# Impact of Modelling Approach on the Estimation of Government Intervention Effects in COVID-19

Lun Liu ( ✉ liu.lun@pku.edu.cn )

Peking University

**Zhu Zhang**

Peking University

**Hui Wang**

Tsinghua University

**Shenhao Wang**

Massachusetts Institute of Technology

**Shengsheng Zhuang**

Peking Union Medical College

---

# Impact of modelling approach on the estimation of government intervention effects in COVID-19

**Authors:** Lun Liu[1,2*], Zhu Zhang[1], Hui Wang[3*], Shenhao Wang[4,5], Shengsheng Zhuang[6]

**Affiliations:**

[1]School of Government, Peking University, China.

[2]Institute of Public Governance, Peking University, China.

[3]School of Architecture, Tsinghua University, China.

[4]Department of Urban Studies and Planning, Massachusetts Institute of Technology; Cambridge, United States.

[5]Media Lab, Massachusetts Institute of Technology; Cambridge, United States.

[6]Peking Union Medical College, China.

*Correspondence to: liu.lun@pku.edu.cn, wh-sa@mail.tsinghua.edu.cn

# Abstract

The efficacy of government interventions in epidemic has become a hot subject since the onset of COVID-19. There is however much variation in the results quantifying the effects of interventions, which is partly related to the varying modelling approaches employed by existing studies. This paper therefore aims to examine how the choice of modelling approach would affect the estimation results of intervention effects, by experimenting with different modelling approaches on a same data set composed of the 500 most affected U.S. counties. We compare the most frequently used methods from the two classes of modelling approaches, which are Bayesian hierarchical model from the class of computational approach and difference-in-difference from the class of natural experimental approach. We find that computational methods are likely to produce larger estimates of intervention effects due to simultaneous voluntary behavioral changes. In contrast, natural experimental methods are more likely to extract the true effect of interventions. Among different difference-in-difference estimators, the two-way fixed effect estimator seems to be an efficient one. Our work can inform the methodological choice of future research on this topic, as well as more robust re-interpretation of existing works, to facilitate both future epidemic response plans and the science of public health.

# Introduction

In fighting COVID-19, non-pharmaceutical interventions aiming at reducing mobility and contact have been implemented by governments around the world repetitively. Given the strong (and mostly negative) impact of these interventions on economy and social life, the efficacy of interventions has become a hot subject of study[1-5], which is informative for both future epidemic response plans and the science of public health. However, there is much variation in the analysis results quantifying the effects of the interventions, which prevents researchers and policy makers from drawing clear and reliable conclusions (estimates with the reproduction number as the outcome of interest are summarized in Table S1). There are two major sources of such variation: the first is the sample data (countries, regions) used in the analyses, as the effects of interventions might be naturally different in different local contexts influenced by the intensity of enforcement, lifestyle, culture and so on; the second is the modelling approach, which might capture different structures in the data thus needs careful evaluation.

The modelling approaches employed by existing analyses can be grouped into two classes: those that incorporate a natural experimental design and those that do not (review of the methods used in 23 highly relevant studies in Table S1). The former usually employ tools from the field of econometrics or biostatistics which have been developed for identifying causal influences in public policy and medical issues. The natural experimental design is featured by comparing regions that implemented or relaxed an intervention in a period (treatment group) to regions that did not (control group), given that the latter group is similar enough to the former group to be considered as the counterfactual. The dominant method used has been difference-in-difference[4,6-11], a method that identifies causal treatment effect through comparing pre-and post-treatment performance of control and treatment groups[12]; and occasionally synthetic control[13] and interrupted time series analysis[14]. The second class of analyses do not incorporate a treatment-control comparison but focus on maximizing model fit to observed data based on assumed model structures. Specific methods that have been used include Bayesian hierarchical model[2,3,15-19], generalized linear regression[5,20-25] and certain machine learning models[5].

The assumed advantage of natural experimental design lies in the ability to rule out common temporal changes in the treatment and control groups, which might otherwise be mistaken as the treatment effect, thus identify the true effect of an intervention. To be more specific, if there is a natural trend in the epidemic happening together with an intervention effect, researchers would be able to rule it out by subtracting the epidemic course of the treatment group with that of the control group. In the case of COVID-19, multiple processes could lead to such a natural trend,

including increased precautionary behavior driven by the severity of the epidemic, decrease in the proportion of susceptible population, virus variants, etc.[9].

This paper therefore aims to empirically examine how the choice of modelling approach would affect the estimation results of intervention effects, by experimenting with different modelling approaches on a same data set. We focus on the most frequently used methods from the two classes, which are difference-in-difference from the class of natural experimental methods and Bayesian hierarchical model from the class of more computational methods (can be checked in Table S1). It should be noted that among the non-natural experimental methods, generalized linear regression is also frequently used, however, most of the model settings are not able to capture the causal relationships and only association can be claimed from the analysis results[20,21,24,26]. More specifically, in terms of Bayesian hierarchical model, we use the model developed by Brauner et al. (2021), which is a later modification in a group of similar models[3,19]. This model uses infection case and death data in each region to backward infer daily reproduction numbers and then the impacts of interventions on the reproduction number. In terms of the difference-in-difference method, we test two estimators. The first is the two-way fixed effect estimator, which is a widely used difference-in-difference estimator in policy analysis and the most used one in relevant studies[12]. The estimation is implemented through a linear model with the outcome of interest as dependent variable and intervention status per region per day as well as region and day fixed effects as independent variables. Estimating the linear model is equivalent to comparing the changes in treatment and control groups before and after intervention when there are two groups and two time periods. However, this estimator could be biased if there are multiple time periods and the treatment effect changes over time, which is possible in the case of COVID-19 as the compliance to interventions may change (e.g. it may take some time for compliance to increase, and compliance may also decrease as time goes)[19,27]. To accommodate the heterogeneous effect, we further experiment with a new difference-in-difference estimator that is robust to temporally heterogeneous treatment effect[28,29]. The robust estimator directly compares the epidemic course after $n$ days of intervention status change in all possible pairs of treatment and control groups across the entire study period and computes the average treatment effect ($n \geq 0$).

We take the United States as the case and analyze the effect of government interventions using county-level data of 500 counties with the largest number of infections. We use data from the first pandemic wave, that is, from March to August 2020, since intervention effect estimates in

later periods could be confounded by more factors including lockdown fatigue, virus variants and vaccination. Six widely applied government interventions are evaluated, which are stay-at-home order, school closure, childcare closure, non-essential retail closure, banning small-size gatherings (below 10 people), and banning large-size gatherings (above 10 people) (Fig. 1).

Accurate knowledge of the effectiveness of government interventions is key to cost-effective policy making not only for controlling the on-going COVID-19 but also for future public health crisis. By directly experimenting with three main stream estimation methods, this work aims to facilitate a better understanding on the strength and weakness of relevant modelling approaches and the potential bias of reported results. While we cannot fully encompass all possible methodology, our analysis is nonetheless informative on refining the analysis on this issue and grounding decision making on sound scientific evidences.
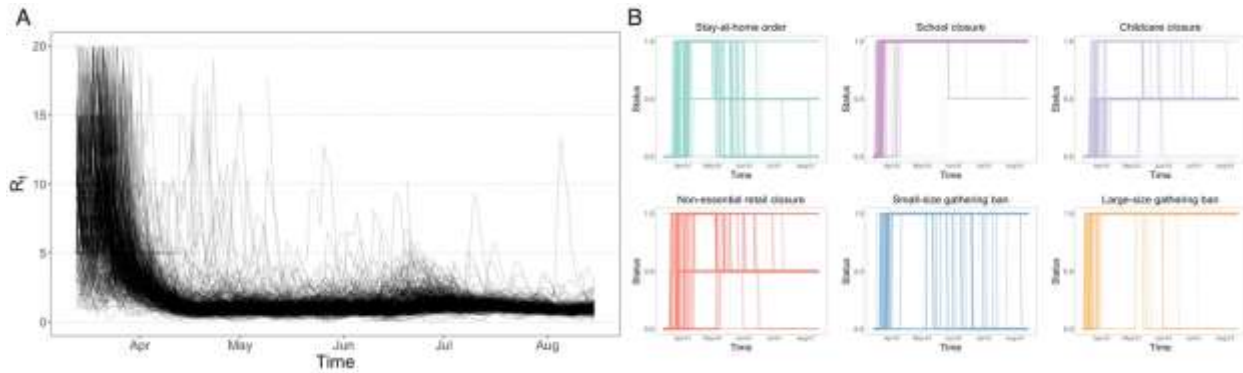


**Fig. 1. Epidemic course and interventions in the sample counties in the United States. (A)** Dynamics of daily $R_t$ in the sample counties. **(B)** Dynamics of the six interventions. The interventions switched between none (0), partial restriction (0.5) and full restriction (1) during the study period. The widths of lines in the graphs are proportional to the number of counties in the corresponding status on the corresponding day.

# Result

**Existence of common behavioral trends**

Since the key difference between natural experimental and non-natural experimental methods lies in the use of a control group to rule out universal trend in the epidemic, we start by testing whether such trend exist and whether it correlates with the implementation or relaxation of interventions. Among the possible contributors to the universal trend, we focus on residents' behavioral changes, which is supposed to be a major contributor and the information is available

through Google's COVID-19 Community Mobility Reports that provides daily indicators on the movements of residents in a region (the county level for the United States)[30].

By regressing the county-specific daily mobility indicators on day variables and intervention status, we find a statistically significant day effect on most days in our study period, indicating a non-zero common change in the amount of travel conducted by residents in different counties on each day (Fig. 2). There was first an upward common trend in the time staying at home from March to mid-April, followed by a gradual decrease afterwards, yet by the end of the period, the time staying at home was still higher than before the pandemic. Actually, the day effects account for a larger proportion of the total variance in mobility than government interventions, as the adjusted $R^2$ of the model is 0.22 with only intervention status as explanatory variables and increases to 0.66 when day effects are added (full results in Table S2). Correlation tests show that this common behavioral trend is statistically significantly correlated with the status of interventions (Pearson's $r$=0.02~0.39, all statistically significant at 0.001 level). The positive correlation suggests that the estimation of intervention effects could be upwardly biased by simultaneous spontaneous behavioral changes without proper controlling measures, which confirms the concern behind natural experimental methods.
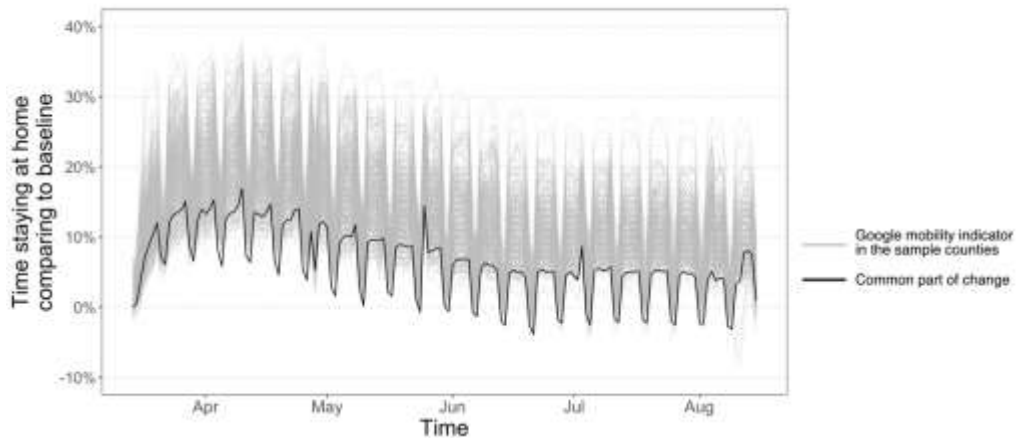


**Fig. 2. Trend of the common change in residents' mobility.** Grey lines are the daily values of the Google mobility indicator in the sample counties and the black line is the common daily mobility change shared by the counties. Please note that the baseline in the Google mobility indicator is the median value on each day of the week from the 5-week period Jan 3 to Feb 6, 2020, in other words, different baselines are used for different days of the week.

**Comparison of model results**

The intervention effect estimates produced by the three methods are shown in Fig. 3, which are different from each other given the diverse modelling strategies. Generally speaking, the semi-mechanistic model produces the largest estimates on interventions' effects in reducing $R_t$, which is consistent with our reasoning that non-natural experimental methods are likely to produce larger estimates. Three interventions are estimated to have a statistically significant impact on reducing $R_t$, which are school closure (reducing $R_t$ by -67.4%, 95% confidence interval: -69.7 to -64.8%), childcare closure (reducing $R_t$ by -17.6%, -24.5 to -9.7%), and non-essential retail closure (reducing $R_t$ by -6.9%, -14.9 to -1.2%). The particularly large effect of school closure might be related to the interaction between the sequence of interventions and structure of the model[31].

However, none of these three interventions are found to have a statistically significant impact according to the two difference-in-difference estimators, which could be a consequence of excluding spontaneous behavioral changes by the natural experiment design. Generally, estimation results acquired from the two difference-in-difference estimators are closer to each other than to the results of the semi-mechanistic model (differences are not statistically significant). Stay-at-home order is found to have the most prominent effect in reducing $R_t$ by both of the two difference-in-difference estimators: -6.6% (-10.6 to -2.5%) as estimated by the two-way fixed effect estimator and -17.3% (-31.7 to 0%) as estimated by the robust estimator, which nonetheless are not statistically different (z-statistics = 1.27). The wider confidence interval of the latter could be because the estimation is on a day-by-day basis so that a smaller sample is used in estimating the impact of implementing or relaxing an intervention for $n$ days ($n \in \{0 \sim 21\}$), resulting in more uncertainties. Besides, banning small-size gatherings is also found to statistically significantly reduce $R_t$ by the two-way fixed effect estimator (-5.1%, -9 to -1.1%), while its estimate is associated with a wide confidence interval when using the robust estimator. The other interventions are not found to have a statistically significant impact by both estimators. All the interventions satisfy the parallel pre-trend assumption, which means that the epidemic course in the control group is similar to that of the treatment group before intervention status changes, so that the former can be proper counterfactuals for the latter (details on the methodology and results of parallel pre-trend test in Supplementary Method and Table S3&4).

Since the difference between the results from the two difference-in-difference estimators should stem from the temporally heterogeneous intervention effects, we then examine the day-by-day estimates of intervention effects produced by the robust difference-in-difference estimator (Fig.

3B). The results suggest that the intervention effects do change with the length of implementation. Among the two interventions that are found to have significant impacts on reducing $R_t$ in the two-way fixed effect model, stay-at-home order shows an increasing effect over time, which might be caused by the gradual change of behavior and enforcement measures; and small-size gathering ban demonstrates a V-shaped trend, suggesting a loss of effect after initial effectiveness.
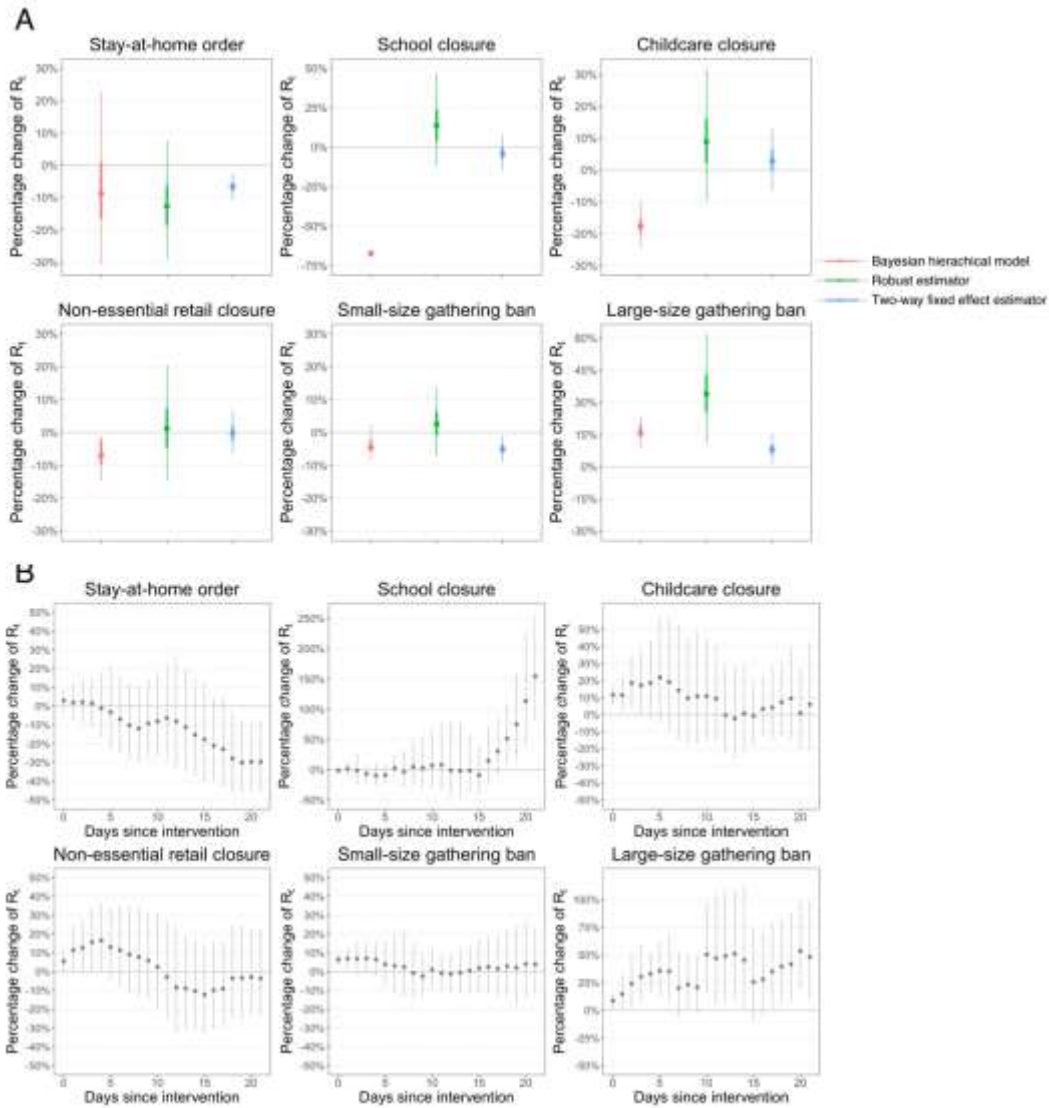


**Fig. 3. Intervention effect estimates from different modelling approaches.** **(A)** Comparison of the results from the three approaches, with 50% and 95% confidence intervals. **(B)** Time-varying intervention effects estimated by the robust difference-in-difference estimator. Point estimates of percentage changes in $R_t$ with 95% confidence intervals are shown.

# Discussion

The wide use of non-pharmaceutical interventions in COVID-19 provides large-scale natural experiments on the impacts of various public health measures, which are valuable information for fostering the knowledge in relevant fields such as public health, urban resilience, public policy and informing future policy making. Diverse methodologies have been developed to evaluate the efficacy of interventions in COVID-19. By testing three major modelling methods on the epidemic data from the United States, we demonstrate that quite different estimates of intervention effects could be acquired with different methods. Particularly, methods without explicit comparison of treatment and control groups tend to over-estimate intervention effects, as we find there is non-negligible spontaneous behavioral change happening together with the interventions. This is further testified by comparing our estimates to previous analyses—our estimates from difference-in-difference methods are generally smaller than estimates from non-natural experimental methods[2,17,20,23], though this comparison is not absolutely valid due to varying geographical extents of the studies. The results suggest that estimates produced by methods with a natural experimental design could be more reliable, despite of differences caused by variance in details of modelling approaches. Our work can inform the methodological choice of future research on this topic as data are still accumulating, and also facilitate more robust re-interpretation of existing works.

Regarding to the two estimators with natural experimental design, the one robust to temporal heterogeneity in intervention effects is methodologically more superior. However, since much less observations are available for each $n$-day effect estimate, there could be less certainty about the estimate thus wider confidence intervals are produced. Considering that the results from the two difference-in-difference estimators are not statistically significantly different, the two-way fixed effect estimator is a preferable choice, which also has the advantage of fast computing. On our data, it takes 42 CPU hours (2.6 GHz) to finish bootstrapping 100 times for the estimation of one intervention. Nonetheless, if longer period data is collected and used, the robust estimator might be able to draw narrower confidence intervals thus become a more preferable choice.

Nonetheless, when employing the two-way fixed effect estimator, one should be aware of the possible bias, which is jointly influenced by the intervention effect in each treatment-control pair and their distribution across time[28,29]. Therefore, the difference between the results of the two difference-in-difference estimators in our analysis might not hold for other countries and time periods thus could not be taken as a reference. Neither might the temporal trends of intervention

effects in this analysis (Fig. 3B) hold in alternative contexts, which could be influenced by the dynamics of enforcement intensity and compliance.

A limitation to our methodological analysis could be the potential signaling or spillover effect of interventions. The former refers to the possibility that increased intervention in treatment regions makes people in the control group alerted and more cautious, or the reversed in the reopening stage[9]. The latter refers to the possibility that intervention in treatment regions also reduces (or increases in reopening) infections in the control group due to change in cross-region travel[7]. Both cases would deflate the intervention effect estimates in all the methods.

Although significant bias is found with non-natural experimental models in estimating intervention effects, this can be modified by including model elements that simulate voluntary behavioral changes in response to the severity of epidemic. Our conclusions on modelling approach comparison may also apply to the evaluation of intervention effects in other areas which involve massive behavioral changes such as disaster recovery and crime mitigation.

# Method

## Data

The analysis involves daily infection case and death data, county-level mobility data, and government intervention data on 500 counties with the most infection cases in the United States between March 13 and August 15 2020. The infection case and death data is downloaded from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[32]. The information on people's mobility is from Google's COVID-19 Community Mobility Reports[30]. The time of implementing and relaxing interventions is collected manually from the official websites of state governments, which is the major level of government in charge of intervention policy making in COVID-19 in the United States. We do not account for the few cases in which county or local governments take alternative actions than the orders of state governments, which are only occasional[7]. The interventions are coded 0, 0.5 or 1 based on the intervention status in each region on a given day: 0 if there was no relevant restriction, 1 if a full restriction was implemented, and 0.5 if there was a partial restriction (e.g. recommending instead of enforcing a restriction, shortening opening hours or limiting occupancy instead of a complete closure).

**Testing common behavioral trends**

We examine whether there is a common behavioral trend across all sample counties using Google's COVID-19 Community Mobility Reports. This dataset provides the daily percentage change in the visits to different categories of places (or the time spent at places of residence) comparing with baselines (different baselines for each day of the week, using the median value from the 5-week period Jan 3 to Feb 6, 2020). We focus on the 'residence' indicator, referring to the percentage change in the time spent at home, since it can reflect the overall change of mobility. Next, we test whether there is a 'common component' in the time people staying at places of residence each day across counties after controlling for interventions, by regressing the mobility indicator against the daily status of interventions in each county and a series of binary variable indicating each day of the study period (i.e. day-fixed effect). If there is a day effect, then the coefficient for the day variable should be significantly different from zero. The regression model is expressed as follows

$$m_{c,t} = \sum_{i \in I} \beta_i x_{i,c,t} + \gamma_t + e_{c,t} \quad (1)$$

where $m_{c,t}$ denotes the value of the mobility indicator in county $c$ on day $t$; $x_{i,c,t}$ denotes the status of intervention $i$ in county $c$ on day $t$ and $\beta_i$ denotes coefficients; $I$ is the set of interventions being studies; $\gamma_t$ denotes the shared change of mobility on day $t$; $e_{c,t}$ is the error term. $\beta_i$ and $\gamma_t$ are estimated using ordinary least square method.

**Experiment 1. Bayesian hierarchical model**

We use the model developed by Brauner et al. (2021), which is an extension of the original model proposed by Flaxman et al. (2020). The semi-mechanistic model is composed of a series of equations that link intervention status to instantaneous reproduction number, and then to the total number of infections and the number of observed cases and deaths. The key equations are as follows.

$$R_{t,c} = R_{0,c} \prod_{i=1}^{I} exp\left(-\alpha_{i,c} x_{i,t,c}\right) \quad (2)$$

$$R_{t,c} = \frac{1}{M\left(-\log\left(1 + g_{t,c}\right)\right)} \quad (3)$$

$$N_{t,c}^{(C)} = N_{0,c}^{(C)} \prod_{\tau=1}^{t} \left[\left(1 + g_{\tau,c}\right) \cdot exp\left(\varepsilon_{\tau,c}^{(C)}\right)\right] \quad (4)$$

$$N^{(D)}_{t,c} = N^{(D)}_{o,c} \prod_{\tau=1}^{t} \left[ \left( 1 + g_{\tau,c} \right) \cdot exp\left( \varepsilon^{(D)}_{\tau,c} \right) \right] \quad (5)$$

$$\bar{y}^{(C)}_{t,c} = \sum_{\tau=0}^{31} N^{(C)}_{t-\tau,c} P_C( delay = \tau) \quad (6)$$

$$\bar{y}^{(D)}_{t,c} = \sum_{\tau=0}^{47} N^{(D)}_{t-\tau,c} P_D( delay = \tau) \quad (7)$$

where $R_{o,c}$ and $R_{t,c}$ denote the basic reproduction number in county $c$ and the instantaneous reproduction number on day $t$ in county $c$; $x_{i,t,c}$ denotes the status of intervention $i$ in county $c$ on day $t$ and $I$ is the number of interventions; $\alpha_{i,c}$ denotes the effect of intervention $i$ in county $c$, which is assumed to be county-specific and independent multiplicative on $R_{t,c}$. $\alpha_{i,c}$ is assumed to follow a normal distribution $\alpha_{i,c} \sim$ Normal($\alpha_i$, $\sigma^2_i$) and the prior distributions for $\alpha_i$ and $\sigma_i$ are $\alpha_i$ ~ Asymmetric Laplace($m=0$, $\kappa=0.5$, $\lambda=10$) and $\sigma_i \sim$ Half Student-T($\nu=3$, $\sigma=0.04$). In Eq. 3, $g_{t,c}$ denotes the daily growth rate of infections and M($\cdot$) is a moment-generating function; the equation can be solved with information on the generation interval, which is the time between successive infections in a transmission chain. In Eq. 4 to 7, $N^{(C)}_{t,c}$ and $N^{(D)}_{t,c}$ denote the number of infections that subsequently become confirmed cases or reported deaths, respectively; $g_{\tau,c}$ is the daily growth rate; $\varepsilon^{(\cdot)}_{\tau,c}$ is a noise term; $P_C(delay)$ and $P_D(delay)$ are the distribution of the delay from infection to confirmation and from infection to death, respectively. $\varepsilon^{(\cdot)}_{\tau,c}$ is assumed to follow a normal distribution $\varepsilon^{(\cdot)}_{\tau,c} \sim$ Normal(0, $\sigma_N$) and a prior distribution is placed on $\sigma_N \sim$ Half Student-T($\nu=3$, $\sigma=0.15$). $N^{(C)}_{0,c}$ and $N^{(D)}_{0,c}$ are placed with uninformative priors. The model estimates the posterior distributions of the parameters, including the effects of interventions ($\alpha_{i,c}$), using a Markov Chain Monte Carlo sampling algorithm. More details about the model can be found in Brauner et al. (2021).

**Experiment 2. Difference-in-difference: Two-way fixed effect estimator**

Two-way fixed effect model is a widely used modelling method to implement difference-in-difference analysis[12]. The model regresses the outcome of interest on the treatment variable and, at the same time, controls for unit fixed effects and time fixed effects, the latter of which helps rule out the common trends in intervention and control groups at each time point. The model is expressed as follows.

$$log( R_{c,t}) = \sum_{i \in I} \beta_i x_{i,c,t} + \alpha_c + \tau_t + \varepsilon_{c,t} \quad (8)$$

where $log(R_{c,t})$ denotes the log-transformed instantaneous reproduction number in county $c$ on day $t$; $x_{i,c,t}$ denotes the status of intervention $i$ in county $c$ on day $t$ and $\beta_i$ denotes their

coefficients; $\alpha_c$ and $\tau_t$ denote the unit and time fixed effects, respectively; and $\varepsilon_{c,t}$ is the error term. We estimate robust standard error in a way that allows $\varepsilon_{c,t}$ to correlate at the county level[33].

We use $log(R_{c,t})$ as the dependent variable to match with the Bayesian hierarchical model, whose parameter estimation is interpreted as the proportional change of $R_t$ by interventions. The coefficients of intervention variables in this model can be interpreted in the same way by taking $exp(\beta_i)$. Daily $R_t$ in each county is estimated with the method proposed by Cori et al.[34], with a 7-day sliding time window to smooth infection numbers and parameters of the serial interval of infections based on previous epidemiological investigation of COVID-19 (mean=7.5 days, standard deviation=3.4 days)[35]. $R_t$ estimates with coefficient of variation greater than 0.3 are excluded from further analysis (mainly due to small case numbers in the corresponding time window).

**Experiment 3. Difference-in-difference: Robust estimator**

The heterogeneity robust difference-in-difference estimator is proposed by de Chaisemartin and D'Haultfoeuille[28]. This estimator estimates the average treatment effect of implementing an intervention for $k$ days ($k \geq 0$) across all the regions where an intervention switched from level $d$ ($d \in \{0, 0.5, 1\}$) on $t$-$1$ to $d'$ ($d' \in \{0, 0.5, 1\}$, $d' \neq d$) on $t$ and remained $d'$ at least till day $t$+$k$ (switching regions), using the regions where the intervention remains $d$ between $t$-$1$ and $t$+$k$ as controls (stable regions). Since the treatment effect is estimated through explicit matching between switching and stable regions and analyzing their difference after a given number of days of treatment, temporal heterogeneity will not create bias. The estimator is computed as follows

$$DID_{i,k} = \sum_{t=2}^{T-k} \sum_{(d,d') \in D^2, d \neq d'} \frac{N_{i,d,d',t,k}}{\sum_{(c,t):t \geq 2} |D_{i,c,t} - D_{i,c,t-1}|} DID_{i,d,d',t,k} \quad (9)$$

$$DID_{i,d,d',t,k} = [1\{d'>d\} - 1\{d>d'\}] \left[ \sum_{(c,t,k):D_{i,c,t+k}=\dots=D_{i,c,t}=d', D_{i,c,t-1}=d, t \geq 2} \frac{e_{i,c,t,k}}{N_{i,d,d',t,k}} - \sum_{(c,t,k):D_{i,c,t+k}=\dots=D_{i,c,t}=d, D_{i,c,t-1}=d, t \geq 2} \frac{e_{i,c,t,k}}{N_{i,d,d,t,k}} \right] (10)$$

$$\log(R_{c,t+k}) - \log(R_{c,t-1}) = \left( X'_{i,c,t+k} - X'_{i,c,t-1} \right) \gamma + b + e_{i,c,t,k} \quad (11)$$

where $DID_{i,k}$ denotes the estimated effect of intervention $i$ after $k$ days of implementation; $t \in \{1,\dots,T\}$ denotes the observed dates; $D_{i,c,t}$ denotes the status of intervention $i$ in county $c$ on day $t$; $N_{i,d,d',t,k}$ and $N_{i,d,d,t,k}$ denote the number of switching counties and stable counties between day $t$

and $t+k$. If either $N_{i,d,d',t,k}$ or $N_{i,d,d,t,k}$ is 0, then $DID_{i,d,d',t,k}$ is defined as 0. $e_{i,c,t,k}$ denotes the change of the outcome variable, which is the logarithm of instantaneous reproduction number, between switching and stable counties after controlling for the impact of other interventions ($X'_{i,c,t}$); $e_{i,c,t,k}$ is estimated through Eq.11 with ordinary least square using only stable observations; $\gamma$ is a vector of coefficients and $b$ is the constant. The standard error of $DID_{i,k}$ is estimated by bootstrapping for 100 times. Further, to account for the correlation of errors at the day level[4,36], we utilized block bootstrap with days as blocks.

To compare with the estimates obtained with the other two methods, we further compute the average intervention effect across $k$ days and the corresponding standard error. The average effect is taken as the mean of $DID_{i,0}$ to $DID_{i,k}$ and the standard error is computed as follows.

$$se_{i,0-k} = \sqrt{\frac{\sum_{a=0}^{k} se_{i,a}^2 + \sum_{b=0, b \neq a}^{k} \sum_{a=0}^{k} cov_{i,a,b}}{(k+1)^2}} \quad (12)$$

where $se_{i,0-k}$ denotes the standard error of the average effect after $k$ days of intervention $i$, $se_{i,a}$ denotes the standard error of $DID_{i,a}$; and $cov_{i,a,b}$ denotes the covariance between $DID_{i,a}$ and $DID_{i,b}$. We estimate $DID_{i,k}$ for 21 days after an intervention, which is a common length of interventions in practice.

## Acknowledgments

## Author contributions

L.L. and H.W. conceptualized the study. Z.Z. collated and cleaned the data. L.L., Z.Z. and S. Z. performed the analyses. L.L. and H.W. wrote the first draft of the manuscript. L.L., H.W. and S.W. revised and finalized the manuscript.

## Competing interests

The authors declare no competing interests.

## Data Availability

The datasets analyzed during the current study are available in the Github repository, https://github.com/lunliu454/covid19_intervention_model_compare.

## Code availability

All codes for analysis and visualization presented in this manuscript is available at https://github.com/lunliu454/covid19_intervention_model_compare. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

# References

1       Sharma, M. *et al.* Understanding the effectiveness of government interventions in Europe's second wave of COVID-19. Preprint at https://www.medrxiv.org/content/10.1101/2021.03.25.21254330v1.full (2021).
2       Brauner, J. M. *et al.* Inferring the effectiveness of government interventions against COVID-19. *Science* **371**, eabd9338 (2021).
3       Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257-261 (2020).
4       Hsiang, S. *et al.* The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**, 262-267 (2020).
5       Haug, N. *et al.* Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **4**, 1303-1312 (2020).
6       Singh, S., Shaikh, M., Hauck, K. & Miraldo, M. Impacts of introducing and lifting nonpharmaceutical interventions on COVID-19 daily growth rate and compliance in the United States. *Proc. Natl. Acad. Sci.* **118**, e2021359118 (2021).
7       Zhao, M., Holtz, D. & Aral, S. Interdependent program evaluation: Geographic and social spillovers in COVID-19 closures and reopenings in the United States. *Sci. Adv.* **7**, eabe7733 (2021).
8       Fowler, J. H., Hill, S. J., Levin, R. & Obradovich, N. The effect of stay-at-home orders on COVID-19 cases and fatalities in the United States. Preprint at https://arxiv.org/abs/2004.06098 (2020).
9       Goodman-Bacon, A. & Marcus, J. Using difference-in-differences to identify causal effects of COVID-19 policies. *Surv. Res. Methods* **14**, 153-158 (2020).
10      Dave, D., Friedson, A. I., Matsuzawa, K. & Sabia, J. J. When do shelter-in-place orders fight COVID-19 best? Policy heterogeneity across states and adoption time. *Econ. Inq.* **59**, 29-52 (2021).
11      Courtemanche, C., Garuccio, J., Le, A., Pinkston, J. & Yelowitz, A. Strong social distancing measures in the United States reduced the COVID-19 growth rate: Study

evaluates the impact of social distancing measures on the growth rate of confirmed COVID-19 cases across the United States. *Health Aff. (Millwood)* **39**, 1237-1246 (2020).

12 Angrist, J. D. & Pischke, J. *Mostly Harmless Econometrics*. (Princeton University Press, Princeton and Oxford, 2008).

13 Cho, S. Quantifying the impact of nonpharmaceutical interventions during the COVID-19 outbreak: The case of Sweden. *The Econometrics Journal* **23**, 323-344 (2020).

14 Islam, N. *et al.* Physical distancing interventions and incidence of coronavirus disease 2019: Natural experiment in 149 countries. *BMJ* **370**, m2743 (2020).

15 Dehning, J. *et al.* Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* **369**, eabb9789 (2020).

16 Banholzer, N. *et al.* Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave. *PLoS One* **16**, e0252827 (2021).

17 Esra, R. T. *et al.* Evaluating the impact of non-pharmaceutical interventions for SARS-CoV-2 on a global scale. Preprint at https://doi.org/10.1101/2020.07.30.20164939 (2020).

18 Olney, A. M., Smith, J., Sen, S., Thomas, F. & Unwin, H. J. T. Estimating the effect of social distancing interventions on COVID-19 in the United States. *Am. J. Epidemiol.* **190**, 1504-1509 (2021).

19 Sharma, M. *et al.* Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nature Communications* **12**, 1-13 (2021).

20 Bo, Y. *et al.* Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020. *Int. J. Infect. Dis.* **102**, 247-253 (2021).

21 Chaudhry, R., Dranitsaris, G., Mubashir, T., Bartoszko, J. & Riazi, S. A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. *EClinicalMedicine* **25**, 100464 (2020).

22 Dreher, N. *et al.* Policy interventions, social distancing, and SARS-CoV-2 transmission in the United States: a retrospective state-level analysis. *Am. J. Med. Sci.* **361**, 575-584 (2021).

23 Koh, W. C., Naing, L. & Wong, J. Estimating the impact of physical distancing measures in containing COVID-19: an empirical analysis. *Int. J. Infect. Dis.* **100**, 42-49 (2020).

24 Leffler, C. T. *et al.* Association of country-wide coronavirus mortality with demographics, testing, lockdowns, and public wearing of masks. *Am. J. Trop. Med. Hyg.* **103**, 2400 (2020).

25 Piovani, D. *et al.* Effect of early application of social distancing interventions on COVID-19 mortality over the first pandemic wave: An analysis of longitudinal data from 37 countries. *J. Infect.* **82**, 133-142 (2021).

26 Pozo-Martin, F. *et al.* The impact of non-pharmaceutical interventions on COVID-19 epidemic growth in the 37 OECD member states. *Eur. J. Epidemiol.*, 1-12 (2021).

27 Goldstein, P., Yeyati, E. L. & Sartorio, L. Lockdown fatigue: The diminishing effects of quarantines on the spread of COVID-19. Preprint at https://doi.org/10.21203/rs.3.rs-621368/v1 (2021).

28 De Chaisemartin, C. & d'Haultfoeuille, X. Two-way fixed effects estimators with heterogeneous treatment effects. *Am. Econ. Rev.* **110**, 2964-2996 (2020).

29 Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. *Journal*

*of Econometrics* (2021).

30  Google. COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/ (2020).

31  Soltesz, K. *et al.* The effect of interventions on COVID-19. *Nature* **588**, E26-E28, doi:10.1038/s41586-020-3025-y (2020).

32  Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. COVID-19 data repository. https://github.com/CSSEGISandData/COVID-19 (2020).

33  Abadie, A., Athey, S., Imbens, G. W. & Wooldridge, J. When should you adjust standard errors for clustering? Preprint at https://www.nber.org/papers/w24003 (2017).

34  Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505-1512 (2013).

35  Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *N. Engl. J. Med.* **382**, 1199-1207 (2020).

36  Bertrand, M., Duflo, E. & Mullainathan, S. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**, 249-275 (2004).

37  Siedner, M. J. *et al.* Social distancing to slow the US COVID-19 epidemic: Longitudinal pretest–posttest comparison group study. *PLoS Med.* **17**, e1003244 (2020).

38  Wibbens, P. D., Koo, W. W.-Y. & McGahan, A. M. Which COVID policies are most effective? A Bayesian analysis of COVID-19 by jurisdiction. *PLoS One* **15**, e0244177 (2020).

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- paper3SM.docx