2/2))}. \tag{6}$$

Given the observed data $H_{1:T}$ and $R_{1:T}$, we used the posterior distribution:

$$\pi(\Theta, I_{1:T} | H_{1:T}, R_{1:T}, N, L_{in})$$

for Bayesian inference. Once the prior distribution of Θ: $\pi(\Theta)$ was given:

$$\pi(\Theta, I_{1:T} | H_{1:T}, R_{1:T}, N, L_{in}) \propto$$

$$\prod_{t=1}^{T-1} \mathrm{Mult}(Z_{t+1}; N, \mathcal{T}_{t,\Theta}(Z_t)/N)\ \pi(\Theta),$$

which could be randomly approximated by Gibbs sampler embedded Metropolis-Hastings steps[10, 6] since

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

$\pi(\Theta, I_{1:T}|H_{1:T}, R_{1:T}, N, L_{in})$ has no closed form. The full conditional distributions for the Gibbs sampler were:

$$p(\Theta|\cdot) \propto \prod_{t=1}^{T-1} \text{Mult}(Z_{t+1}; N, \mathcal{T}_{t,\Theta}(Z_t)/N)\ \pi(\Theta),$$

$$p(I_{1:T}|\cdot) \propto \prod_{t=1}^{T-1} \text{Mult}(Z_{t+1}; N, \mathcal{T}_{t,\Theta}(Z_t)/N).$$

The initial values for Gibbs sampling were chosen as: $\Theta^{(0)} = (1, 1, 1, 1)$ and $I_{1:T}^{(0)} = (1, ..., 1)$. After a burn-in period in the Markov chain, we used the remained samples to approximate the posterior distribution[41, 42].

For the prediction of spread, we used the posterior distribution ($s > T$):

$$\pi(\mu_{Z_s}|H_{1:T}, R_{1:T}, N, L_{in}), \tag{7}$$

where μ_{Z_s} was the conditional mean of Z_s given Z_T. Notice that:

$$\begin{aligned}
&\pi(\mu_{Z_s}|H_{1:T}, R_{1:T}, N, L_{in}) \\
&\propto \int p(\mu_{Z_s}, \Theta, I_{1:T}|H_{1:T}, R_{1:T}, N, L_{in})\ \mathrm{d}\Theta\ \mathrm{d}I_{1:T} \\
&\propto \int I(\mu_{Z_s} = \mathcal{T}_{s,T,\Theta}(Z_T))\ \pi(\Theta, I_{1:T}|H_{1:T}, R_{1:T}, N, L_{in})\ \mathrm{d}\Theta\ \mathrm{d}I_{1:T},
\end{aligned}$$

where $\mathcal{T}_{s,T,\Theta}(Z_T)$ was the vector at time s determined by the dynamic system (A1) in the Appendix with the initial value Z_T and parameters Θ. Hence the posterior distribution of μ_{Z_s} was calculated from $\pi(\Theta, I_{1:T}|H_{1:T}, R_{1:T}, N, L_{in})$.

The point estimation of parameters Θ, latent variables $I_{1:T}$ and the predicted $\mu_{Z_s}, s > T$ were presented as the median of the posterior distribution while 95% credible intervals were constructed with 2.5% and 97.5% quantiles.

Parameter m_1 reflected the effect of interventions over time. We varied the estimated values of m_1 to investigate the likely outcomes of delayed interventions. We considered the delay operator $\mathcal{D}_h$: $m_1 \rightarrow m_1 + h$ and let $\mu_{Z_s}^{(h)} := \mathcal{T}_{s,m_1,\Theta^h}(Z_{m_1}))$, where $\Theta^h = (\beta_0, \mathcal{D}_h(m_1), m_2, \gamma)$, $s > m_1$. Similarly, we used the posterior (7) of $(\mu_{Z_s}^{(h)})_{s>m_1}$ to simulate the potential outcomes of the delays of the interventions, where h varied from 0 to 5.

3. RESULTS

3.1 Synthetic control method (SCM) outcomes

Sixty-nine areas (68 counties in the United States and Shenzhen) were selected by using their corresponding population density over 500 people per square kilometre. SCM was used to combine 68 counties in the United States to construct a synthetic region of Shenzhen based on their latitude, population density and first two components from the PCA procedure described above, where the first two components explained over 95% of covariances of $\mathcal{C}$ in (3) for the pre-intervention period.

The synthetic and average values of latitude, population density, the first component, and the second component of $\mathcal{C}$ in (3) were shown in Table 1. During the pre-intervention period, the latitude was higher in the average of the 68 counties than in Shenzhen, while the population density was substantially lower on average in the 68 counties than in Shenzhen. Moreover, the values of two principal components of $\mathcal{C}$ in (3) were negative on the average of the 68 counties. In contrast, the "synthetic Shenzhen" exactly reproduced the values of population density, and two principal components of $\mathcal{C}$ in (3) which had in Shenzhen. A "highly resembling" region of Shenzhen was obtained by mainly combining Broward (Florida), New York (New York), Santa Clara (California), Pinellas (Florida), and Westchester (New York) with their corresponding weights 0.423, 0.335, 0.123, 0.032 and 0.025 as an optimal solution, respectively. The weights of all possible counties (removed counties with value of weight 0) with their latitude and population density were represented in Table A1 in the Appendix.

Table 1. The synthetic and average values of latitude, population density, the first component and the second component for 68 counties in the pre-intervention period.

Features	Shenzhen		Average of
	Real	Synthetic	68 counties
Latitude	22.55	33.16	38.15
Population density	6670	6669	1348
The first component	0.107	0.107	-0.002
The second component	0.035	0.035	-0.001

Before the mild intervention policies were implemented on January 23, 2020, the trends of actual cases per 100,000 in both Shenzhen and "synthetic Shenzhen" were highly similar, suggesting that such a synthetic region could be used to estimate the "counterfactual" results of Shenzhen. Based on Figure 2, after 1 day of the implementation of the policies, the growth rate of actual cases per 100,000 in Shenzhen was relatively slower than that of "synthetic Shenzhen" until January 29, 2020. After that, there was a sharp increase in the gap of cases per 100,000 between Shenzhen and "synthetic Shenzhen". On the 16th day from January 19, 2020, i.e. February 3, 2020, the estimated number of cases per 100,000 in "synthetic Shenzhen" was 11.87, which is 5.75 times of the actual COVID-19 cases per 100,000 observed (2.07) in Shenzhen (Figure 2).

A placebo test was performed to determine the significance level of the difference in the trends of COVID-19 cases per 100,000. To this end, we plotted the gap curves between Shenzhen and "synthetic Shenzhen" by in turn exchanging Shenzhen and each one of the 68 counties in the United States. The gap of COVID-19 cases per 100,000 between

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

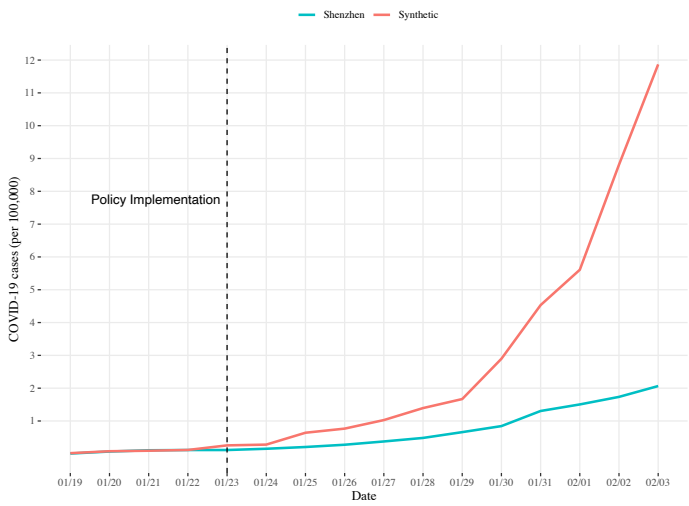

Figure 2. The trends of COVID-19 cases per 100,000 between Shenzhen and "synthetic Shenzhen" from January 19 to February 3, 2020. The black dashed line indicated the day of starting the intervention policies in Shenzhen.

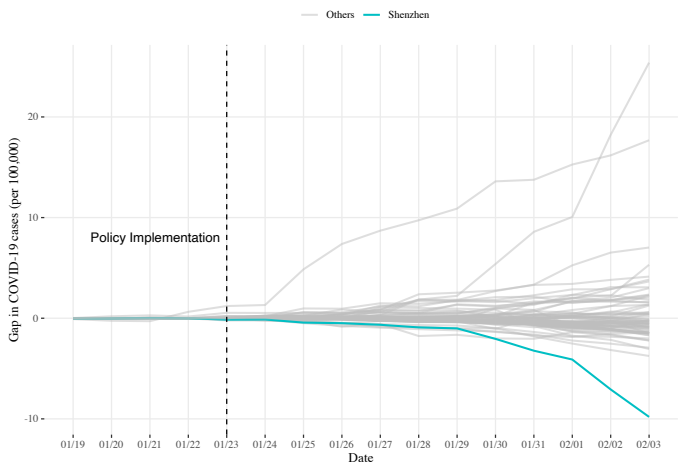

Figure 3. The permutation test of the treatment effects of implementing the policies in Shenzhen and the 68 control regions in the United States. All grey curves represent the placebo tests of COVID-19 cases per 100,000 gaps between a control area (an arbitrary random county of the 68 counties) and synthetic control city (a combination of remaining 67 counties and Shenzhen), and the blue curve represents the placebo test of COVID-19 cases per 100,000 gap between Shenzhen and "synthetic Shenzhen".

Shenzhen and our "synthetic Shenzhen" was the negatively largest among the negative gaps, i.e. the negative effect of the intervention on COVID-19 per 100,000 in Shenzhen was the lowest of all. For those 68 counties, the probability of having a gap for Shenzhen under a random permutation of the control measures was less than 5% (i.e. $1/69 = 0.014$), which is statistically significant in a conventional test[2]. This suggested that the mild intervention policies of Shenzhen might have significantly reduced the COVID-19 cases per 100,000 (Figure 3).

3.2 The likely outcomes of delayed interventions

There were mild but early interventions implemented in Shenzhen. It is worthy examining the likely outcomes of those interventions if they were delayed by different days (Figure 4). The observed COVID-19 cases per 100,000 from January 19 to February 3, 2020 (red points in Figure 4) were very close to the 0-day delay curve (i.e. the estimated cumulative confirmed cases per 100,000). Based on our simulation by delaying the interventions, the expected number of cumulative confirmed cases per 100,000 would be 2.32 times for a 4-day delay and, a larger, 2.51 times for a 5-day delay on February 3, 2020 (Figure 4).

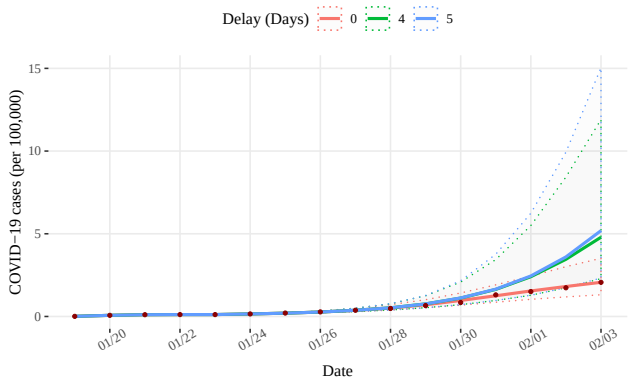

Figure 4. The expected number of cumulative confirmed cases per 100,000 with corresponding confidence intervals for mild interventions if they were delayed by 0, 4 and 5 days . The points represent the actual cumulative confirmed cases per 100,000 of Shenzhen.

3.3 Revisiting the results in terms of cumulative confirmed cases

We have presented our analysis and results based on the number of cumulative confirmed cases per 100,000 persons in the population. Here, we summarize our results based on

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

the number of cumulative confirmed cases.

If Shenzhen had not implemented mild intervention policies on January 23, 2020, the projected number of cumulative confirmed cases would be reach 1546 on February 3, 2020, which would be approximately 5.75 times of the actual number of cases.

The expected number of cumulative confirmed cases with corresponding 95% credible interval (CI) for different days of delay are summarized from January 29 to February 3, 2020 in Table 2. The full simulation results from January 19 to February 29, 2020 are presented in supplementary Appendix Figure A1. According to Table 2, the expected cumulative confirmed cases were 624 with the corresponding 95% CI (298,1551) for the 4-day delay and 676 with the corresponding 95% CI (303,1959) for the 5-day delay. Based on Figure A1, if the mild interventions were delayed for 4 or 5 days, the epidemic of COVID-19 in Shenzhen could be more severe than that in "synthetic Shenzhen" (1546 cumulative confirmed cases on February 3, 2020).

4. DISCUSSION

We used the daily reported cumulative confirmed case data for 16 days since January 19, 2020, in Shenzhen and the corresponding data for 16 days since March 1, 2020 in 68 counties in the United States serving as a control group, i.e., the "synthetic Shenzhen", those 68 United States counties were selected to match Shenzhen by latitude, population density and the first two components of $\mathcal{C}$ in (3) for the pre-intervention period. Those 68 United States counties did not implement systematic interventions during 16 days period.

By applying PCA to compress COVID-19 confirmed cases per 100,000 from the untreated units (counties in the United States), and combining the PCA with the latitude and population density in the estimation of weights through data assimilation, we were able to effectively estimate the model parameters based on the synthetic control method. In doing so, the latitude and population density as well as the principal components of $\mathcal{C}$ in (3) during pre-intervention period were considered as the characteristics of COVID-19.

The cases of COVID-19 per 100,000 in Shenzhen were significantly lower than those in "synthetic Shenzhen" after January 29, 2020. This indicated that the implementation of the mild intervention policies on January 23, 2020 in Shenzhen had a significant effect on controlling the epidemic of COVID-19. The reduction in the COVID-19 cases per 100,000 became larger and larger as time moved on (Figure 2).

The predicted cumulative confirmed cases (0-day delay) were very close to the observed cumulative confirmed cases of Shenzhen, indicating that our simulation could provide precise estimates[41]. Therefore, our simulated likely outcomes of delayed interventions were expected to be reliable. According to our simulation, there was little difference from January 23 to January 29, but the gap in the expected cumulative confirmed cases started increasing after January 29, 2020. This was consistent with the trend of the epidemic between Shenzhen and "synthetic Shenzhen" (Figure 2).

Therefore, it took about a week to see clear effects of interventions. If the mild interventions were not implemented, or were delayed by a week or longer, the epidemic of COVID-19 could have been severe. When the mild interventions were implemented early, the epidemic situation was significantly mitigated.

A novel step of our analysis strategy was to extend the use of SCM, which is a classical causal effect method and used it to create a synthetic reference population. Then, we introduced an epidemic compartmental SIHR model to evaluate the effectiveness of mild and early interventions in Shenzhen. Furthermore, the importance of the early interventions was revealed by simulating delays and assessing the consequences in the SIHR model. In conclusion, through the use of both SCM and SIHR models, the lessons were learned from the effect of timeliness of interventions to control the epidemic of COVID-19.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

Table 2. The likely outcome of intervention implemented delay in different days including the expected cumulative confirmed cases with corresponding 95% CI from January 29 to February 3, 2020.

Date	Actual	0 day	1 day	2 days	3 days	4 days	5 days
01/29	86	92 (68,132)	96 (69,146)	98 (69,154)	99 (69,159)	99 (69,163)	99 (69,165)
01/30	110	126 (89,185)	136 (91,216)	141 (93,238)	144 (93,258)	145 (93,271)	146 (93,281)
01/31	170	163 (112,249)	187 (122,309)	202 (125,363)	209 (126,407)	213 (126,448)	215 (126,490)
02/01	196	201 (135,318)	244 (155,415)	281(164,522)	300 (166,610)	313 (167,716)	319 (167,813)
02/02	226	236 (154, 392)	302 (188,518)	370 (211,685)	417 (225,862)	450 (226,1098)	469 (227,1294)
02/03	269	269 (171, 458)	355 (216, 616)	456 (260,865)	554 (288, 1175)	624 (298,1551)	676 (303,1959)

Table A1. The pool of counties in the United States to synthesize Shenzhen with their corresponding population density, latitude and weights.

States	Counties	Latitude	Population density	Weights
Guangdong	Shenzhen	22.55	6670	
Florida	Broward	26.19	570	0.423
New York	New York	40.78	18721	0.335
California	Santa Clara	37.22	571	0.123
Florida	Pinellas	27.90	619	0.032
New York	Westchester	41.15	747	0.025
Florida	Orange	28.51	536	0.016
Florida	Seminole	28.69	528	0.014
Texas	Bexar	29.45	616	0.008
Texas	Harris	29.86	1024	0.007
Georgia	Fulton	33.79	769	0.002
California	Los Angeles	34.20	816	0.001
California	Orange	33.68	1293	0.001
Georgia	Clayton	33.55	782	0.001
Georgia	Cobb	33.94	852	0.001
Georgia	DeKalb	33.77	1081	0.001
Georgia	Gwinnett	33.96	828	0.001
Texas	Dallas	32.77	1120	0.001
Texas	Tarrant	32.77	900	0.001

APPENDIX A

A.1 Effect of mild interventions

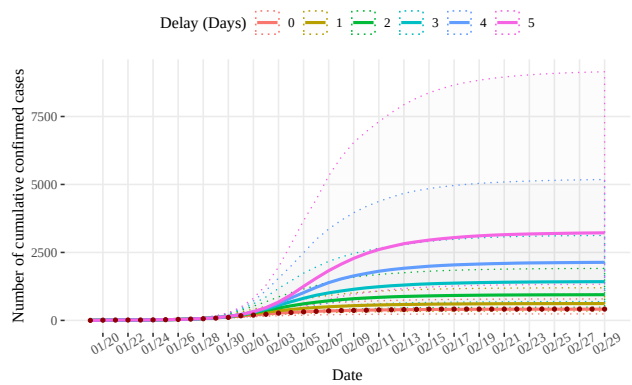

Figure A1. The expected number of cumulative confirmed cases with corresponding confidence intervals for mild interventions if they were delayed by different days from January 19 to February 29, 2020. The points represent the actual cumulative confirmed cases of Shenzhen.

A.2 Delay effect of mild interventions

The dynamic system of compartments S, I, H, R are as follows:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta(t)\frac{S(t)}{N}I(t), \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta(t)\frac{S(t)}{N}I(t) - \frac{I(t)}{L_{in}}, \\ \frac{\mathrm{d}H}{\mathrm{d}t} &= \frac{I(t)}{L_{in}} - \gamma H(t), \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma H(t), \end{aligned} \tag{A1}$$

where $(S(t), I(t), H(t), R(t))$ are the numbers of corresponding compartments at the time t. Given initial $(S(t_0), I(t_0), H(t_0), R(t_0))$ and the parameters $(\beta(t), L_{in}, \gamma)$, we can simulate the $(S(t_1), I(t_1), H(t_1), R(t_1))$, $t_1 > t_0$ by the deterministic dynamic system (A1). For simplicity, let $\beta(t) \equiv \beta(k)$, if $k \leq t < k+1$.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

A.3 Online Materials

The data and R files supporting the conclusions of this study are available in the https://github.com/tingT0929/The-timing-and-effectiveness-of-implementing-mild-interventions-of-COVID-19

ACKNOWLEDGEMENTS

We would like to thank all individuals who are collecting epidemiological data of the COVID-19 outbreak around the world.

Received 16 July 2020

REFERENCES

[1] Abadie, A. (2019). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature* 1–19.

[2] Abadie, A., Diamond, A. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program. *Journal of the American Statistical Association* **105** 493–505.

[3] Abadie, A., Diamond, A. and Hainmueller, J. (2011). Synth: An r package for synthetic control methods in comparative case studies. *Journal of Statistical Software* **42**.

[4] Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review* **93** 113–132.

[5] World Health Organization and others (2020). Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19).

[6] Andrieu, C., De Freitas, N., Doucet, A. and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning* **50** 5–43.

[7] Backer, J. A., Klinkenberg, D. and Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* **25**.

[8] Baidu (2020). Baidu Migration Data 2020. https://qianxi.baidu.com/.

[9] Becker, N. G. (1989). *Analysis of infectious disease data* **33**. CRC Press.

[10] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[11] Bjørnstad, O. N., Finkenstädt, B. F. and Grenfell, B. T. (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* **72** 169–184.

[12] The United States Census Bureau (2020). Population and Housing United Estimates Datasets.

[13] The United States Census Bureau (2020). Table of United States counties.

[14] Cantó, B., Coll, C. and Sánchez, E. (2017). Estimation of parameters in a structured SIR model. *Advances in Difference Equations* **2017** 33.

[15] Carozzi, F. (2020). Urban density and COVID-19. *IZA Discussion Paper Series* **13440**.

[16] Covid, C., COVID, C., COVID, C., Bialek, S., Gierke, R., Hughes, M., McNamara, L. A., Pilishvili, T. and Skoff, T. (2020). Coronavirus Disease 2019 in ChildrenUnited States, February 12–April 2, 2020. *Morbidity and Mortality Weekly Report* **69** 422.

[17] Ferguson, N. M., Donnelly, C. A. and Anderson, R. M. (2001). The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science* **292** 1155–1160.

[18] The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (2020). COVID-19 Data Repository). https://github.com/CSSEGISandData/COVID-19.

[19] Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning* **1**. Springer Series in Statistics New York.

[20] Wenzhou Municipal People's Government (2020). Notice of the Leading Group for the Prevention and Control of COVID-19 in Wenzhou (No. 6) 2020. http://www.wenzhou.gov.cn/art/2020/2/2/art_1219304_41867466.html.

[21] Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D. S. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine* **382** 1708–1720.

[22] The White House (2020). 30 Days to slow the spread.

[23] Izenman, A. J. (2008). *Modern multivariate statistical techniques*. Springer.

[24] Jones, J. H. (2007). Notes on R0. *California: Department of Anthropological Sciences* **323** 1–19.

[25] Kang, D., Choi, H., Kim, J.-H. and Choi, J. (2020). Spatial epidemic dynamics of the COVID-19 outbreak in China. *International Journal of Infectious Diseases* **94** 96–102.

[26] Kuznetsov, Y. A. and Piccardi, C. (1994). Bifurcation analysis of periodic SEIR and SIR epidemic models. *Journal of Mathematical Biology* **32** 109–121.

[27] Lagarias, J. C., Reeds, J. A., Wright, M. H. and Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization* **9** 112–147.

[28] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G. and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine* **172** 577–582.

[29] Li, M. Y. and Muldowney, J. S. (1995). Global stability for the SEIR model in epidemiology. *Mathematical Biosciences* **125** 155–164.

[30] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y. et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* **382** 1199–1207.

[31] Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., Yuan, B., Kinoshita, R. and Nishiura, H. (2020). 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** 538.

[32] Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H. et al. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300** 1966–1970.

[33] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7** 308–313.

[34] Health Commission of Guangdong Province (2020). First-level public health emergency response 2020. http://wsjkw.gd.gov.cn/zwyw_yqxx/content/post_2878895.html.

[35] National Health Commission of the People's Republic of China (2020). Distribution of COVID-19 cases in the world 2020. http://2019ncov.chinacdc.cn/2019-nCoV/global.html.

[36] Potterat, J. J., Zimmerman-Rogers, H., Muth, S. Q., Rothenberg, R. B., Green, D. L., Taylor, J. E., Bonney, M. S. and White, H. A. (1999). Chlamydia transmission: concurrency, reproduction number, and the epidemic trajectory. *American Journal of Epidemiology* **150** 1331–1339.

[37] Robinson, A. R. and Lermusiaux, P. F. (2000). Overview of data assimilation. *Harvard Reports in Physical/Interdisciplinary Ocean Science* **62** 1–13.

[38] Rubin, D. B. (1978). Bayesian inference for causal effects: The

medRxiv preprint doi: https://doi.org/10.1101/2020.06.22.20137380; this version posted September 5, 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-NC-ND 4.0 International license .

role of randomization. *The Annals of Statistics* 34–58.

[39] Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F. and Amoroso, A. (2020). Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. *Available at SSRN 3550308*.

[40] Sarmadi, M. et al. (2020). Association of COVID-19 global distribution and environmental and demographic factors: An updated three-month study. *Environmental Research* 109748.

[41] Tan, J., Jiang, Y., Tian, T. and Wang, X. (2020). P-SIHR Probabilistic Graphical Model: an Applicable Model of COVID-19 in Estimating the Number of Infectious Individuals. *Acta Mathematicae Applicatae Sinica* **43** 365–382.

[42] Tian, T., Zhang, J., Lin, S., Jiang, Y., Tan, J., Li, Z. and Wang, X. (2020). Data-driven analysis on the simulations of the spread of COVID-19 under different interventions of China. *medRxiv*.

[43] Wu, J. T., Leung, K. and Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* **395** 689–697.

[44] Wu, T., Ge, X., Yu, G. and Hu, E. (2020). Open-source analytics tools for studying the COVID-19 coronavirus outbreak. *medRxiv*.

Ting Tian
School of Mathematics
Sun Yat-Sen University
Guangzhou, 510275
China
E-mail address: tiant55@mail.sysu.edu.cn

Wenxiang Luo
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: luowx8@mail2.sysu.edu.cn

Jianbin Tan
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: tanjb6@mail2.sysu.edu.cn

Yukang Jiang
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: jiangyk3@mail2.sysu.edu.cn

Minqiong Chen
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: mcp04chm@mail3.sysu.edu.cn

Wenliang Pan
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: panwliang@mail.sysu.edu.cn

Songpan Yang
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: yangsp5@mail2.sysu.edu.cn

Jiashu Zhao
School of Mathematics
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: zhaojsh@mail2.sysu.edu.cn

Xueqin Wang
School of Statistics
Capital University of Economics and Business
Beijing, 100070
China
E-mail address: hawkingwang@gmail.com

Heping Zhang
School of Public Health
Yale University
New Haven, CT 06520-8034
the United States
E-mail address: heping.zhang@yale.edu