Fear, Partisanship and the Spread of COVID-19 in the United States

Robert Kubinec[3,*] Luiz Max Carvalho[1] Cindy Cheng[2] Joan Barceló[3]
Luca Messerschmidt[2] Derek Duba[5] Matthew Sean Cottrell[6]

August 31, 2020

Abstract

In this paper we use a Bayesian latent variable model to identify the effect of sociopolitical covariates on the historical COVID-19 infection rate among the 50 states. The model is calibrated using serology surveys issued by the Center for Disease Control. The model is able to show important associations between cellphone mobility and daily polls of concern over COVID-19 spread with the spread of the pandemic. We use mediation analysis to show how other covariates hypothesized to affect disease spread, including 2016 Trump vote share, public health spending, smoking rates, per capita income and concern over the state of the economy predict COVID-19 spread. We are able to show stark associations between higher Trump approval and less spread of COVID-19, but these effects are not mediated by mobility or fear of COVID-19. Rather, the association between partisanship for President Trump and a weaker pandemic persists despite, rather than because of, recommended social distancing measures, signifying that residents of Republican-leaning states likely inferred that they did not need to adopt strategies of left-leaning states to protect themselves from COVID-19.[1]

[1] School of Applied Mathematics, Getúlio Vargas Foundation

[2] Hochschule für Politik at the Technical University of Munich (TUM) and the TUM School of Governance, Munich, Germany

[3] Social Science Division, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[4] Department of Political Science, University of Southern California

[1]To reproduce the model and to access the underlying Stan code, please see our Github page. This paper is part of the CoronaNet project collecting data on government responses to the COVID-19 pandemic. For helpful comments we thank participants of the 2020 Polnet/Politics and Computational Social Science Conference. We acknowledge funding from New York University Abu Dhabi and the Technical University of Munich. We thank Tesea Conte, Muhannad AlRamlawi, Shiva Teerdhala, and Luke Burkholder for invaluable research assistance in evaluating state-level COVID-19 policies.

[5] School of Politics and Global Studies, Arizona State University

[6] University of California Riverside

[*] Correspondence: Robert Kubinec <rmk7@nyu.edu>

In this paper, we present a Bayesian latent variable model to track COVID-19 spread in the United States, and apply the model to measure associations between state-level sociopolitical factors and the spread of COVID-19. The utility of the model, compared to existing approaches, lies in its direct parameterization of the well-known observation bias stemming from incomplete testing for the disease in the United States. Rather than attempt to use statistical methods to eliminate the bias, we model this structure directly by jointly estimating the effect of covariates on COVID-19 cases and tests via the unobserved infection rate.

We use this model to measure important associations between U.S. state-level factors and the disease as of July 14, 2020 to better understand how political, social and economic factors correlate with the spread of the disease. We investigate the increasingly clear partisan division over the disease, and we show in this paper that there is a robust correlation between partisanship and the disease's spread, with states that voted for Trump in 2016 much less likely to see wide scale outbreaks. To explain this association, we employ mediation analysis in our models to understand how much a particular factor suppresses the disease by reducing mobility (measured through cell phone signals) or by increasing people's concern about the disease (measured through daily polling) as opposed to an unmediated direct pathway. As a result, we show in this paper that the Trump effect on the disease does not come through reducing mobility or increasing concern over the pandemic, but rather through some other unmeasured mechanism. We interpret this finding to imply that the Trump association with COVID-19 is largely driven by exogenous conditions affecting the early spread of the outbreak, and that residents of pro-Trump states likely made the erroneous conclusion that the pandemic was a problem associated with more liberal areas of the country, and would not affect them as strongly. As a result, people in pro-Trump states have tended to practice fewer social distancing behaviors and have displayed less concern over the pandemic overall.

In addition to this finding, we examine an array of other important political behaviors and COVID-19 and the disease. We show that social justice protests when measured as a proportion of the state population do show associations with increased COVID-19 spread through lowering people's fear of the disease and also via a direct pathway. This result suggests that epidemiologists' fears over outdoor protest activity were justified, although the effect sizes are relatively small compared to other factors. If a state were to experience significant protest activity every day following the COVID-19 outbreak–which did not of course occur–the cumulative infection rate would increase to a maximum of 0.27%.

Finally, we show that the effect of state-level suppression policies targeted at the epidemic varies significantly over the time, making summary statements difficult. While some policies like stay-at-home orders seem to have an ability to suppress the epidemic consistently, other policies like business restrictions show declining effectiveness over time. On the other hand, individual-level measures of personal behavior, such as polling data about individual concern over COVID-19 and polling data about mask-wearing prevalence, show the

strongest association with reduced disease spread. These results suggest that the link between state policies and individual behavior cannot be assumed but rather must be estimated to know what effect policies will actually have on the course of the pandemic.

1 Partisanship, Policies, and COVID-19

The partisan divide in American politics has become a serious concern in political science and the broader community as identities have hardened in a process ongoing since the 1990s or even earlier (Alesina and Rosenthal 1995; Poole and Rosenthal 2007, 1997; Grossman and Hopkins 2016; Iyengar and Westwood 2015). The powerful effect of partisanship on American politics has grown even stronger since the polarizing presidency of Donald J. Trump and the hardening of racial identities in the United States (Horowitz, Brown, and Cox 2019). More recently, political scientists have investigated to what extent partisanship has inhibited preventive measures against the COVID-19 pandemic as President Trump has argued against public health policies like face masks. Research has already shown that Republicans are less likely than Democrats to practice public health behaviors like hand washing (Gadarian, Goodman, and Pepinsky 2020), to practice social distancing (Andersen 2020; Alcott et al. 2020; Painter and Qiu 2020), and to comply with policies targeted against COVID-19 (Fan, Orhun, and Turjeman 2020; Grossman et al. 2020). These results seem to imply that partisanship is perhaps more than just a "hell of a drug"; it may even make the user insensitive to the risks of a deadly pandemic.

However, at the same time, research probing this relationship suggests observed behavior is more than just partisanship–or at least that partisanship has multiple dimensions. For example, Harper and Rhodes (2020) find that conservatives do dislike flouting of public health rules, though they treat violations by out-group members (i.e. liberals) more harshly than they do violations in-group members. Hart, Chinn, and Soroka (2020) argue that the politicization of news media early in the pandemic contributed to the perception of COVID-19 as a partisan issue, and that if media had featured more commentary from public health experts, the issue could have been painted differently. Similarly, Koetke, Schumann, and Porter (2020) find that trust in science is an important moderator for the partisanship-social distancing relationship: when trust in science is high, conservatives also want to support in prudent public health measures. Finally, Cornelson and Miloucheva (2020) show that the partisanship relationship may be more of an issue at the state level rather than national level, as people with a different party ID than their state governor tend to disobey their state's policies.

In summary, partisanship is an important variable for understanding the unfolding of the pandemic. It is important to note that the effect of partisanship is not limited to conservatives. A wave of protests following

the death of George Floyd at the hands of police officers is an example of progressive movements for political change leading to risky public health behavior. However, research to date suggests the protests have not had an adverse effect on COVID-19 infections (Dave et al. 2020). Nonetheless, there does seem to be at least some evidence that partisanship can affect willingness to follow public health directives on both side of the political divide.

The theory proposed in this article is that partisanship became a powerful variable not only because of President Trump's embrace of pseudo-science and the difficulties faced by state governors with hostile citizens (though these are quite influential). We argue that COVID-19 infections were unusually low in areas where Trump's approval rating was high and where more people voted for Trump in 2016, though this reprieve in the epidemic occurred despite, rather than because of, more prudent health behaviors. Instead, it would appear that the association is due to time-persistent effects related to where the outbreak initially occurred. The random chance by which coastal liberal areas experienced early outbreaks led to an enduring partisan difference in where infections arose, a difference that persists as of the date of our data (July 14th), approximately five months into the pandemic.

To establish this proposition, we will test the following two hypotheses:

> H1: COVID-19 infections are unusually low in areas with high Trump vote share and high Trump approval rating.

and

> H2: Low COVID-19 infections in areas that voted for Trump happened despite, rather than because of, pathways known to prevent COVID-19 infections via individual behaviors like social distancing.

In the next section, we discuss our statistical method for testing these hypotheses.

2 Methods

As more and more data has become available on observed case counts of the SARS-CoV2 coronavirus, there have been increasing attempts to infer how contextual factors like government policies, partisanship, and temperature affect the disease's spread (Carleton and Meng 2020; Sajadi et al. 2020; Dudel et al. 2020; Tasnim, Hossain, and Mazumder 2020; Seth Flaxman 2020; Brzezinski et al. 2020). The temptation to make inferences from the observed data, however, can result in misleading conclusions. For example, some

policy makers have publicly questioned whether the predictions of epidemiological models are far worse than the observed case count.[2] By contrast, in this paper we show that the unobserved infection rate obscures any estimates of covariates because the infection rate influences counts of both COVID-19 cases and tests. For this reason, in this paper we present a retrospective Bayesian model that can adjust for this bias by estimating the unseen infection rate up to an unidentified constant. Furthermore, by incorporating informative priors based on serological surveys of infection prevalence, it is possible to put an informative prior on the unobserved infection rate and estimate both recent disease trends and the effect of covariates on the historical spread of the disease.

In this section we present a formal definition of the model. We refer the reader to the supplementary materials for details of Monte Carlo simulations showing recovery of the latent infection rate.

Compartmental models employed by epidemiologists to study disease, and in particular SARS-CoV2 (Peak et al. 2020; Riou et al. 2020; Robert Verity 2020; Perkins et al. 2020; Jose Lourenco 2020; Ruiyun Li 2020; Neil M Ferguson 2020), suppose different classes (compartments) of individuals in the population, denoted S for susceptible, I for infectious, and R for removed (other compartments may be added, such as E for exposed). The model is usually written in the form of a system of ordinary differential equations (ODEs) and assumes a fixed population size, as seems reasonable during a relatively quick epidemic. The number infected individuals can then be obtained from the solution of the ODE system for the I compartment. These models guide our understanding of the disease and its progression, and have made warnings about the disease's spread that are proving true on a daily basis.

By contrast, this paper endeavors to estimate a much simpler quantity than the entire evolution of the outbreak. Many researchers and the general public often want to learn about what has already happened, or the *empirical* infection rate (also called the attack rate in the epidemiological literature). For a number of time points $t \in T$ since the outbreak's start and countries/regions $c \in C$, we aim to identify the following quantity:

$$f_t\left(\frac{I_{ct}}{S_{ct} + R_{ct}}\right)$$

Assuming a fixed population size, this quantity is simply the marginal rate of infections in the population up to the present. The function f_t determines the historical time trend of the rate of infection (which is assumed to be same across countries/regions) in the population up to time T, the present. Because the denominator is shifting over time due to disease progression dynamics, this model is only useful for retrospection, i.e., to examine factors that may be influencing the empirical time trend f_t. As S_{ct} and R_{ct} are exogenous to

[2]See article available at https://www.realclearpolitics.com/video/2020/03/26/dr_birx_coronavirus_data_d

the model, the model cannot predict future prevalence of the disease given that it does not determine these crucial factors. In other words, this model can be seen as a local linear approximation to the I_{ct} curve from an SIR model.

However, we do not have estimates of the actual infected rate I_{ct}, only positive COVID-19 cases a_{ct} and numbers of COVID-19 tests q_{ct}. Given this limitation, the aim of the model is to backwards infer the infection rate I_{ct} as a latent process given observed test and counts. Modeling the latent process is necessary to avoid bias in using only observed case counts as a proxy for I_{ct}. The reason for this is shown in Figure 2 in which a covariate X_{ct}, such as temperature, is hypothesized to affect the infection rate I_{ct}. Unfortunately, increasing infection rates can cause both increasing numbers of observed counts a_{ct} and tests q_{ct}. As more people are infected, more tests are likely to be done, which will increase the number of cases independently of the infection rate. As a result, due to the back-door path from the infection rate I_{ct} to case counts a_{ct} via the number of tests q_{ct}, it is impossible to infer the effect of X_{ct} on I_{ct} from the observed data alone without modeling the latent infection rate.

Figure 1: Directed Acyclic Graph Showing Confounding of Covariate X_{ct} on Observed Tests q_{ct} and Cases a_{ct} Due to Unobserved Infection Rate I_{ct}

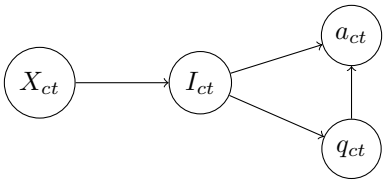


Figure shows the relationship between a covariate X_{ct} representing a policy or social factor influencing the infection rate I_{ct}. Because the infection rate I_{ct} influences both the number of reported tests q_{ct} and reported cases a_{ct}, any regression of a covariate X_{ct} on the reported data will be biased.

Given this overview of the intuition behind our approach, we turn to a more formal definition. Our observed outcomes are the cumulative total of tests and cases reported on a given day t in a given state c. We assume that the unobserved state-specific cumulative infection rate I_{ct} can be modeled as a Beta-distributed random variable. We also assume that the over-time change in the disease can be modeled as a 3-order polynomial time trend that is a function of the number of post-outbreak time periods $T_O < T$, where an outbreak begins at the first reported case in a given area. The third-order polynomial reflects the fact that epidemics occur in waves, although the curve is unlikely to be symmetric as a simpler quadratic function would require. We allow the polynomial trends to vary by states hierarchically, i.e., the information about the trends is partially pooled across states. This time function permits reasonable flexibility while still constraining the overall trajectory of the disease to a wave-like shape. We note that varying start dates for the disease allow us to incorporate state-level heterogeneity in the early spread of the pandemic.

We define the conditional distribution of the unobserved infection rate I_{ct} as:

$$\Pr(I_{ct} \mid t = T) \sim \text{Beta}(\mu\phi, (1-\mu)\phi) \tag{1}$$

$$\mu = g^{-1}(\alpha_1 + \beta_{O1} \sum_{c=1}^{C} \sum_{t=1}^{T-14} a_{ct} + \tag{2}$$

$$\beta_{I1} t_o + \beta_{I2} t_o^2 + \beta_{I3} t_o^3 + \beta_C X_{ct}) \tag{3}$$

This parameterization of the Beta distribution in terms of μ and ϕ follows from the Beta regression literature (Ferrari and Cribari-Neto 2004) so that we can model the expected value $E[I_{ct}]$ directly via μ. As such, we use $g^{-1}(\cdot)$, the inverse logit function, to scale the linear model in μ to the $(0, 1)$ interval. For the parameters, $\beta_{O1} \sum_{c=1}^{C} \sum_{t=1}^{t-14} a_{ct}$ are the sum of observed cases in the country with a 14-day lag, which represents the possibility of cross-border spread in infections. The three β_{Ii} are polynomial coefficients of the number of post-outbreak time periods t_o.

The parameter vector β_C represents the effect of independent covariate matrix X_{ct} on the latent infection rate. These are our main variables of interest, and are estimated marginal of the polynomial time trends. Finally, the parameter ϕ is a dispersion parameter governing the variability of latent infection rate.

Because we do not have measures of I_{ct}, we need to used the observed data, tests q_{ct} and cases a_{ct}, to backwards infer I_{ct}. First, we propose that the number of infections is associated with the count of tests as states try to identify who may have the disease. Second, we can assume that a rising infection rate is associated with a higher ratio of positive results (reported cases) conditional on the number of tests. We model both of these observed indicators, tests and cases, jointly to simultaneously adjust for the infection rate's influence on both factors.

To model the number of tests, we assume that each state has an unobserved level of testing parameter, β_{cq}, indicating how strongly each state is willing and able to perform tests as a factor of the unobserved infection rate. We further allow this testing parameter to vary linearly over time as testing capacity increases (or potentially decreases) within states. The cumulative number of observed tests q_{ct} for a given time point t and state c and as a fraction of the states' population, c_p, then has a binomial distribution:

$$q_{ct} \sim \text{Binomial}(c_p, g^{-1}(\alpha_2 + \beta_b I_{ct} + \beta_{cq} I_{ct} L_t + \beta_L L_t)). \tag{4}$$

The parameter β_{cq} serves to scale the infection rate I_{ct} so that an increasing infection rate has heterogeneous effects on the number of tests by state. The parameters β_b and β_L permit the baseline rate of testing to

infected to vary over time as well. The intercept α_2 indicates how many tests would be performed in a state with an infection rate of zero, and as such is likely to be very low.

Given the parameter β_{cq}, a state could test almost no one or test far more than are actually infected depending on their willingness to impose tests. Because the capacity to test changed significantly over time, we include a linear time interaction (denoted L_t) to allow testing capacity to adjust accordingly.[3]

The binomial model for the number of observed tests q_{ct} provides some information about I_{ct}, but not enough for useful estimates. We can learn much more about I_{ct} by also modeling the number of observed cases a_{ct} as another binomial random variable expressed as a proportion of the state population, c_p:

$$a_{ct} \sim \text{Binomial}(c_p, g^{-1}(\alpha_3 + \beta_a I_{ct})), \tag{5}$$

where $g^{-1}(\cdot)$ is again the inverse logit function, α_3 is an intercept that indicates how many cases would test positive with an infection rate of zero (approximately equal to the false positive rate of the test), and $\beta_a > 0$ is a parameter that determines how hard it is to find the infected people and test them as opposed to people who are not actually infected. We impose a positivity constraint on this parameter to identify the latent variable so that an increasing infection rate is always associated with a non-decreasing proportion of cases in the population. The multiplication of this parameter and the infection rate determines the cumulative number of cases, a_{ct}, as a proportion of the state population, c_p.

To summarize the model, infection rates determine how many tests a state is likely to undertake and also the number of positive tests they receive as cases. This simultaneous adjustment helps takes care of mis-interpreting the observed data by not taking into account varying testing rates, which has made it hard to generalize findings concerning the disease and also led some policy makers to claim that rising case rates are solely due to increasing numbers of tests. It also allows us to learn the likely location of the infection rate conditional on what we observe in terms of tests and cases.

Because sampling from a model with a hierarchical Beta parameter can be difficult, we can simplify the final likelihood by combining the beta distribution and the binomial counts into a beta-binomial model for tests:

$$q_{ct} \sim \text{Beta-Binomial}(c_p, \mu_q \phi_q, (1 - \mu_q)\phi_q) \tag{6}$$

$$\mu_q = g^{-1}(\alpha_2 + \beta_b I_{ct} + \beta_{cq} I_{ct} L_t + \beta_L L_t) \tag{7}$$

and cases:

[3]For a very compelling visualization of this process with empirical data from the COVID-19 pandemic, we refer the reader to this website: https://ourworldindata.org/grapher/covid-19-tests-cases-scatter-with-comparisons.

$$a_{ct} \sim \text{Beta-Binomial}(q_{ct}, \mu_a\phi_a, (1-\mu_a)\phi_a) \tag{8}$$

$$\mu_a = g^{-1}(\alpha_3 + \beta_a I_{ct}). \tag{9}$$

where I_{ct} is now equal to the linear model vector μ shown in (3) and mapped to $(0, 1)$ via the inverse logit function.

2.1 Identification

This model contains an unobserved latent process I_{ct}, and as such there are further constraints necessary in order to have a unique scale and rotation of the latent variable. Three restrictions are necessary to identify the rotation of the latent variable. First, as noted earlier, we impose a positivity constraint on the parameter $\beta_a > 0$ so that the latent variable is always increasing in rising case counts. In addition, we impose a constraint on the latent variable so that it is always increasing in t within a given state c. Because we know that the count of cases and tests is cumulative, we need to require that I_{ct} is always non-decreasing relative to itself for a given state c. We do so by adjusting I_{ct} using the ordered transformation in (10):

$$I_{ct} = \begin{cases} I_{ct} & \text{if } t = 1 \\ I_{ct-1} + e^{I_{ct}} & \text{if } 1 < t < T \end{cases} \tag{10}$$

Because I_{ct} is used as a right-hand side variable in the cases/tests beta-binomial models, we do not need to include a Jacobian adjustment to permit correct sampling given this transformation.

The third and final step in identifying the model is to add on further prior information concerning the location and scale of I_{ct}. Without further information, I_{ct} will rank the states relative to each other in terms of latent infection rates, but it will not reflect any meaningful scale in terms of percent of the state population. In other words, we need a way to relate the latent space to an empirically defined space.

To add in this crucial information, we employ serology surveys undertaken by the Centers for Disease Control. Though these surveys are opt-in samples, they were adjusted using post-stratification to match population totals, providing a reasonably accurate assessment of the state of infection for a given state (Havers and Krapiunaya 2020). The surveys we employ are listed in Table 1. In cases where only a portion of the state was sampled, we project the infection rate to the entire state by assuming that the cases/infected ratio (i.e. the observation bias) is constant within the state at that time point.

The survey information is added as a strongly informative prior on the transformed infection scale I_{ct}:

Table 1: Geographic Serological Surveys from the Centers for Disease Control

State	% Infected	N	Date Started	Date Ended
Washington	0.69%	3265	2020-03-23	2020-04-01
New York	3.71%	2482	2020-03-23	2020-04-01
Florida	0.77%	1742	2020-04-06	2020-04-10
Missouri	2.65%	1882	2020-04-20	2020-04-26
Utah	2.18%	1132	2020-04-20	2020-05-03
Connecticut	4.94%	1431	2020-04-26	2020-05-03
Pennsylvania	1.37%	824	2020-04-13	2020-04-25
Pennsylvania	2.5%	1743	2020-05-26	2020-05-30
Minnesota	1.93%	860	2020-04-30	2020-05-12
Louisiana	4.48%	1184	2020-04-01	2020-04-08
Connecticut	5.02%	1800	2020-05-21	2020-05-26
Missouri	2.72%	1831	2020-05-25	2020-05-30
California	0.93%	1224	2020-04-23	2020-04-27
Washington	2.14%	1719	2020-04-27	2020-05-11
Florida	1.29%	1280	2020-04-20	2020-04-24

$$I_{ct} \sim \text{Normal}(g^{-1}(s_{ct}, .01)) \tag{11}$$

where s_{ct} is the value of seroprevalence at state s and time t and $g(\cdot)$ is the logit function. The use of the logit function allows us to assign this prior before transforming I_{ct} to the $(0, 1)$ scale and is done for computational convenience.

Because the serology surveys are relatively early in the time series, we add a semi-informative prior on the infected to reported case-population ratio for all days following the survey. This prior provides general bounds on the reporting bias revealed by the serology surveys:

$$\frac{I_{ct}}{\frac{a_{ct}}{c_p}} \sim \text{logNormal}(2.1, .4) \tag{12}$$

This time-invariant prior suggests that the ratio of infected individuals to reported cases is somewhere between 2 and 20 with a median of value of 8. This prior puts density on all of the infected-case ratios from

serology surveys in Table 1

As we show in the appendix with simulations, no other identification restrictions are necessary to estimate the model beyond weakly informative priors assigned to parameters.

These are:

$$\beta_a \sim \text{Normal}(30, 10), \tag{13}$$

$$\beta_{qc} \sim \text{Normal}(\mu_q, \sigma_q), \tag{14}$$

$$\sigma_q \sim \text{Normal}(0, 5), \tag{15}$$

$$\mu_q \sim \text{Normal}(0, 20), \tag{16}$$

$$\beta_C \sim \text{Normal}(0, 5), \tag{17}$$

$$\beta_{Ii} \sim \text{Normal}(\mu_{Ii}, \sigma_{Ii}), \tag{18}$$

$$\mu_{Ii} \sim \text{Normal}(0, 10), \tag{19}$$

$$\sigma_{Ii} \sim \text{Normal}(0, 5), \tag{20}$$

$$\alpha_1 \sim \text{Normal}(0, 10), \tag{21}$$

$$\alpha_2 \sim \text{Normal}(0, 10), \tag{22}$$

$$\alpha_3 \sim \text{Normal}(0, 10) \tag{23}$$

where the normal distribution is parameterized in terms of mean and standard deviation.

The priors to note are the hierarchical regularizing prior put on the varying testing adjustment parameters β_{qc} and varying polynomial trends β_{Ii} with shared means and standard deviations. This partial pooling permits a reasonable degree of heterogeneity in the parameters while still constraining overall dispersion.

We note that an advantage of this framework is providing a way to measured the count of infected adjusting for known biases in the number of tests. By comparing numbers of tests per capita and growth rates in cases across regions, the model is able to backwards infer a likely number of infected individuals in a given area. As such it exploits both within-area and between-area variance to adjust for the biases of imperfect testing. The wide variety of covariates we add to the model, which we describe in the next section, provide the mechanism through which the model can infer test/case relationships even in states which have not had a CDC serology survey.

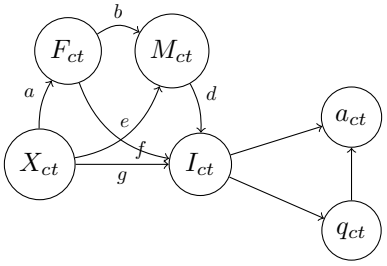
We also extend this model in order to analyze the mediation of a subset of covariates X'_{ct} by adding mediators M_{ct} for mobility and F_{ct} for fear of the disease to the causal diagram, as in Figure 2.1. Figure 2.1 has several paths due to the fact that the influence of covariates X_{ct} affects the two mediators differently. Given that beliefs and preferences precede actions, the covariates X'_{ct} first influence I_{ct} along the ae and abd path through

perceptions of how dangerous the disease is. These beliefs both affect the chance of an individual getting infected and thus I_{ct} directly on the path ae, such as by causing an individual to adopt social distancing behaviors, and also on an indirect path abd by which an increase in a people's fear of the disease reduces mobility as people prefer to stay home.

In addition to pathways through the fear mediator F_{ct}, a covariate could influence infections along the pathway through mobility ed without increasing or decreasing fear. This situation could arise if government policies forced people to stay at home against their will and despite their unconcern about the disease. Finally, a covariate could have an unmediated direct effect g on the infection rate. The total effect of a covariate X_{ct} on the spread of the disease is then the sum of all the paths, $abd + af + ed + g$. To calculate the indirect effects and direct effects given the use of the inverse logit function $g^{-1}(\cdot)$, I employ the chain rule as in Winship and Mare (1983) to calculate the marginal effect of covariates with respect to different pathways to I_{ct}.

Adding the mediators to the model is relatively simple as they do not have link functions and can be included as Normal distributions (i.e., OLS regression) as in Yuan and MacKinnon (2009). It should be noted that there are in fact five mobility covariates as explained in the following section, and so we explicitly model the covariance in mobility via a multivariate Normal distribution with a covariance matrix parameter Σ_m.

Figure 2: Directed Acyclic Graph for Latent Infection Rate with Mediators



This figure adds mediators M_{ct} (mobility data) and F_{ct} (fear of COVID-19) that mediate the relationship between state-level covariates X'_{ct} and the latent infection rate I_{ct}. Because beliefs precede actions, F_{ct} is causally prior to M_{ct} and can affect infections both via reducing mobility (path abd) and directly apart from mobility (path ae), such as by encouraging individuals to remain socially distant.

To add our mediation covariates M_{ct} and F_{ct}, which we describe in more detail in the next section, we multiply the following likelihoods with the joint posterior:

$$M_{ct} \sim MVN(\alpha_m + \beta_m X_{ct}, \Sigma_m) \tag{24}$$

$$F_{ct} \sim N(\alpha_f + \beta_f X_{ct}, \sigma_f) \tag{25}$$

We also include all of M_{ct} and F_{ct} as linear predictors in (3).

We fit this model using Markov Chain Monte Carlo sampling in the Stan software package (Carpenter et al. 2017). We fit the model for 1000 iterations with 500 warmup iterations and two chains to test for convergence.

3 Data

The only data required to fit the model, in addition to the covariates of interest, are observed cases and tests for COVID-19 by day. In this section, we fit the model to numbers of COVID-19 case counts on US states and territories provided by The New York Times. By doing so, we can use the differences in trajectories across states to help identify the effect of state-level covariates on the infection rate. We supplement these observed case counts with testing data by day from the COVID-19 Tracking Project. As there are discrepancies where the reported number of cases or tests decreases for a given day, we impute these cases and tests through linear interpolation as the number is likely an under/over count of neighboring days. We then take the 7-day rolling average of both series to account for reporting fluctuations and weekly reporting effects.

To analyze the effect of suppression policies, we use data on counts of social distancing policies, restrictions on mass gatherings, restrictions on businesses, mandatory mask orders, restrictions on government services, and stay-at-home orders from the CoronaNet Government Response dataset (Cheng et al. 2020). For each type of policy, we include a variable representing the count of policies in that category effective for a particular day. For each update to an existing policy, we code it as +1 if the update increases the scope of the policy or -1 if it decreases the scope of the policy (down to a minimum of 0). While this is a simplification of the underlying data, we are still able to capture relative complexity over time without having to make judgments about stringency or other qualitative criteria. We then interact these policy counts with a linear trend to examine time-varying policy effects. We separately include policies designed to increase health resources like personal protective equipment (PPE) and also policies requiring mask use as we do not examine time-varying effects of these covariates. The use of a variety of policy types is important as the adoption of policies is correlated and so including only stay-at-home orders could mask other distinct policies that were implemented at the same time.

The policy data is plotted by state in Figure 3. As can be seen, there is a rise in policies after the pandemic begins in the middle of March, though the number of policies varies across categories. The count of policies is an admittedly imperfect measure though it communicates more information about policy activity than a simple binary coding. Generally speaking, states imposed many more policies designed to increase their access to PPE for health staff than they were willing to take on lockdowns, social distancing, and restrictions

on businesses and government services. This difference likely has to do with the increased cost and salience of these policies vis-a-vis relatively less politically difficult options like gathering more masks and face shields for health care workers (Cheng et al. 2020).

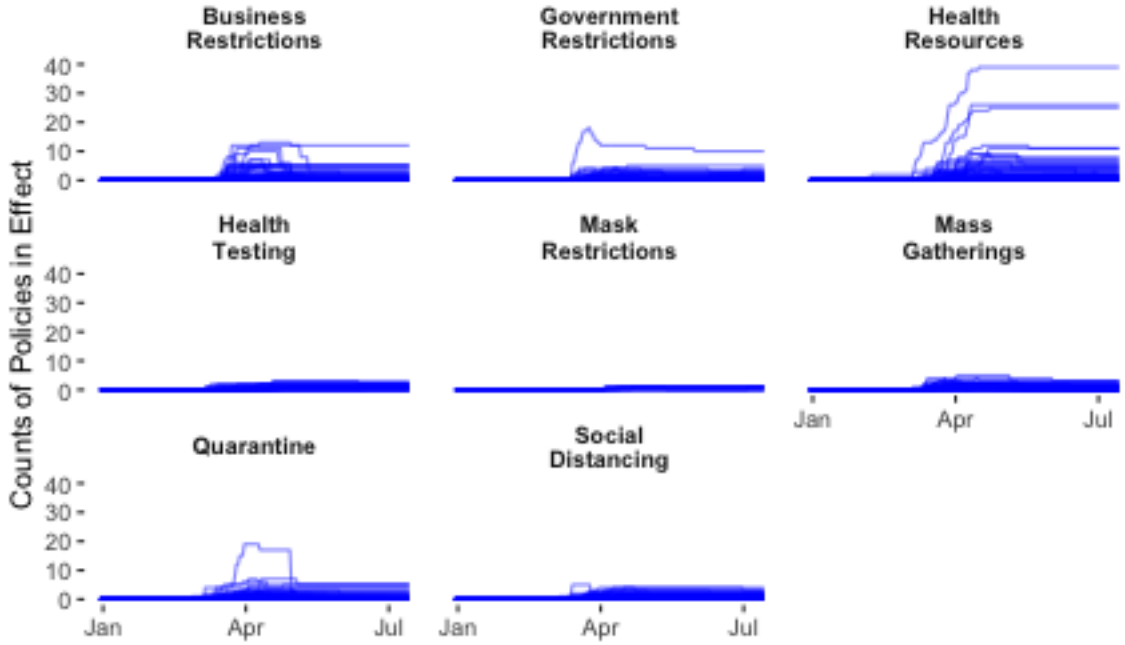


Figure 3: Count of Policies in Effect by Day and by State from the CoronaNet Dataset

To better understand over-time factors that may also affect COVID-19, we include polling data from Civiqs and YouGov at the state level. From Civiqs we include state-level polling averages by day for the percentage of respondents favoring Trump, percentage reporting the economy is "very good", and the percentage reporting that they are "extremely concerned" about the coronavirus. From YouGov we use a poll from May 8th reporting average number of respondents who said they used masks by U.S. state. As this poll does not vary over time, we set the mask prevalence at one-half the minimum value of the poll prior to the WHO's revision of guidance concerning wearing masks on April 3rd, and equal to the poll's values thereafter. As described in the previous section, the poll asking respondents whether they are "extremely concerned" about COVID-19 represents our fear mediator, and is also included as a separate outcome with other covariates as predictors.

To better understand the mediating effects of suppression policies, we include Google mobility data[4] for retail, residential, parks, workplaces, transit and retail establishments. These estimates are by day and aggregated to the state level. They are measured in terms of an index that is initialized with a value of 100 at the index start on February 15th, 2020. To test for mediation, we include these as predictors of

[4]See https://www.google.com/covid19/mobility/

the infection rate, and separately fit a likelihood with each mobility covariate as an outcome and the other covariates as predictors.

We note that it is important to measure mediation for mobility because mobility is hypothesized to affect the spread of COVID-19 (Gao et al. 2020). As such, measuring the simultaneous effect on mobility for covariates in our model is important as the covariates could be affecting mobility, which subsequently affects COVID-19 spread. Ignoring this association would result in post-treatment bias that deflates the effect of predictors in the model, though our main interest in including these variables is because this mediation is substantively interesting to decompose.

To measure protest activity, we include a covariate reflecting the proportion of a state's population engaged in social justice protests following the death of George Floyd on May 25, 2020. This data is drawn from publicly available information about the number and size of protests from three online sources: Wikipedia protest data, the Count Love protest web-crawling web site,[5] and list of protests compiled by Ipsos.[6] For protests present in only one of the three sources, we used information on both size and location. If a protest was present in three sources, we averaged reported protest size. If the sources had contradictory information about the type of protest, we had research assistants re-code the protest using secondary sources. For protests for which size was not available, we imputed missing data using random forest algorithms (Stekhoven and Bühlmann 2012).

All time-varying covariates–polling, protests, policies and mobility data–are lagged by 14 days to account for the likely delay in events showing up in reported cases. This 14-day lag comes from the epidemiology literature (Seth Flaxman 2020) and is meant to take into the account the amount of time required for people to be infected, be tested and then have the test results reflected in case counts.

We further add in non-varying state-level data on Donald Trump's vote share for the 2016 election from the MIT Election Lab, a 2019 estimate of state GDP from the Bureau of Economic Analysis, the 2018 percentage of foreign born residents, population under 18 years of age and population density from the U.S. Census Bureau, 2019 state-level average data on air pollution,[7] cardiovascular deaths per capita, percentage of residents under age 18, number of dedicated health care providers, public health funding, and smoking rates provided by the United Health Foundation ("America's Health Rankings 2019 Report" 2019). All variables are standardized to permit comparability.

We note before turning to the results that we cannot make claims of causal identification as we can with our claims of statistical identification of the latent infection rate. COVID-19 is not a very likely candidate for

[5] https://countlove.org/

[6] See https://www.ipsos.com/en-us/knowledge/society/Protests-in-the-wake-of-George-Floyd-killing-touch-all-50-states

[7] Defined as average exposure of the general public to particulate matter of 2.5 microns or less ($PM_{2.5}$) measured in micrograms per cubic meter (3-year estimate).

meeting any kind of assumption about ignorable selection into treatment; it is a disease that is indirectly caused by human behavior. Our identification strategy primarily relies on including as many relevant adjustment variables as is prudent to isolate factors which are likely to or known to have an effect on COVID-19 spread and could be confounding variables.

That being said, it is of course impossible to know for sure whether an association reported in this paper represents the effect of that variable or some other confounding factor. We make limited claims to causal identification in two cases. First, the time-varying variables included in the model with a 14-day lag are less likely to be confounded as they represent precisely measured day-to-day changes, and we can also rule out reverse causality. Second, we can separate out the possible channels of effect between covariates affecting the outcome directly and an effect mediated through mobility data and through changes in beliefs about the threat of the pandemic. While this does not allow us to state confidently whether the direct effect is identified, it does allow us to know whether a variable seems to to be linked through the outcome via social distancing behaviors or through some other means. While this may seem like a modest point, it will in fact help us to determine whether our hypotheses are supported as well as learn substantive information about how variables seem to affect the spread of COVID-19.

Finally, we would argue that covariate adjustment is and is likely to remain the best strategy for making causal claims from aggregated measurement with COVID-19. Intentionally manipulating the spread of the disease is ethically monstrous. Quasi-experimental methods are unlikely to work as they either suppose that time-varying confounders do not change (difference-in-difference) or that forcing variables might cause some to suffer more exposure than others (regression discontinuity). In the first case, the pandemic and the factors associated with it change on a daily basis (hence our use of daily data), which renders difference-in-difference estimates suspect as they assume that units follow parallel paths–at least, without extensive use of covariate adjustment. In the second case, if a forcing variable did cause some people to have less exposure, such as a geographical area, given the severity of the disease, it is very likely that people would self-select out of the higher exposed area. This has already been seen to occur as people migrate around the United States in response to rising infection counts.[8] For studying COVID-19, there simply does not seem to be any statistical equivalent of a free lunch.

For these reasons, we present these findings as observational associations with appropriate covariate adjustment so that we can at least say which associations are not related to well-known potential confounders. That is, while we cannot always say whether an effect is identified, we can at least plausibly rule out some other explanations. In general, we believe that the best strategy for understanding the spread of the epidemic

[8]For example, see https://www.wsj.com/articles/people-were-leaving-new-york-city-before-the-coronavirus-now-what-11587916800.

is to obtain the best data available and the clearest interpretations possible from models. There are issues which may never be resolved in terms of COVID-19 spread due to the difficulty of causal identification in a rapidly changing environment. However, we do think that we can learn from observational data so long as we remain aware of the ever-present possibility of alternative explanations.

4 Results

We first examine model performance to see how well the model is able to reproduce the observed data. While this kind of predictive validity is only partially useful given that we are interested in a latent quantity, it is still helpful to know whether the model is able to reproduce the empirical distribution. If the model could not fit the observed data very well, then we might be suspicious about whether we are informing our latent estimate correctly. To do we estimate the *posterior predictive distribution*, i.e., we draw from the posterior distribution using the empirical data.

The results of drawing from posterior values for the beta-binomial distribution of cases and tests are shown relative to the original observed values in Figure 4 for five states. The plots show that although there is noise in the predictions (represented by the black shaded region), the model is generally able to capture the empirical values (represented by a red line) with high probability and reasonable fit. It is important to note as well that the predictions are not always as accurate near the end of the sample time series. This is due to the use of linear time adjustments that will fit the full series but not necessarily the tails. Adding a tighter fit to the observed data would be more useful for predictive validity but at the cost of increased bias due to over-fitting (i.e., the well-known bias/variance trade-off). As Figure 4 shows, the model likely under-fits the data, which we take as a reasonable trade-off because the aim of this model is identification of the latent infection rate, not forecasting future cases and tests.

Given this check on the model's fidelity to the data, we can then report the model's estimates of infected counts for the U.S. population as a whole in Figure 5. Panel A in this plot shows the cumulative total both for reported cases (thin black line) and for the model's estimate of total infected (blue line). The interval in this plot, as with all figures presented, are the 5% and 95% quantiles of the empirical posterior distribution. As can be seen, the model estimates that there are approximately 3-4 times as many infected people in the United States as reported cases, with the total cumulative number of infected persons reaching 10 million. Early expert estimates are shown as confidence intervals in panel A, revealing that even epidemiologists largely under-estimated the spread of the disease in its early stages, largely due to limitations in testing and case reporting.

We compare these estimates with a popular COVID-19 forecaster employing SEIR models from Gu (2020)

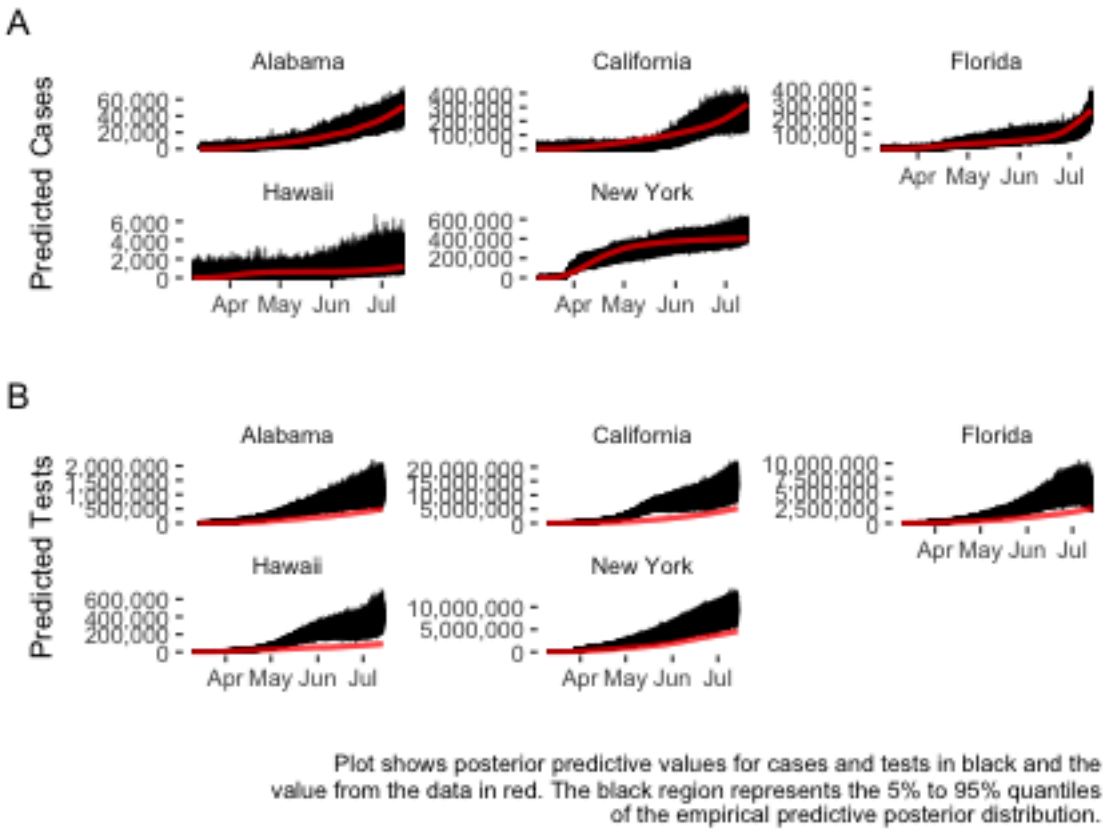


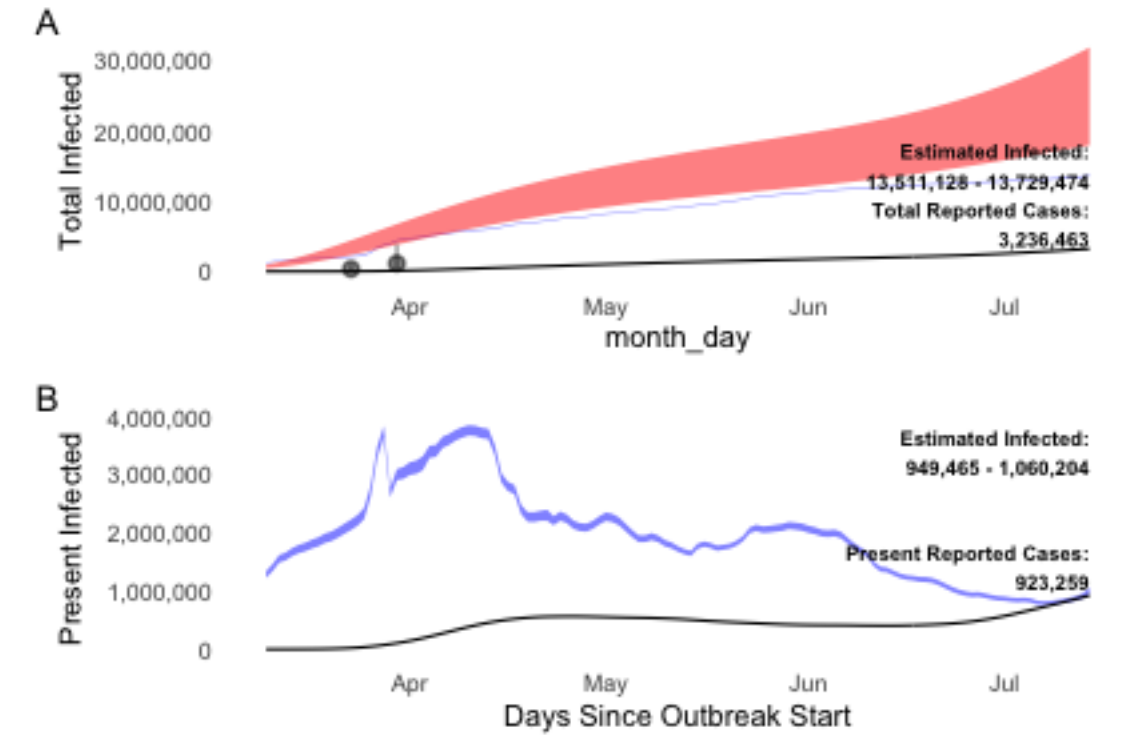
Figure 4: Predictive Validity of Model Vis-a-vis Observed Cases and Tests

by plotting their estimates as a red ribbon on the plot. As can be seen, the trajectories are similar although they diverge slightly at the end of the series in mid-July. This is likely due to the fact that our model is not designed to predict into the future, and so it does not pick up as strongly on the increase in cases in the last week of July. However, on the whole it would seem that our estimate of infected individuals is on the conservative end compared to other approaches–in other words, while we do not know for certain what the true number is, we are unlikely to be under-estimating the total. Again, as our aim here is interpretation and inference rather than forecasting, we believe the conservative nature of our model is a benefit. The estimates of covariate associations we present later in this paper could be slightly too low, but they are unlikely to be too large given what we can model. Furthermore, our intervals are far more precise than the SEIR model, which is likely because we are only trying to identify a single parameter as opposed to the full range of outcomes, including deaths and hospitalizations.

Panel B in the plot shows our estimates of infected individuals, excepts that it adjusts the cumulative number with a 19-day lag to account for the approximate time that recovery from COVID-19 requires (deaths are first subtracted). This plot displays an imperfect but useful formulation of the likely number of people infected at any given time point. As of July 14, it would appear that there were approximately 1 million infected individuals in the United States, while the number peaked at about four million in late April. It should be noted that the data used to fit the model did not include the latest wave of infections, and because the model is not predictive, it does not weight as strongly the last few observations in the series. Given these limitations, the estimate shows a declining case/infection ratio as testing increased, with the bias reaching a very low number at the end of the series. We expect given recent data that the bias would increase as case totals surged and testing lagged.

By comparison, Figure 6 shows the cumulative totals of estimated infections by state. Plot A in this figure has the count of infections by state, while plot B shows the percentage of the population infected by state. Both the overall S-shape of the epidemic can be seen along with the substantial heterogeneity in infections, with early infected states like New York and New Jersey still in the top quartile of states with infections even though they successfully reduced the rate of disease spread.

In addition to the estimation of the cumulative count of infected individuals, the model provides further useful information by parameterizing the relationship between the unobserved infection rate and the number of tests conducted in a given state. These individual parameters are shown in Figure 7. The scale of the y axis shows the number of people that a state was able to test relative to each person infected. The plot shows that some states have been able to test far more people than have been infected (New York, Rhode Island), while other states like Texas, Arizona and Pennsylvania have tested barely twice as many as those who have been infected. The fact that new outbreaks have been seen in Texas and Arizona suggests that



Blue 5% - 95% HPD intervals show estimated infected and the black line shows observed cases from he New York Times. These estimates are based on CDC seroprevalence data and a Bayesian model of how cases and tests are influenced by infection rates. Black dots in Panel A show early expert estimates of COVID-19 prevalence in the United States. Red ribbon shows 5% - 95% predicted cumulative infections from covid19-projections.com hybrid SEIR model.

Figure 5: Total Cumulative and Present COVID-19 Infections in the United States

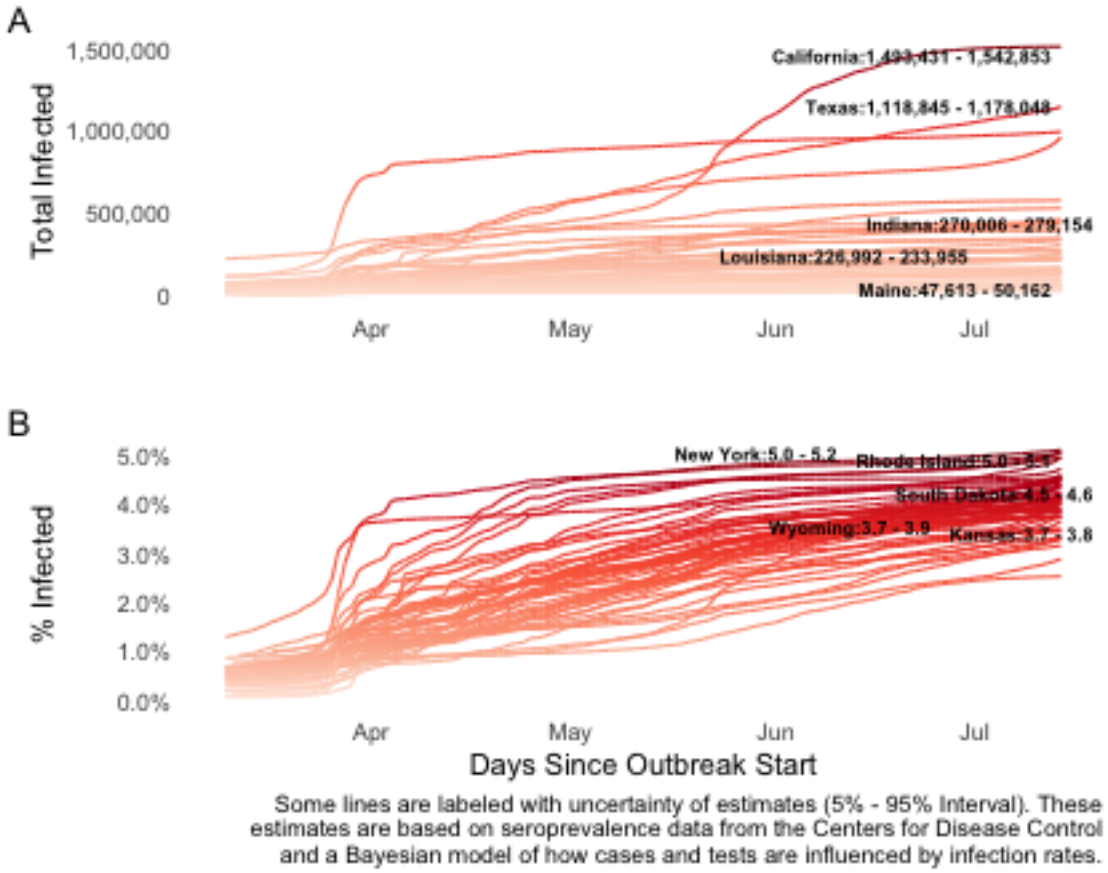


Figure 6: Average Cumulative Count of Infected People by U.S. State as of July 14th

this shortfall in testing likely disguised early outbreaks that could have been detected otherwise.

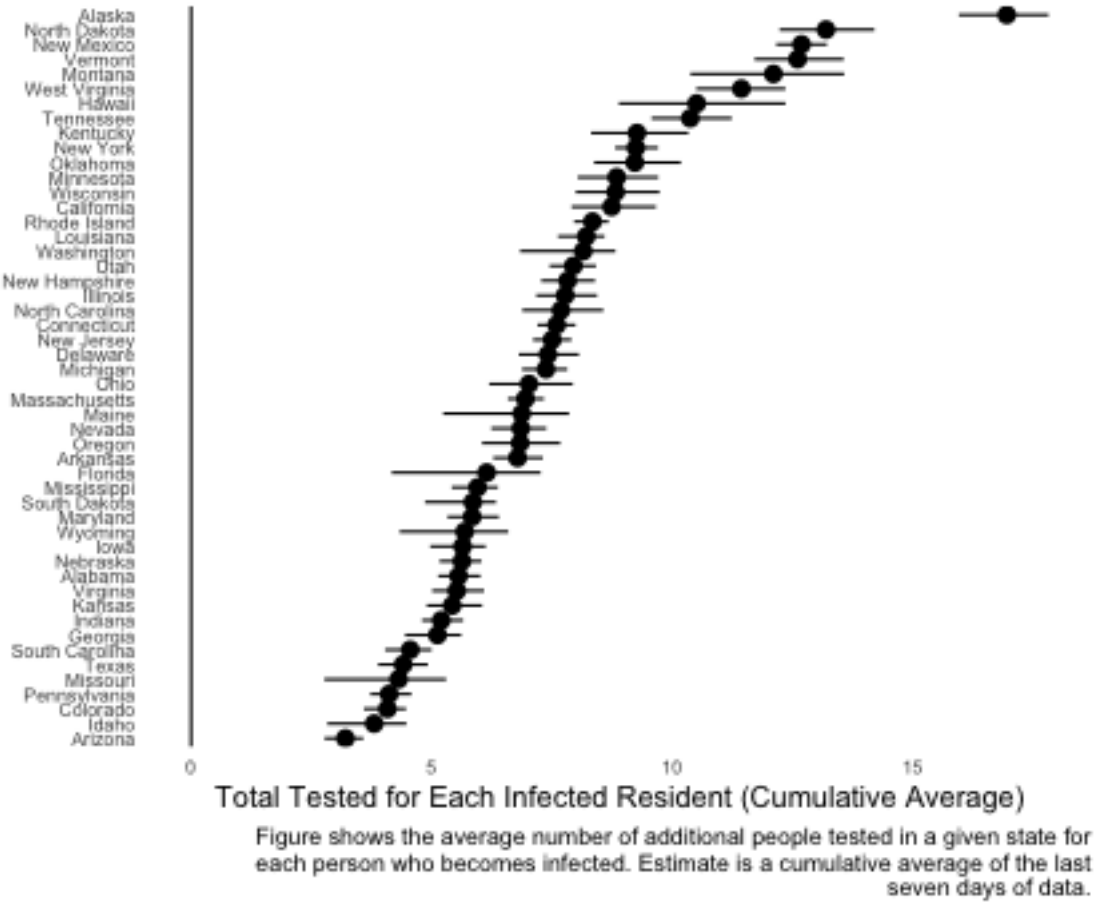


Figure 7: Measuring States' Testing Rates Relative to Infection Rates

We would note that this information is also helpful to policy makers and others trying to make sense of observed case counts given the limitation in testing thus far. Our estimates help take into account these known biases and adjust them based on differences between states and within states in terms of disease trajectories. We believe this model can be used to help understand disease trends and factors associated with it even in the relatively data-poor environment many countries find themselves in.

To calculate the effect of covariates on the infection rate, we report here average cumulative marginal effects by state. We report cumulative marginal effects rather than the sample average marginal effect because the outcome monotonically increases, and so the marginal effect at any one point in time is not as meaningful a statistic. The way to interpret the coefficients presented is how a 1-unit (and mainly 1-SD) change would affect the infection rate if that increase were sustained for an average state's entire time series (March to July).

We first show the association of mobility types with the infection rate. In Figure 8 we show the marginal

effect of a 1-SD increase in different types of Google mobility on the infection rate expressed as a fraction of a state's population. In line with the growing research on cellphone mobility and the epidemic, there are strong positive effects of some types of mobility on the spread of the disease, especially grocery stores, residential mobility and to a lesser extent workplace and retail mobility. Movement in parks and via transit has an estimated zero association with infections. While these results are somewhat surprising given how large the grocery store effect is relative to the other series, other results confirm with prior suspicions that outdoor activities like attending parks are relatively low-risk for COVID exposure. It should be noted as well that all the effects are estimated simultaneously and the variables are correlated, so the effect of workplace mobility could be partially masked by the other mobility measures. In any case, it is clear that the mobility measures have the single largest estimated effect on suppressing the disease for a time-varying covariate, reaching up to a 0.3% cumulative increase in population infected for a 1-SD increase in grocery store mobility.

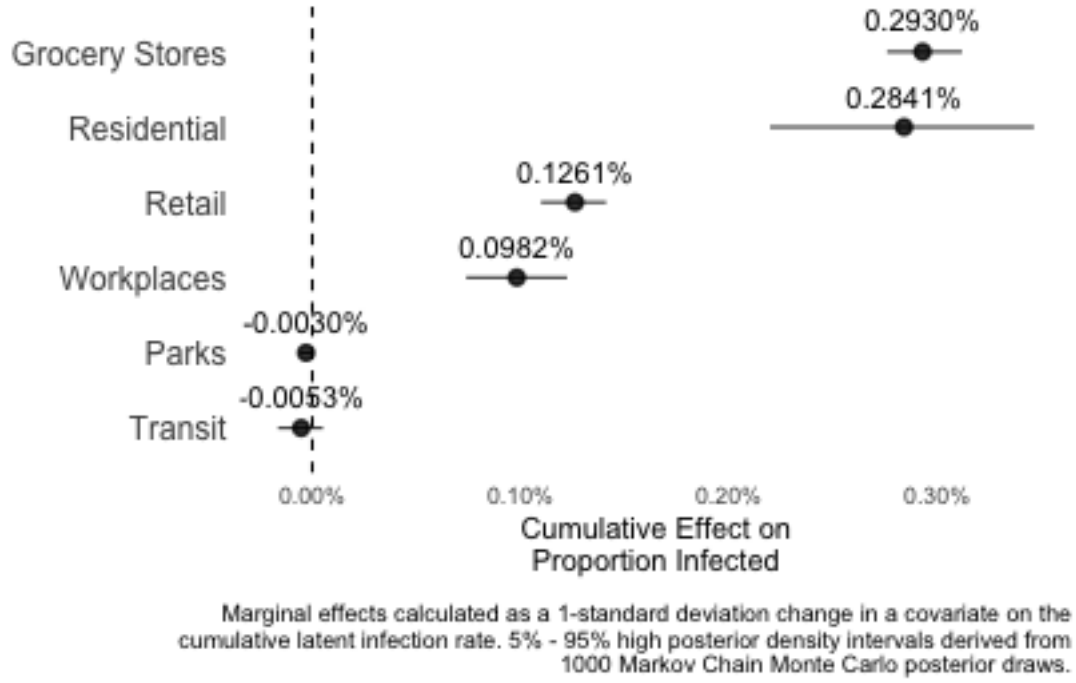


Figure 8: Effect of Google Mobility Data on COVID-19 Spread

Figure 9 shows the marginal effect of all other covariates in the model on the latent infection rate expressed as average cumulative marginal effects. The estimates are further broken out in terms of mediation. The mobility effect is equivalent to the ed path in Figure 2.1, i.e., it is the path from the covariates to mobility that does not go through increased fear. The fear effect, on the other hand, is equivalent to the $abd + ae$ paths, or the sum of the path from fear through mobility and the path from fear to infections apart from mobility. The direct effects are equivalent to the g path in Figure 2.1, and the total effects are the sum of

all paths. The direct and indirect effects are in panel A while the total effects are in panel B.

The use of mediation analysis shows substantial heterogeneity in the types of associations and whether direct and indirect effects tend to complement or substitute each other. First, it is important to note that the single strongest associations in panel B come from the YouGov mask-wearing poll, the Civiqs concern over coronavirus poll and the economy poll, and the percent of a state's residents that are foreign-born. As these are cumulative average marginal effects, that number reflects what an *average* state might experience; the effect could well be larger for states with higher infection rates than average. On the other hand, as these effects are cumulative, they reflect a state that experienced a sustained increase in the covariates and so it might overstate the effects somewhat.

In general, the factors that are most strongly associated with infection suppression are concern over COVID-19, wearing masks, and living in a state with more smokers, more wealth and more population density. Conversely, a higher percentage of foreign-born, more concern over the economy, more PPE policies and social justice protests are associated with more infections.

Given that these are associations, we can learn more about the meaning of the results when we can identify effects through pathways which we have a theoretical reason to believe matter for fighting the epidemic: individual concern over COVID-19 and individual mobility. Furthermore, there is more reason to think that an association might be identified if it is a time-varying factor given that the 14-day lag can rule out reverse causality. For this reason, while the association with percent foreign-born is quite strong, we cannot say as much about the reasons underpinning the association as it is mainly a direct effect. Indirectly, the percentage foreign born is associated with *reduced* COVID-19 spread via the fear pathway.

Considering that the share of foreign-born was fixed before the outbreak, we have limited ability to know why the direct association is so strong. However, we have reason to think that the share of foreign-born could be a proxy for international travel and hence may be estimating the long influence of the outbreak of the epidemic in states with easy access to international travel. In fact, this connection was the original motivation for including this variable as a control covariate.

In contrast to other research, we find that social justice protests are positively associated with COVID-19 spread, though the effect is not particularly large. Furthermore, as we report cumulative marginal effects, it is unlikely that states experienced protests every day in the sample, suggesting that the reported effect is more of an upper bound for what most states experiencd. There is some limited evidence, as Dave et al. (2020) suggest, that the positive effect of the protests was offset by reduced mobility by non-protesters as the indirect effect is weakly negative, though it is small (-0.017%). There is also a positive association via the fear pathway, suggesting that protest activity tended to reduce people's fear of COVID-19, and this may have led to increased infections independently of social contact.

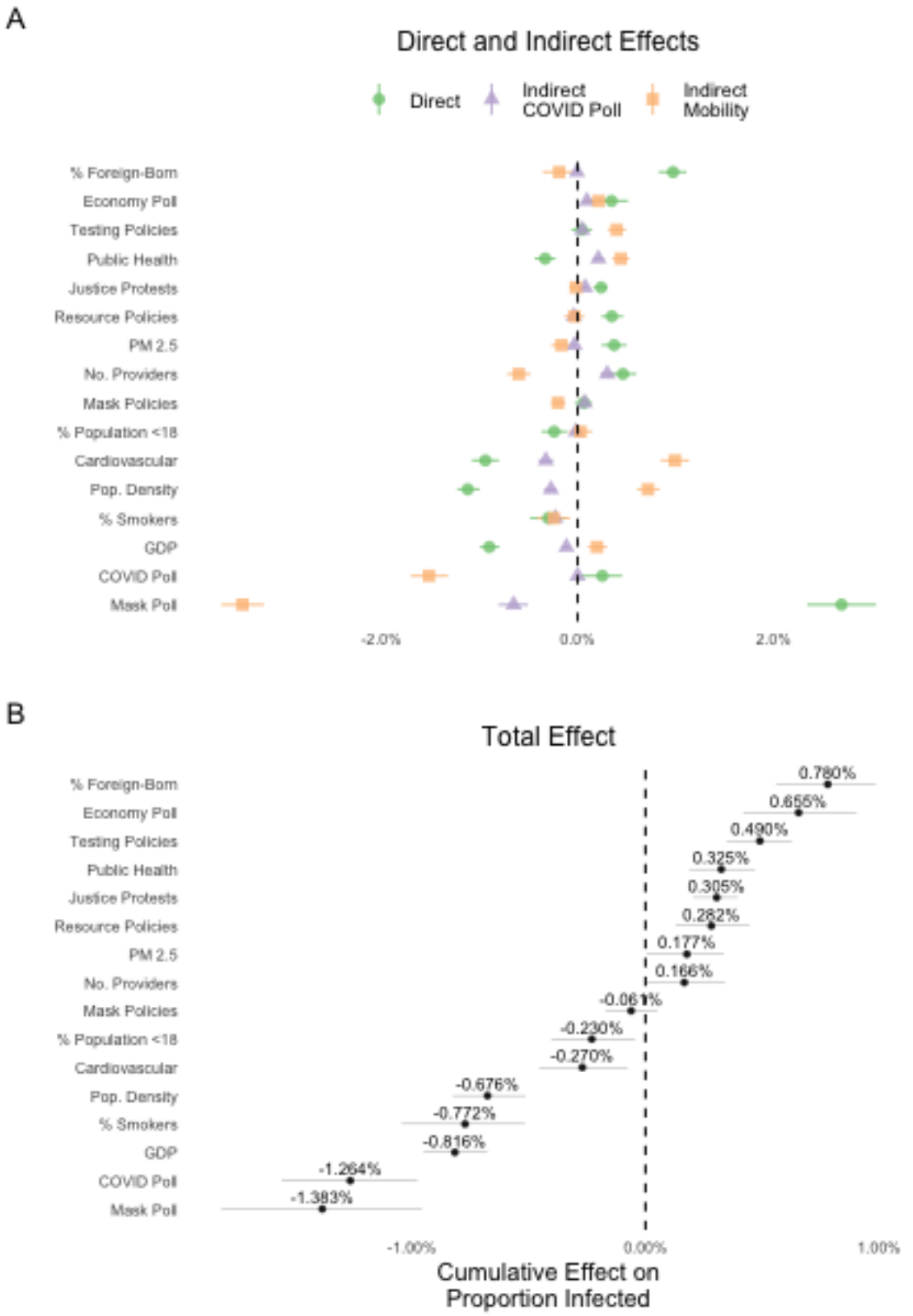


Figure 9: Marginal Effects of Covariates on Latent Infection Rates for U.S. States

While it is difficult to generalize given the variety of associations, it does seem possible to rule out some possibilities. The number of young people does not (yet) appear to be a positive predictor of the disease despite accusations that younger people tend to ignore social distancing and other policies. Health-related variables including the prior level of public health funding and the number of providers are either positively or weakly associated with the disease. They also tend to have countervailing influences via different pathways and direct effects, suggesting rather complicated processes through which they are associated with the spread of the pandemic. At the very least, there is no simple path from increased public health spending or number of healthcare providers and a more limited outbreak of COVID-19.

What is clear is that the strongest time-varying factors present in the model concern individual behavior more than policies or state preparedness. Considering that the percentage of foreign residents and the number of smokers were determined long before COVID-19 arrived, the most important manipulable factors are those involving beliefs, such as in the strength of the economy and the relative threat of COVID-19, along with personal behaviors like mask-wearing.

It is interesting finally to note differences in indirect and direct effects of the covariates in panel A of Figure 9. The large effect from the COVID poll primarily comes from mobility data; people who are more concerned about COVID are less likely to frequent places where they could contract the disease. The mask poll is associated with repressing COVID through mobility *and* fear of COVID-19, while the direct effect is positive. This may suggest that the direct association has to do with overall higher rates of COVID-19 in places where people adopted masks. These associations suggest that masks do influence how we think about the disease, though the increased awareness of the disease leads to more prudent social behavior overall rather than masks substituting for social distancing as some supposed would happen (Abaluck et al. 2020).

Given the prominence of Trump-related variables in explaining the spread of the disease, in addition to the importance of the question to the study of partisanship, we explore the interaction between Trump vote share and Trump approval polls in Figure 10. In this figure, the effect of Trump 2016 vote share is plotted conditional on the relative level of daily Trump approval polling on the x axis. The effects are shown aggregated in panel A and disaggregated across mobility types in panel B. Panel A shows that in general, the effect of partisanship for Trump is largely a direct effect, and is highly conditional on the above/below polling average of approval for Trump in a given state (which has a maximum swing of about +/- 4 pp). States that voted for Trump in 2016 tend to see more infections when Trump approval is low, and fewer infections when Trump approval is high. These effects are quite large, reaching cumulative numbers of +/- 5% infections.

However, it is important to note opposite effects through the mediated pathways. Figure 10 shows that Trump vote share mediated through mobility and fear is *positive*. While the effects are not as large as the

direct effects, they are still substantial. Trump vote share's effect on COVID-19 mediated through these important channels shows that pro-Trump states tend to implement social-distancing behaviors at lower rates, as previous research has shown, with consequent relative increases in infections. Furthermore, these associations are relatively constant given Trump approval polls, although there is a more stronger association for Trump approval polling and Trump vote share in dampening fears over COVID-19.

What we can say is that this finding points to very strong associations between partisanship and the spread of the COVID-19, but not via behaviors that we know to suppress the epidemic. States with higher Trump vote shares have seen significantly fewer infections; when Trump approval increases in these states they generally observe fewer infections, which cannot be explained through decreased mobility nor concern over COVID-19. Our contention, as expressed through our hypotheses, is that the evidence suggests this relationship is spurious. After all, it is well-known that the early states that were infected with COVID tended to vote against Trump, although partisanship is not why they were more vulnerable to COVID initially. We believe that pro-Trump states received fortuitous outcomes by happening to not be on major travel routes from early COVID-19 hot spots; rising Trump approval in these states occurred as pro-Trump residents believed their president's dismissal of the virus' threat. In other words, the unexplained direct effect justified the relative inattention to important behaviors that could prevent infection. Given the increase in COVID-19 infections in the last two months in heavily Republican states, it would seem that this tendency would lead pro-Trump states to suffer in the long run as behavior caught up with initial conditions.

Finally, we can also use estimates of cell phone mobility on COVID-19 to understand how policies have had mediated effects on the disease through increasing or decreasing mobility. Figure 11 shows the disaggregated mediation effects for two types of policies for the sake of space, restrictions on businesses and stay-at-home orders. The plots reveal how indirect mediation effects differ substantially between policy types. Panel A shows that business restrictions had a powerful suppressive effect on grocery store and to a lesser extent retail establishments during the early part of the epidemic, though that association changed over time. By contrast, panel B indicates that stay-at-home orders have had more durable effects on mobility that have suppressed the disease, particularly in grocery stores, retail establishments and transit. Furthermore, these effects seem to be increasing rather than decreasing over time. On the other hand, stay-at-home orders seem to be increasing disease infections via increasing residential mobility and decreasing time spent in parks, trade-offs that were noted in some early epidemiological modeling of COVID-19 (Seth Flaxman 2020).

The fact that we can find indirect effects of state policies, but not from Trump vote share, provides further evidence that the "Trump effect" is likely due to fortuitous circumstances. Indeed, the recent increase in COVID-19 among southern states is likely a reflection of the end of this long-standing trend, though it may take some time to fully reverse itself. Due to factors that were quite beyond the control of individual states,

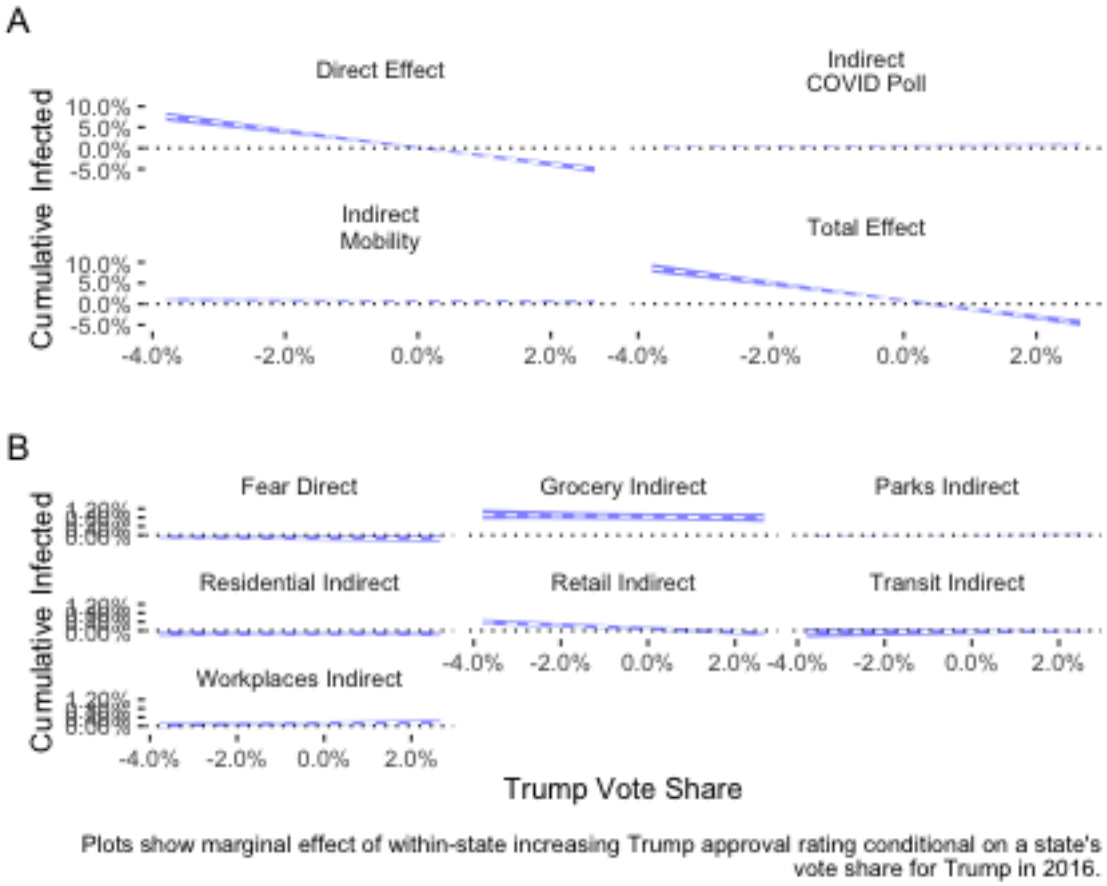


Figure 10: Marginal Effects of Trump Vote Share in 2016 Conditional on State Approval Polls

Figure 11: Mediated Effects of Lockdowns on Google Mobility Data

the spread of COVID-19 occurred far more in states with fewer Trump voters and consequently led to a perception that the disease was associated with liberal states.

5 Conclusion

These empirical results indicate that partisanship is strongly associated with the spread of the disease, though we believe it is very unlikely that the relationship can be described as causal. Although states with higher vote shares for Donald Trump in 2016 and rising Trump approval polls have observed fewer infections, these associations are not explained primarily by mobility data nor personal concern over COVID-19, which are powerful predictors of reduced infections. We believe that further research is necessary to understand whether people have been making flawed inferences about the spread of COVID-19 and partisanship, allowing politicians to use the epidemic as a wedge issue.

The model employed in this article was devised to permit the statistical identification of suppression measures and social, political and economic factors on the spread of COVID-19. It is not intended to be a replacement or alternative to the disease forecasting literature. If anything, this modeling exercise shows why structural epidemiological models are so important: without them it is impossible to project the total number of infected people on a given day. This model's simplicity and ability to use empirical data are its main features, and the hope is that it can be used and extended by researchers looking at government policies and other tertiary factors on the spread of the disease. At the very least, the model provides realistic uncertainty intervals taking into account very real biases in the observed data.

In addition, the model provides insight into how the number of tests undertaken by a given country or area compares to the probably number of infections. These parameter estimates can be used to understand whether a state's testing exceeds, is the same as or is less than the number of infected individuals. Given the wide problem of data scarcity in understanding the disease's spread, we hope this model can be used to make the most of empirical evidence.

Bibliography

Abaluck, Jason, Judith A. Chevalier, Nicholas A. Christakis, Howard Paul Forman, Edward H. Kaplan, Albert Ko, and Sten H. Vermund. 2020. "The Case for Universal Cloth Mask Adoption and Policies to Increase Supply of Medical Masks for Health Workers." *SSRN*.

Alcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. 2020.

"Polarization and Public Health: Partisan Differences in Social Distancing During the Coronavirus Pandemic." *Journal of Public Economics*.

Alesina, Alberto, and Howard Rosenthal. 1995. *Partisan Politics, Divided Government and the Economy*. Cambridge University Press.

"America's Health Rankings 2019 Report." 2019. United Health Foundation. https://www.americashealthrankings.org/learn/reports/2019-annual-report.

Andersen, Martin. 2020. "Early Evidence on Social Distancing in Response to Covid-19 in the United States." *SSRN*.

Brzezinski, Adam, Guido Deiana, Valentin Kecht, and David Van Dijcke. 2020. "The Covid-19 Pandemic: Government Versus Community Action Across the United States." *CEPR Press*, no. 7: 115–47.

Carleton, Tamma, and Kyle C. Meng. 2020. "Causal Empirical Estimates Suggest Covid-19 Transmission Rates Are Highly Seasonal." *Working Paper*. https://t.co/69vR0LUGsT?amp=1.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Cheng, Cindy, Joan Barcelo, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. "COVID-19 Government Response Event Dataset (Coronanet V.1.0)." *Nature Human Behavior*. https://doi.org/https://doi.org/10.1038/s41562-020-0909-7.

Cornelson, Kirsten, and Boriana Miloucheva. 2020. "Political Polarization, Socialfragmentation, and Cooperation During Apandemic." *Working Paper*.

Dave, Dhaval M., Andrew I. Friedson, Kyutaro Matsuzawa, Joseph J. Sabia, and Samuel Safford. 2020. "Black Lives Matter Protests, Social Distancing, and Covid-19." *NBER*.

Dudel, Christian, Tim Riffe, Enrique Acosta, Alyson A. van Raalte, and Mikko Myrskyla. 2020. "Monitoring Trends and Differences in Covid-19 Case Fatality Rates Using Decomposition Methods: Contributions of Age Structure and Age-Specific Fatality." *Working Paper*. https://doi.org/10.31235/osf.io/j4a3d.

Fan, Ying, A. Yesim Orhun, and Dana Turjeman. 2020. "Heterogeneous Actions, Beliefs, Constraints and Risk Tolerance During the Covid-19 Pandemic." *NBER*.

Ferrari, Silvia, and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7): 799–815.

Gadarian, Shana Kushner, Sara Wallace Goodman, and Thomas B. Pepinsky. 2020. "Partisanship, Health Behavior and Policy Attitudes in the Early Stages of the Covid-19 Pandemic." *SSRN*.

Gao, Song, Jinmeng Rao, Yuhao Kang, and Yunlei Liang andJake Kruse. 2020. "Mapping County-Level Mobility Pattern Changes in the United States in Response to Covid-19." *SIGSPATIAL Special* 12 (1): 16–26.

Grossman, Guy, Soojong Kim, Jonah M. Rexer, and Harsha Thirumurthy. 2020. "Political Partisanship Influences Behavioral Responses to Governors' Recommendations for Covid-19 Prevention in the United States." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2007835117.

Grossman, Matt, and David J. Hopkins. 2016. *Assymetric Politics: Ideological Republicans and Group Interest Democrats*. Oxford University Press.

Gu, Youyang. 2020. *Covid19-Projections.com*. https://covid19-projections.com/about/#about-the-model.

Harper, Craig, and Darren Rhodes. 2020. "Ideological Responses to the Breaking of Covid-19 Social Distancing Recommendations." *Psyarchiv*.

Hart, P. Sol, Sedona Chinn, and Stuart Soroka. 2020. "Politicization and Polarization in Covid-19 News Coverage." *Science Communication*.

Havers, C.; Lim, F. P.; Reed, and I Krapiunaya. 2020. "Seroprevalence of Antibodies to Sars-Cov-2 in 10 Sites in the United States." *JAMA Internal Medicine*.

Horowitz, Juliana Menasce, Anna Brown, and Kiana Cox. 2019. "Race in America 2019." *Pew Forum*.

Iyengar, Shanto, and Sean J. Westwood. 2015. "Fear and Loathing Across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59 (3): 690–707.

Jose Lourenco, Mahan Ghafari, Robert Paton. 2020. "Fundamental Principles of Epidemic Spread Highlight the Immediate Need for Large-Scale Serological Surveys to Assess the Stage of the Sars-Cov-2 Epidemic." *medRxiv*. https://doi.org/https://doi.org/10.1101/2020.03.24.20042291.

Koetke, Jonah, Karina Schumann, and Tenelle Porter. 2020. "Trust in Science Increases Conservative Support for Social Distancing." *Open Science Foundation*.

Neil M Ferguson, Gemma Nedjati-Gilani, Daniel Laydon. 2020. "Impact of Non-Pharmaceutical Interventions (Npis) to Reduce Covid19 Mortality and Healthcare Demand." *Imperial College of London Working Paper*. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf.

Painter, Marcus, and Tian Qiu. 2020. "Political Beliefs Affect Compliance with Covid-19 Social Distancing Orders." *SSRN*.

Peak, Corey M., Rebecca Kahn, Yonatan H. Grad, Lauren M. Childs, Ruoran Li, Marc Lipsitch, and Caroline O. Buckee. 2020. "Modeling the Comparative Impact of Individual Quarantine Vs. Active Monitoring of Contacts for the Mitigation of Covid-19." *medRxiv*. https://doi.org/https://doi.org/10.1101/2020.03.05.20031088.

Perkins, T. Alex, Sean M. Cavany, Sean M. Moore, Rachel J. Oidtman, Anita Lerch, and Marya Poterek. 2020. "Estimating Unobserved Sars-Cov-2 Infections in the United States." *Working Paper*. http://perkinslab.weebly.com/uploads/2/5/6/2/25629832/perkins_etal_sarscov2.pdf.

Poole, Keith, and Howard L. Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*.

Poole, Keith T., and Howard L. Rosenthal. 2007. *Ideology & Congress*. Transaction Publishers.

Riou, Julien, Anthony Hauser, Michel J. Counotte, and Christian L. Althaus. 2020. "Adjusted Age-Specific Case Fatality Ratio During the Covid-19 Epidemic in Hubei, China, January and February 2020." *medRxiv*. https://doi.org/https://doi.org/10.1101/2020.03.04.20031104.

Robert Verity, Ilaria Dorigatti, Lucy C Okell. 2020. "Estimates of the Severity of Covid-19 Disease." *medRxiv*. https://doi.org/https://doi.org/10.1101/2020.03.09.20033357.

Ruiyun Li, Bin Chen, Sen Pei. 2020. "Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (Sars-Cov2)." *Science*. https://doi.org/10.1126/science.abb3221.

Sajadi, Mohammad M., Parham Habibzadeh, Augustin Vintzileos, Shervin Shokouhi, Fernando Miralles-Wilhelm, and Anthony Amoroso. 2020. "Temperature, Humidity and Latitude Analysis to Predict Potential Spread and Seasonality for Covid-19." *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3550308.

Seth Flaxman, Axel Gandy, Swapnil Mishra. 2020. "Estimating the Number of Infections and the Impact of Non-Pharmaceutical Interventions on Covid-19 in 11 European Countries." *Working Paper*. https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-13-europe-npi-impact/.

Stekhoven, Daniel J., and Peter Bühlmann. 2012. "MissForest-Non-Parametric Missing Value Imputation for Mixed-Type Data." *Bioinformatics* 28 (1): 112–18.

Tasnim, Samia, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. "Impact of Rumors or Misinformation on Coronavirus Disease (Covid-19) in Social Media." *SocArchiv*. https://doi.org/10.31235/osf.io/uf3zn.

Winship, Christopher, and Robert D. Mare. 1983. "Structural Equations and Path Analysis for Discrete Data." *The American Journal of Sociology* 89 (1): 54–110.

Yuan, Ying, and David P. MacKinnon. 2009. "Bayesian Mediation Analysis." *Psychological Methods* 14 (4): 301–22. https://doi.org/10.1037/a0016972.