A Retrospective Bayesian Model for Measuring Covariate Effects on Observed COVID-19 Test and Case Counts

Robert Kubinec[1,*] Luiz Max Carvalho[2]

April 20th, 2020

Abstract

As the COVID-19 outbreak progresses, increasing numbers of researchers are examining how an array of factors either hurt or help the spread of the disease. Unfortunately, the majority of available data, primarily confirmed cases of COVID-19, are widely known to be biased indicators of the spread of the disease. In this paper we present a retrospective Bayesian model that is much simpler than epidemiological models of disease progression but is still able to identify the effect of covariates on the historical infection rate. The model is validated by comparing our estimation of the count of infected to projections from expert surveys and extant disease forecasts. To apply the model, we show that as of April 20th, there are approximately 3 million infected people in the United States, and these people are increasingly concentrated in states with more wealth, better air quality, fewer smokers, more residents under the age of 18, more public health funding and a history of more cardiovascular deaths. On the other hand, the timing of state declarations of emergency and the proportion of people who voted for President Trump in 2016 are not clear predictors of COVID-19 trends. In addition, we find that the US states have increased testing at approximately the same level in line with infections, suggesting that testing has not yet increased significantly above infection trends.[1]

[1] New York University Abu Dhabi

[2] School of Applied Mathematics, Getulio Vargas Foundation

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

[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 Cindy Cheng and Joan Barcelo.

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 is a confounding variable affecting any estimates of covariates on the observed counts of 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 from the susceptible-infected-recovered (SIR)/susceptible-exposed-infected-recovered (SEIR) papers on 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), 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.

We also show how the model can be applied by measuring the association between U.S. state-level factors and the disease as of April 20th, 2020, including the timing of state of emergency declarations, vote share in the 2016 election for President Donald Trump, the percentage of foreign born and younger residents, gross domestic product (GDP) per capita, and health-related factors. The results show that as of April 20th, there are approximately 3 million infected people in the United States, and these people are increasingly concentrated in U.S. states with more wealth, better air quality, fewer smokers, more people under the age of 18, and a history of more cardiovascular deaths per capita. On the other hand, the proportion of people who voted for President Trump in 2016, the amount of public health funding in a state and the date that a state declared an emergency are not clear predictors of COVID-19 trends.

In addition, the model is able to capture the relationship between the unobserved number of infected individuals and US states' testing capacity. It shows that testing capacity has increased in tandem with infections, suggesting that the growth in tests has not yet exceeded growth in infections.

1 Methods

In this section we present an intuitive overview of the model, and we refer the interested reader to the supplemental materials for a more complete exposition combined with Monte Carlo simulations showing

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

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 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 1 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}

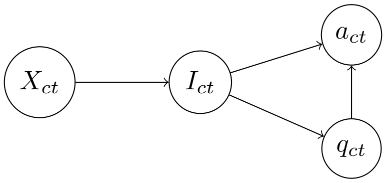


Figure shows the relationship between a covariate X_{ct} representing a policy or social factor influencing the infection rate I_{ct}. Because the infection rate I_{ct} influences both the number of reported tests q_{ct} and reported cases a_{ct}, any regression of a covariate X_{ct} on the reported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

The Bayesian model presented in the supplementary materials provides a full explanation of how to model the infection rate's influence on both cases and tests simultaneously. We further show in the supplementary materials that two restrictions are necessary to identify the sign and rank of the effect of covariates X_{ct} on I_{ct}: the paths $I_{ct} \rightarrow q_{ct}$ and $I_{ct} \rightarrow a_{ct}$ must be strictly positive so that an increasing infection rate will have a non-decreasing effect on cases and tests. Given these restrictions and weakly informative priors on the parameters, it is possible to know whether X_{ct} is associated with increasing or decreasing I_{ct}, though not with respect to the actual number of infected people, only in terms of the covariate's sign or rank relative to other covariates.

Furthermore, we show in the supplementary materials that if further prior information can be put on the ratio between the true number of infected individuals I_{ct} and the number of tests q_{ct}, it is possible to transform inferences from the model to approximate counts of infected people up to the present. Thankfully, this information is available through epidemiological models of the disease, which offer inferences on the number of un-diagnosed cases (Ruiyun Li 2020; Peak et al. 2020). Based on these estimates, we can put an informative prior in the model that the number of tests is likely to be at least 10% of the total number of infected individuals, reaching as high as 10 times the number of infected. With this prior information from disease simulations, it is possible to obtain an empirical estimate of infected rates and covariate effects that are more useful to the general public than solely observed tests and cases, as we demonstrate in the next section.

Finally, 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.

2 Results

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. The testing data starts at March 4th, so we impute the testing data back in time by assuming that the average case/tests ratio stays the same to the origin of the outbreak. Furthermore, as there are discrepancies where the reported number of tests in some states like New Jersey is less than the total number of cases, we impute the number of tests via the case/test ratio for the sample as a whole.

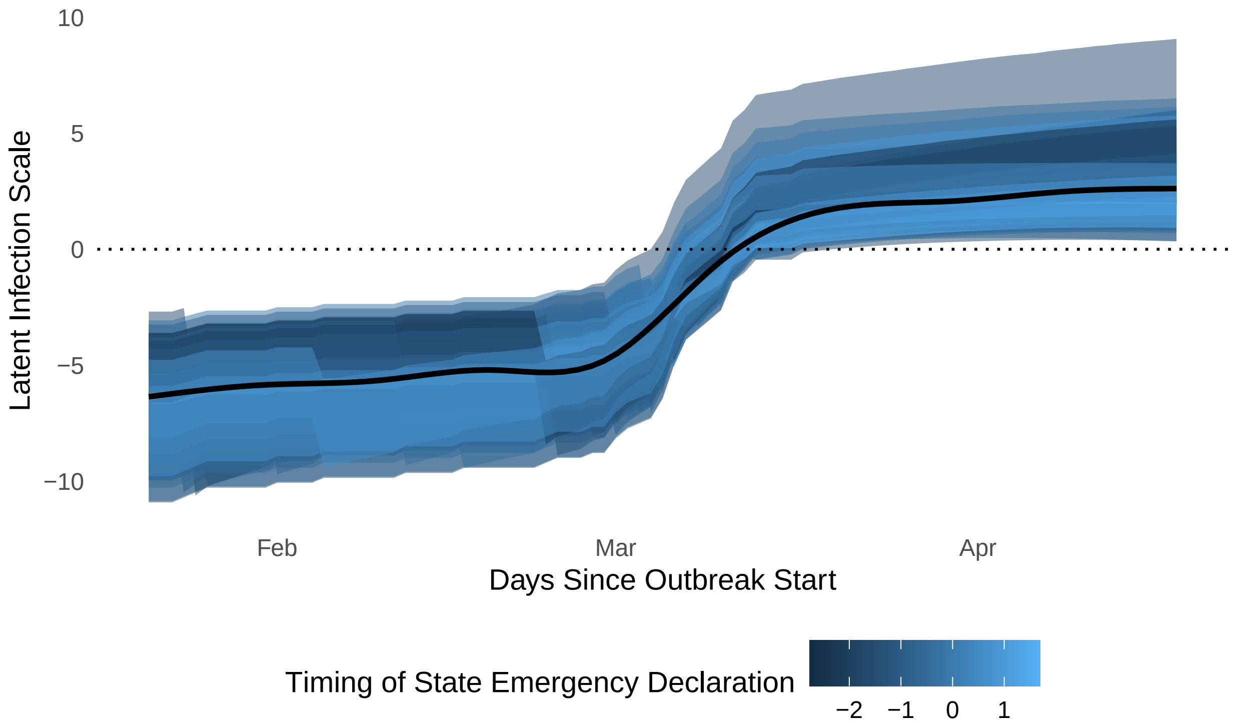
To analyze the effect of suppression policies, we proxy for preparedness to fight the epidemic by including the date that states of emergencies were declared across U.S. states and territories in the model.[3] The suppression covariate is then equal to a vector of days. We further add in state-level 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 from the U.S. Census Bureau, and 2019 state-level average data on air pollution,[4] 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 mean-centered and standardized.

In this section we first fit a partially-identified model without any information about the true number of infected people to demonstrate that it is possible to obtain an estimate of the infection trend, though with substantial uncertainty. We then show how we can use insight from SIR/SEIR models to add in informative prior information on the likely number of infected people and translate the estimates into probable infection rates. We then show that these estimates in fact closely track the predictions of SIR/SEIR models for the U.S. population from March 2020, providing external validity for the method and for these predictions.

Figure 2 shows the 5% - 95% high posterior density (HPD) intervals of the latent infection rate by state

[3] See this site for dates and relevant sources: https://en.wikipedia.org/wiki/U.S._state_and_local_government_response_to_the_2020_coronavirus_pandemic

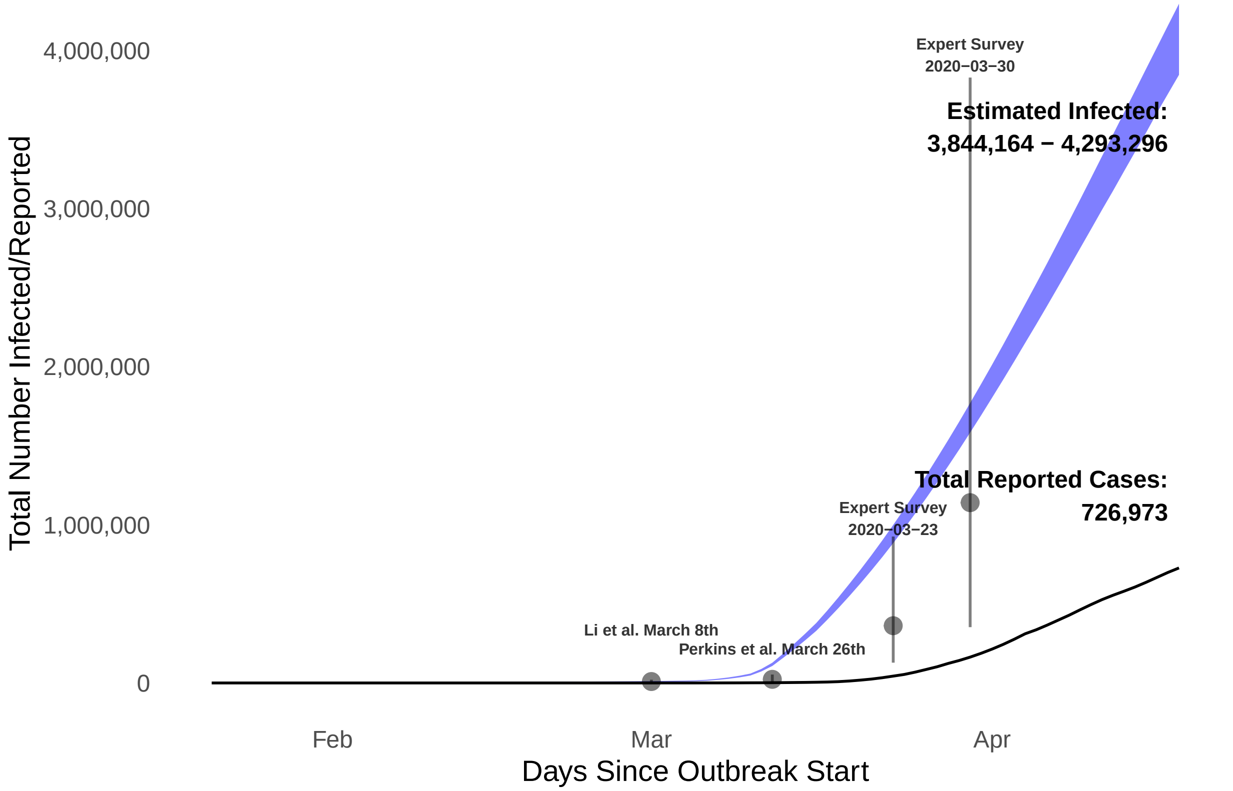
[4] Defined as average exposure of the general public to particulate matter of 2.5 microns or less ($PM_{2.5}$) measured in micrograms per cubic meter (3-year estimate).



5% − 95% HPD Intervals are colored by
when a state declared a state of emergency (standardized count of days).
As the total number of infected people is unknown,
this chart measures the relative marginal infection rates between states.

Figure 2: 5% to 95% HPD Uncertainty Intervals of Partially-Identified Infection Rates by U.S. State with Total Average

since January 1st. The intervals are shaded by the relative time when a state declared a state of emergency, which reveals that state of emergency declarations are correlated with higher infection rates. As can be seen, there is a sharp discontinuity in the plot around March 1st when infection rates began to increase. While it would appear that the rate of increase has leveled off in the last week, that is not a supported inference as the scale is the logit scale, which is similar to the log scale in that higher numbers are farther away than they appear visually. Because the latent scale is not identified, the figure is only showing how the infection rates have evolved from zero to the true but unknown top infection rate. As such, it appears to be slowing as it reaches the top of the scale, but that is simply an illusion of the underlying sigmoid function.



Blue 5% − 95% HPD intervals show estimated infected and the black line shows observed cases from the New York Times.
These estimates are based on the assumption that as few as 10% of cases may be reported based on SIR/SEIR models.

Figure 3: Approximate Total Number of COVID-19 Infected Individuals in the U.S. as of April 20th

By comparison, Figures 3 and 4 show fully-identified models incorporating informative prior information suggesting that the ratio of tests to infected ratio individuals is probably no less than 10% of those infected (though it could very high). This information, as previously mentioned, was derived from simulation and statistical modeling of COVID-19 outbreaks so far suggesting that a large proportion of infected individuals are undetected (Ruiyun Li 2020; Peak et al. 2020). Based on this information, the scale of the latent infection process shown in Figure 2 can be further identified. Figure 3 shows that the likely cumulative number of infected cases, including those who may have recovered or died, is likely approaching 4 million infected individuals in the United States, in line with SIR/SEIR and expert survey projections released recently.[5] Furthermore, Figure 4 shows significant state by state heterogeneity, with New York showing the greatest number of infected, followed by California. There is some suggestive evidence that California's infected count growth may be slowing, though the large uncertainty interval suggests that this inference would be unwise without more data or assumptions.

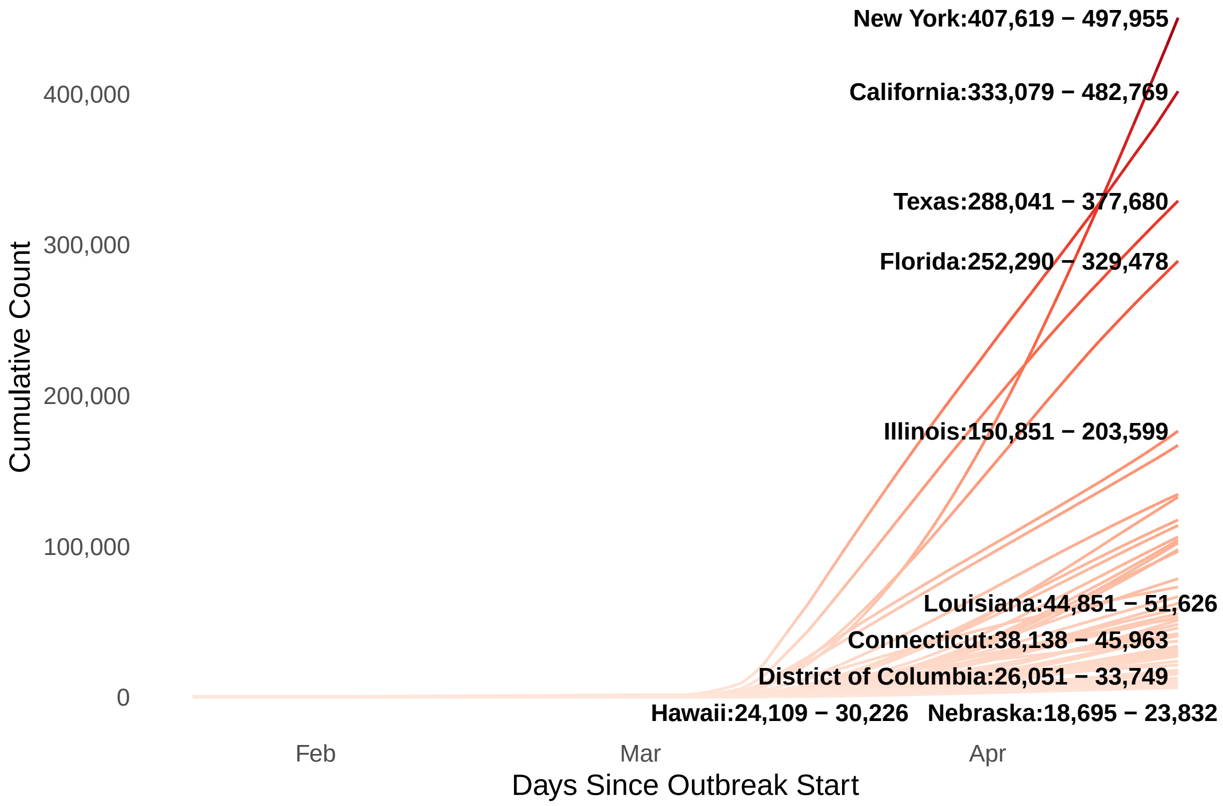
As described in the modeling section, the model does not provide estimates that can be extrapolated into the future. Because the model is estimating empirical infection rates, it is primarily useful for adjusting empirical data, or the count of tests and cases and other background factors. However, because the model is substantively different than the SIR/SEIR simulations used to guide policy choices, it provides helpful external validation of these models incorporating observed information with minimal assumptions.

Figure 7 shows the marginal effect of covariates in the model on the latent infection rate, expressed as the marginal increase in proportions for a standard deviation increase in the covariate. Each covariate has two kinds of marginal effects: a cumulative effect in the top panel and a time-varying effect in the bottom panel, where time is coded as a linear counter since the start of the outbreak (at least one case recorded) in each state. As can be seen, the cumulative effects are generally less precise, but the over-time effects show clear trends. States with larger young and foreign-born populations, a higher GPD per capita and more health care providers are seeing increasingly higher infection rates. By contrast, states with more smokers and better air quality are seeing increasingly fewer infections. These set of associations are not necessarily causal, as they are influenced by the spatial spread of the disease thus far, with outbreaks starting in wealthy coastal states and progressively moving inland. As the disease continues, we expect these associations to shift as we learn more about the effect of suppression policies targeting the disease.

However, it is important to note that state-level Trump vote share is not associated with increasing COVID-19 infections, despite public opinion polling showing that Americans who are more conservative tend to discount the danger posed by the virus.[6] In addition, those states that declared an earlier state of emergency have not yet witnessed a slowing infection rate. It is important to notice, however, that there is a roughly 14 day lag between action and effect. Anything states do, be it beneficial or detrimental, and however big is its effect, will take about a fortnight to manifest itself in the data. At present, however, there is no reason to believe that states that declared emergencies later or have more Trump voters are facing increasing disease

[5] See https://www.nytimes.com/2020/04/01/world/coronavirus-news.html?action=click&module=Spotlight&pgtype=Homepage for a recent overview.

[6] See https://www.vox.com/2020/3/15/21180506/coronavirus-poll-democrats-republicans-trump.



Some lines are labeled with uncertainty of estimates (5% − 95% Interval).
These estimates are based on the assumption that as few as
10% of cases may be reported based on SIR/SEIR models. Does not exclude people
who may have recovered or died.

Figure 4: Average Cumulative Count of Infected People by U.S. State as of April 20th

trends. While we expect these trends to change over time, it is important to note what the empirical data support at present.

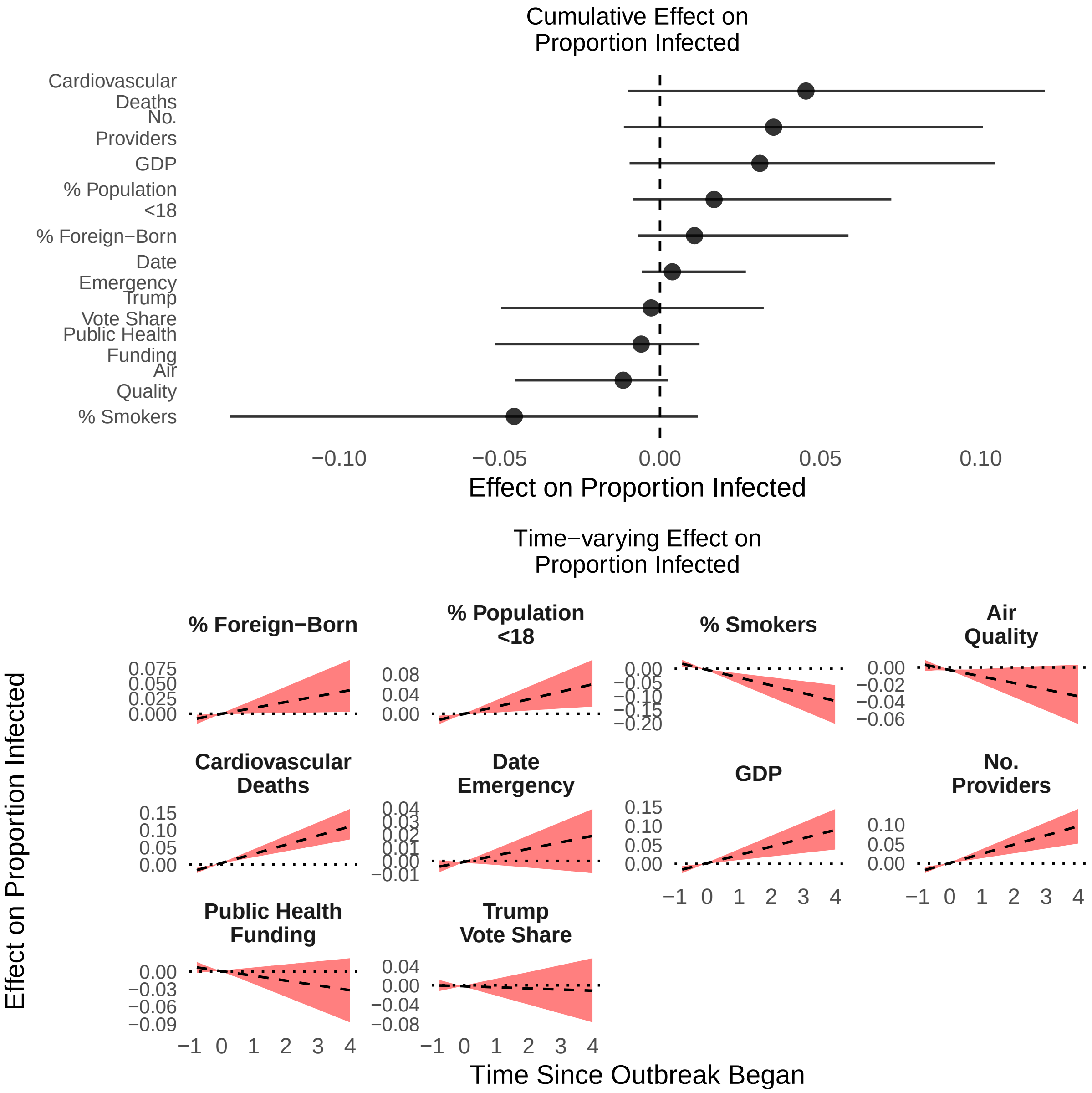
While the use of informative priors is very helpful for obtaining estimates that can be mapped back to actual number of infected persons, as we show in the supplementary information, we can in fact identify covariate effects (or at least their signs) without information about the ratio of tests to infected persons. To demonstrate this, we calculate time-varying marginal effects on the latent infected rate from both the partially-identified and fully-identified models and show them in Figure 6. While there are certainly effect size and uncertainty differences between the models, the estimates are quite similar, and the signs are always in the same direction. For these reasons, we believe this model to be useful even in the case where there is no useful or credible information about the ratio of testing to infected.

Finally, Figure 7 compares the underlying parameter estimates from the latent Bayesian model and a binomial model of the observed case counts with the same covariates as predictors (including time trends). As can be seen, the estimates of these models can wildly diverge, with the observed case count model showing far larger and implausibly precise associations between covariates and case counts. Of particular worry is when covariates are correlated with a state's ability or willingness to test for the virus.[7]

For example, the coefficient for per capita GDP shows an implausibly large positive association (+20 on the *logit* scale) in the observed cases model, suggesting that the more wealth in a state, the lower the infection rate. To show the potential confounding in this result, Figure 8 plots GDP per capita against state COVID-19 tests per capita, revealing states with more infected people–Washington and New York–have implemented many tests and are also more wealthy on average. As such, the observed data model is likely obfuscating the strength of the correlation between GDP per capita and the number of tests with the spread of the disease.

In addition to the estimation of covariates, 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 9. The scale of the y axis shows how much tests will increase for a unit increase of the infection rate on the logit scale. Given the multiplicative nature of the logit scale, the best way to understand the parameter is given a particular infection rate. For example, if we use a recent estimate of the US population at 330 million, and take the recent observed case count (700,000) as the quite conservative estimate of infected, then we have an infected rate of 0.002 for the US population. Given a testing parameter of 1.03 and a baseline of equal testing/infection capacity, then a unit increase (logit scale) in the infection rate from 0.002 to 0.0057451 would increase the testing rate to 0.006. As such, the model provides a relatively nuanced relationship that can help provide insight into how much states are

[7] We note that testing may be conducted for the virus using PCR, or for the disease, which would also include serology. This also may explain why the observed model in Figure 7 shows such a high effect of GDP per capita on reducing infection counts.



Marginal effects calculated as a 1−standard deviation change in a covariate on the latent infection rate. 5% − 95% high posterior density intervals derived from 100 Markov Chain Monte Carlo posterior draws.

Figure 5: Marginal Effects of Covariates on Latent Infection Rates for U.S. States

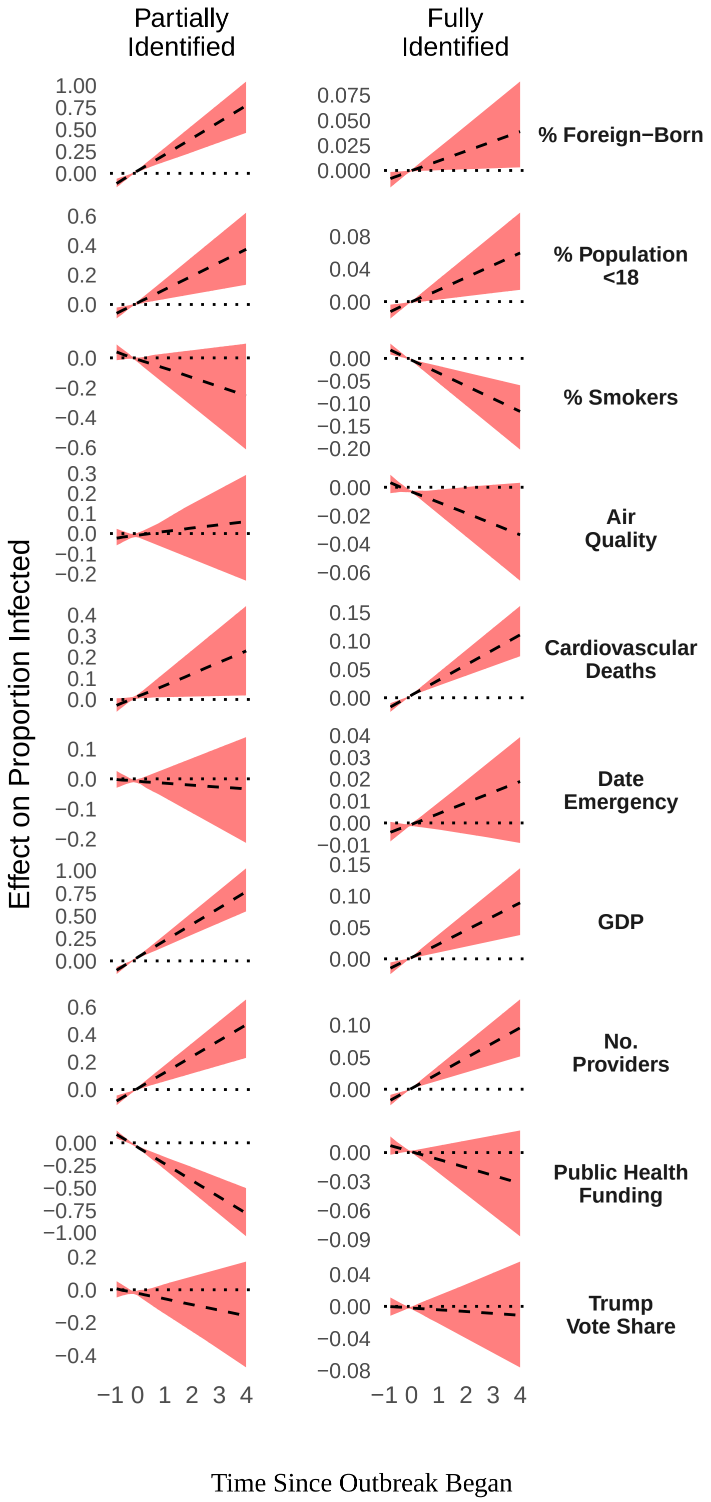
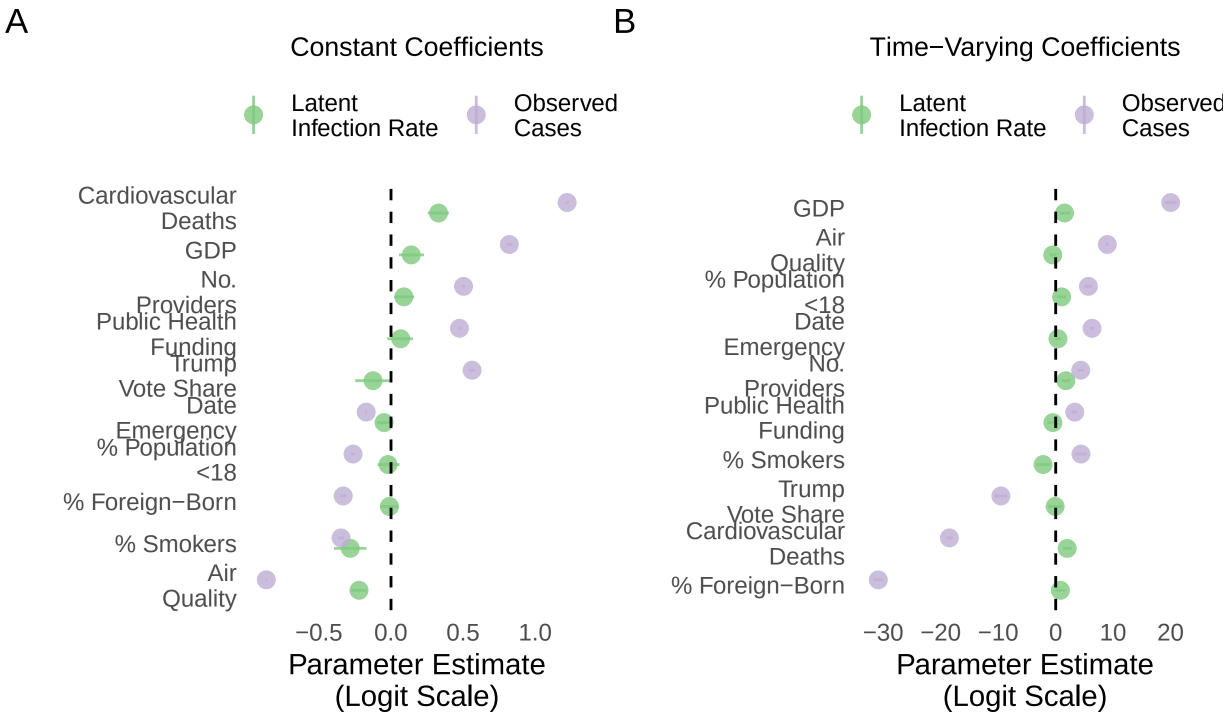


Figure 6: Comparison Of Time-Varying Covariate Marginal Effects from Partially- and Fully-Identified Models



Both models were fit with the same covariates and specification, and using the same Markov Chain Monte Carlo samplers. Parameter values are on the logit scale. Intervals are 5% − 95% high posterior density intervals.

Figure 7: Comparison of Effects from Latent and Observed Data Models

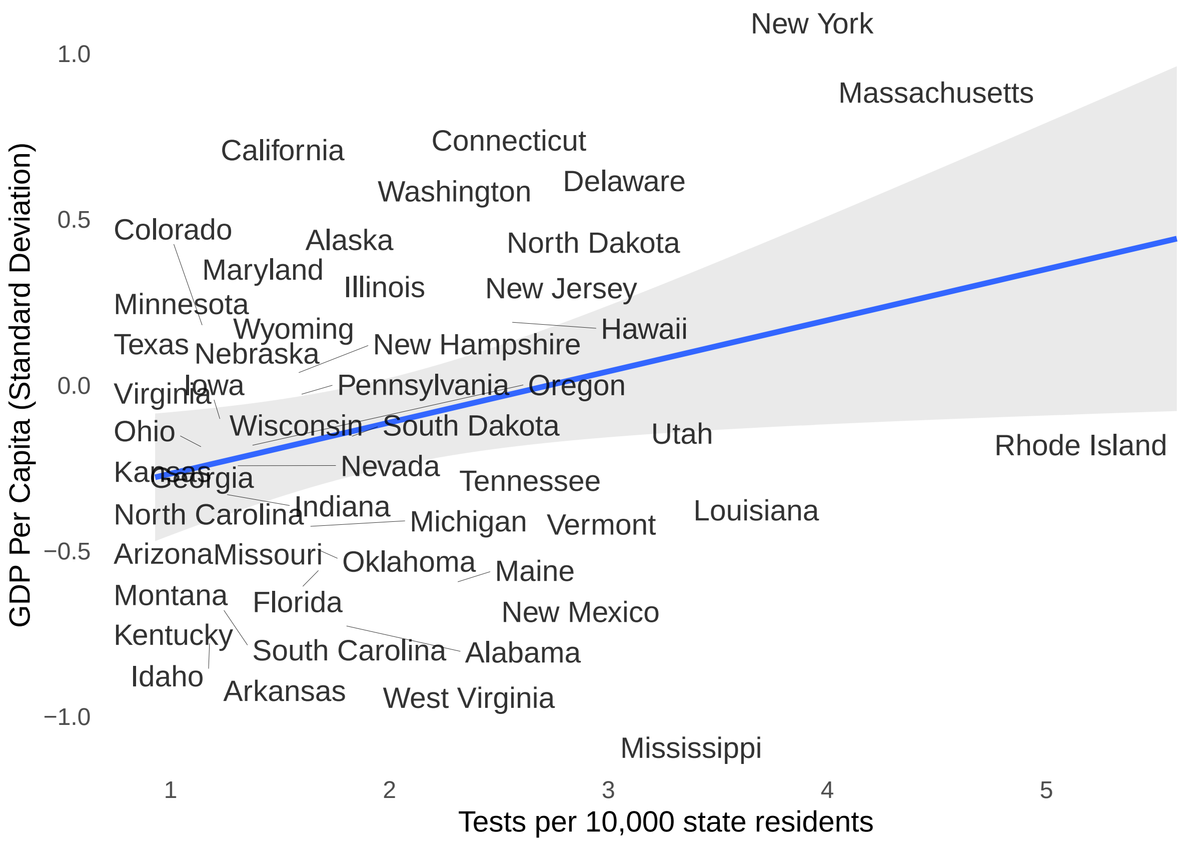


Figure 8: Comparison of GDP Per Capita and COVID-19 Tests per 10,000 Residents by State

able to increase tests in proportion to infection rates.

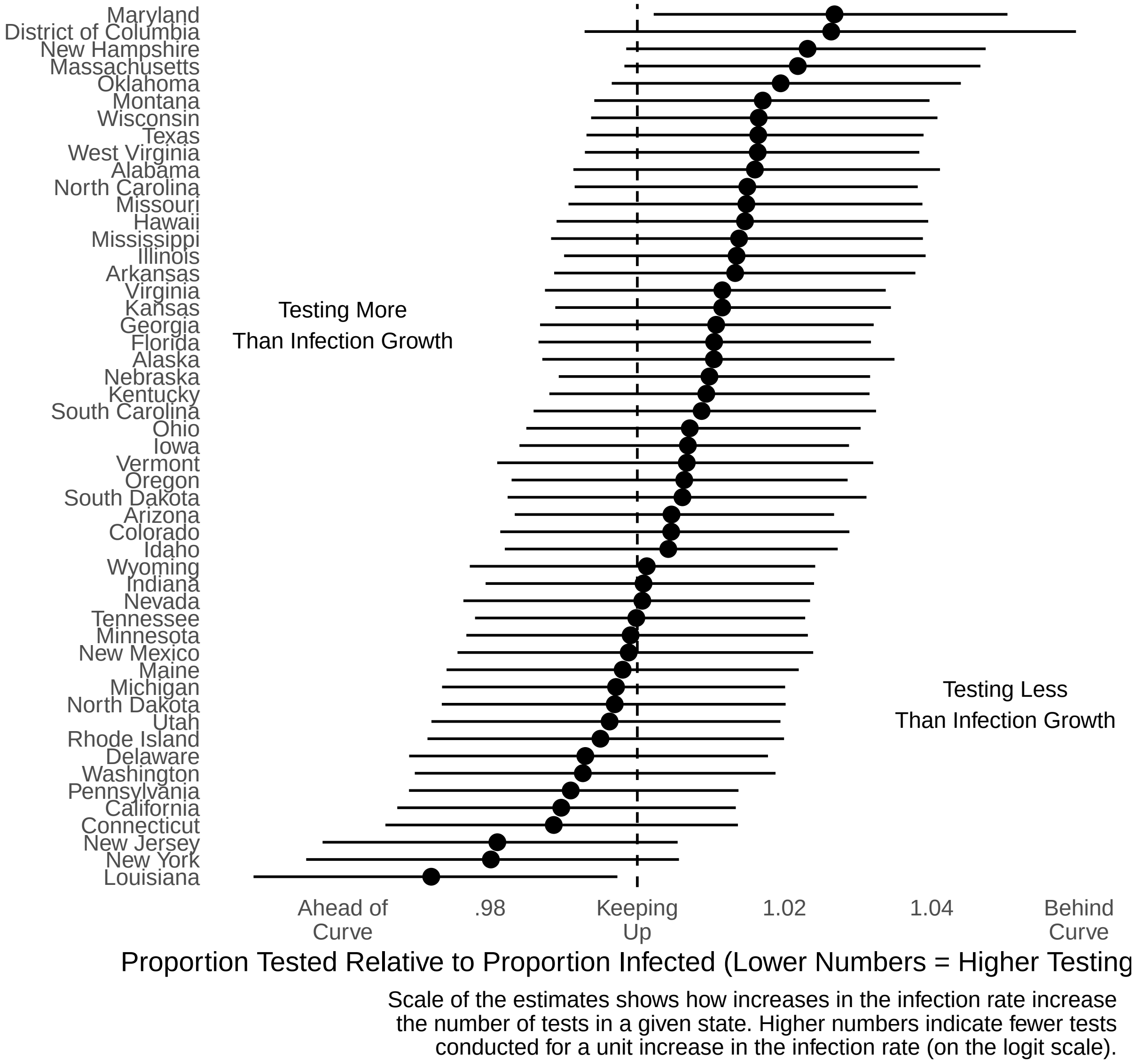


Figure 9: Measuring States' Testing Rates Relative to Infection Rates

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.

3 Conclusion

This model was devised to permit the 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, especially as this model relies on SIR/SEIR estimates for full identification. If anything, this modeling exercise shows why explicit mechanistic epidemiological models are so important: without them it is literally impossible to know 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.

To fit the model, it is necessary to have at least an estimate of how many tests have been conducted. The CoronaNet project is currently working to obtain testing data, in addition to information about government policy responses to COVID-19, in an effort to better understand the role and success of variation in country policy responses to date.

Bibliography

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

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.

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.

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.

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.

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.

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/.

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.