Estimating the number of SARS-CoV-2 infections in the United States

Dayton G. Thorpe[1]* and Kelsey Lyberger[2]
April 7 2020
1. Independent researcher
2. Department of Evolution and Ecology, University of California, Davis, Davis, California
*Corresponding author: dayton@berkeley.edu

Abstract

We apply a model developed by The COVID-19 Response Team [S. Flaxman, S. Mishra, A. Gandy, *et al.*, "Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries," tech. rep., Imperial College London, 2020.] to estimate the total number of SARS-CoV-2 infections in the United States. Across the United States we estimate as of 6 April 2020 the fraction of the population infected was 4.8% [3.2%, 8.0%], 39 times the portion of the population with a positive test result. Excluding New York state, which we estimate accounts for over half of infections in the United States, we estimate an infection rate of 2.0% [1.7%, 2.6%].

We include the timing of each state's implementation of interventions including encouraging social distancing, closing schools, banning public events, and a lockdown / stay-at-home order. We assume fatalities are reported correctly and infer the number and timing of infections based on the infection fatality rate measured in populations that were tested universally for SARS-CoV-2. Underreporting of deaths would drive our estimates to be too low. This model does not include effects of herd immunity; in states where the estimated infection rate is very high - namely, New York - our estimates may be too high.

I. Introduction

The first case of SARS-CoV-2, the virus that causes COVID-19, diagnosed in the United States was discovered on January 21st, 2020 [1]. Since then, limited availability of SARS-CoV-2 tests has interfered with efforts to track the spread of the virus. As of April 6, 2020, according to Johns Hopkins University there were 396,200 reported cases in the United States [2]. The true number of cases could be far higher.

Europe has faced similar constraints on SARS-CoV-2 testing. The COVID-19 Response Team (CRT) at Imperial College London has responded to this challenge, in part, by developing a Bayesian hierarchical model to estimate the true number of infections and the impact of policies intended to slow the epidemic [3]. In this paper, we apply the CRT model to individual states in the United States, plus Puerto Rico and Washington, D.C. We remove states with fewer than 10 reported fatalities. We also reproduce the CRT estimates for 11 European countries with an additional week of data on reported fatalities.

We briefly summarize the CRT model here, but refer to [3] for full details. The model takes three inputs from each modeled population: 1) the age distribution, 2) the time series of known COVID-19 fatalities, and 3) the start dates of four mitigation policies (encouraging social distancing, closing schools and universities, banning public events, and a lockdown / stay-at-home order). It also relies on four values estimated from prior research: 1) the infection fatality ratio (IFR) by age, 2) the serial interval distribution, 3) the infection-to-onset distribution, and 4) the onset-to-death distribution. These values are not extracted from population data based on testing people who are already symptomatic. Instead, they are measured from several cases in which all members of a small population at high risk of infection were tested and observed, whether or not they showed symptoms. We discuss the IFR further in section III.

There are several possible sources of error we do not account for. First, we assume COVID-19 fatalities are reported correctly. If they are underreported, the infection estimates here will be underestimates. Second, the model extrapolates exponential growth rates without reference to the size of the population. If the estimated attack rate approaches the level of herd immunity, the model likely overestimates the total number of infections. Third, we assume the IFR is known exactly. Propagating the error from the IFR would increase the width of our confidence intervals.

The model estimates the baseline reproduction number of the virus without intervention, R_0, and the impact of the considered interventions. It fits one value for each of these variables across all populations simultaneously. The impact of any given intervention is the same for all populations. Most of the interventions were implemented in Europe before the United States. Therefore, the model estimates the impact of these interventions mostly based on the trajectory of the epidemic in Europe. Using the number of fatalities as of a given time, we can infer the number of infections in the past, and extrapolate them to the present using the fitted reproduction numbers under different policy regimes. The model does not use the reported number of infections.

This approach differs from the CRT's model in [4], which is a detailed SIR model. Instead, this model more closely matches that of the UW IHME [5], which forecasts exponential growth rates after different policies are implemented based on historical growth rates in countries and states that previously implemented similar policies.

II. Results

As in Europe, we estimate infections in the United States are much higher than reported infections. As of 6 April 2020 we estimate the attack rate in the United States was 4.8% [3.2%, 8.0%], 39 times the reported infections of 396,200. The difference between estimated and reported infections varies substantially by state. In New York, there are an estimated 9,000,000 infections, 70 times the reported value of 130,000. Whereas, in Oregon, our estimate of infections is only 8.9 times the reported value.

In Figure 1 and Table 1 we report the estimated fraction of the population infected in each state. The estimated value in New York is so high that it is likely close to herd immunity. If the spread is already slowing in New York due to herd immunity, the model would overestimate the number of infections. The state with the second highest attack rate, New Jersey, at 15.60% [7.12%, 29.29%], is well below herd immunity.

Estimated Percentage Infected by State

Figure 1. Log-transformed percent of the population infected for each state as of 6 April 2020, estimated from our model.

	Estimates as of 6 April 2020		
State	**Infections**	**% of population infected**	**95% confidence interval**
Alabama	30,003	0.62%	[0.13%, 1.76%]
Arizona	63,510	0.91%	[0.20%, 2.45%]
Arkansas	12,164	0.41%	[0.03%, 1.46%]
California	273,501	0.70%	[0.30%, 1.38%]
Colorado	117,723	2.13%	[0.79%, 4.30%]
Connecticut	164,097	4.58%	[1.92%, 9.27%]

Delaware	5,378	0.57%	[0.05%, 1.89%]
DC	22,615	3.30%	[0.53%, 10.16%]
Florida	371,700	1.80%	[0.70%, 3.62%]
Georgia	213,385	2.07%	[0.88%, 4.11%]
Idaho	6,483	0.38%	[0.02%, 1.36%]
Illinois	310,798	2.42%	[1.05%, 4.87%]
Indiana	116,644	1.76%	[0.64%, 3.52%]
Iowa	32,106	1.02%	[0.18%, 3.51%]
Kansas	18,260	0.63%	[0.06%, 1.64%]
Kentucky	50,501	1.14%	[0.26%, 3.20%]
Louisiana	462,699	9.92%	[4.56%, 19.90%]
Maine	3,373	0.25%	[0.00%, 1.31%]
Maryland	136,895	2.28%	[0.65%, 5.57%]
Massachusetts	237,397	3.48%	[1.22%, 7.96%]
Michigan	780,136	7.83%	[3.41%, 15.86%]
Minnesota	16,887	0.31%	[0.04%, 1.05%]
Mississippi	64,649	2.16%	[0.44%, 5.81%]
Missouri	48,440	0.80%	[0.14%, 2.44%]
Nevada	26,696	0.91%	[0.16%, 2.82%]
New_Jersey	1,385,696	15.60%	[7.12%, 29.28%]
New_Mexico	5,946	0.28%	[0.02%, 0.97%]
New_York	9,118,672	46.48%	[21.33%, 97.37%]
North_Carolina	33,882	0.33%	[0.06%, 0.96%]
Ohio	99,303	0.85%	[0.31%, 1.80%]
Oklahoma	29,952	0.76%	[0.13%, 1.87%]
Oregon	10,104	0.25%	[0.05%, 0.69%]
Pennsylvania	171,352	1.34%	[0.38%, 2.96%]
Puerto_Rico	10,065	0.30%	[0.04%, 0.92%]
South_Carolina	27,557	0.56%	[0.08%, 1.42%]

Tennessee	73,586	1.11%	[0.27%, 2.53%]
Texas	230,096	0.83%	[0.27%, 1.93%]
Utah	57,683	1.89%	[0.21%, 6.37%]
Virginia	28,969	0.34%	[0.08%, 1.05%]
Washington	82,491	1.13%	[0.63%, 1.72%]
Wisconsin	67,542	1.17%	[0.42%, 2.13%]

Table 1. Fraction of the population infected for each state as of 6 April 2020, estimated from our model.

In Table 2, we compare the weighted average attack rate in the United States to the same 11 European countries as in [3].

	Estimates as of 6 April 2020		
Country	Infections	% of population infected	95% confidence interval
Austria	92,577	1.03%	[0.43%, 2.14%]
Belgium	1,229,813	10.61%	[4.48%, 19.10%]
Denmark	92,307	1.59%	[0.67%, 3.54%]
France	4,862,580	7.45%	[3.74%, 11.73%]
Germany	1,193,005	1.42%	[0.66%, 2.64%]
Italy	5,751,956	9.51%	[5.37%, 16.20%]
Norway	38,386	0.71%	[0.20%, 1.61%]
Spain	8,937,038	19.11%	[9.89%, 38.08%]
Sweden	587,680	5.82%	[2.76%, 11.95%]
Switzerland	515,631	5.96%	[2.84%, 9.94%]
United_Kingdom	5,112,660	7.53%	[3.89%, 14.10%]
United States*		4.74%	[3.15%, 7.94%]
United States ex. New York		2.00%	[1.7%, 2.6%]

Table 2. Fraction of the population infected for each of 11 European countries and the United States as of 6 April 2020, estimated from our model. *The fraction of the population infected for the United States is the population-weighted average of the states we are able to model. Those states sum to 96% of the US population.

In each population, the model infers the time series of infections that most likely gave rise to the observed time series of fatalities. In the appendix, in Figure A1 through Figure A52, we show the estimated and reported time series of infections and fatalities. Model parameters are fit simultaneously across all populations. The second panel in each of these figures shows how well the model reproduces the observed fatalities. The gap between estimated and reported fatalities can be interpreted, in part, as a measure of how well the model reproduces the epidemic in each population.

III. Methods

We retrieved dates for policy interventions as follows. We took lockdown dates from the *New York Times* rounded to the nearest day (lockdowns going into effect past noon are recorded as the next day) [6]. School closure dates are from [7]. Following [3], if a school closure began on a Monday, we set the date of the school closure to the previous Saturday. We set social distancing and the ban of public events to Mar 16, 2020 for all states, when federal guidelines were put into place [8]. See the dates we assign for each region in each state in the Appendix (Table A1). Age distributions for each state come from the US Census Bureau [9]. The time series of reported fatalities and cases in the United States was crowdsourced from Wikipedia [1]. The time series of reported fatalities and cases in the 11 European countries comes from the European Centre of Disease Control (ECDC) [10]. We change the name of the country "Georgia" to "Georgia(country)" in the ECDC data set to avoid confusion with the state of Georgia. In the ECDC data we also delete the entries for Puerto Rico, for which we get duplicate data as a US territory. For the 11 European countries, IFR comes from [3]. All other IFR's are age-adjusted from [11].

All of our code and input data are available at https://drive.google.com/drive/folders/1DDST4e3B875wxsVApJTKLn94seM5lZbx?usp= sharing.

References

[1] "2019–20 coronavirus pandemic data/United States medical cases," *Wikipedia*. Retrieved 6 April 2020, from

https://en.wikipedia.org/wiki/Template:2019–20_coronavirus_pandemic_data/United_States_medical_cases

[2] E. Dong, H. Du, L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect Dis*. 2020.

[3] S. Flaxman, S. Mishra, A. Gandy, *et al.*, "Estimating the number of infections and the impact of non- pharmaceutical interventions on COVID-19 in 11 European countries," tech. rep., Imperial College London, 2020.

[4] N.M. Ferguson, D. Laydon, G. Nedjati-Gilani, *et al.*, "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand," tech. rep., Imperial College London, 2020.

[5] IHME COVID-19 health service utilization forecasting team, "Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months," tech. rep., Institute for Health Metrics and Evaluation, 2020.

[6] "See Which States and Cities Have Told Residents to Stay at Home," *The New York Times*. Retrieved 5 April 2020, from
https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html

[7] "Map: Coronavirus and School Closures," *Education Week*. Retrieved 5 April 2020, from
https://www.edweek.org/ew/section/multimedia/map-coronavirus-and-school-closures.html

[8] "15 Days to Slow the Spread," 16 March 2020. Retrieved 5 April 2020, from
https://www.whitehouse.gov/articles/15-days-slow-spread/

[9] "2018 ACS 1-Year Supplemental Estimates with a Population Threshold of 20,000 or More," *United States Census Bureau*. Retrieved 5 April 2020, from
https://www.census.gov/acs/www/data/data-tables-and-tools/supplemental-tables/

[10] "Download today's data on the geographic distribution of COVID-19 cases worldwide," *European Centre of Disease Control*. Retrieved 6 April 2020, from
https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-geographic-disbtribution-worldwide-2020-04-06.xlsx

[11] R. Verity, L.C. Okell, I. Dorigatti, *et al.*, "Estimates of the severity of coronavirus disease 2019: a model-based analysis," *Lancet Infect Dis*. 2020.

Appendix

Country	schools_universities	public_events	lockdown	social_distancing_encouraged	self_isolating_if_ill

Alabama	3/14/20	3/16/20	4/5/20	3/16/20	4/5/20
Arizona	3/14/20	3/16/20	4/1/20	3/16/20	4/1/20
Arkansas	3/17/20	3/16/20	5/1/20	3/16/20	5/1/20
California	3/19/20	3/16/20	3/19/20	3/16/20	3/19/20
Colorado	3/21/20	3/16/20	3/26/20	3/16/20	3/26/20
Connecticut	3/17/20	3/16/20	3/24/20	3/16/20	3/24/20
Delaware	3/14/20	3/16/20	3/24/20	3/16/20	3/24/20
DC	3/14/20	3/16/20	4/1/20	3/16/20	4/1/20
Florida	3/14/20	3/16/20	4/3/20	3/16/20	4/3/20
Georgia	3/18/20	3/16/20	4/3/20	3/16/20	4/3/20
Idaho	3/21/20	3/16/20	3/26/20	3/16/20	3/26/20
Illinois	3/17/20	3/16/20	3/22/20	3/16/20	3/22/20
Indiana	3/19/20	3/16/20	3/25/20	3/16/20	3/25/20
Iowa	3/14/20	3/16/20	5/1/20	3/16/20	5/1/20
Kansas	3/18/20	3/16/20	3/30/20	3/16/20	3/30/20
Kentucky	3/14/20	3/16/20	3/26/20	3/16/20	3/26/20
Louisiana	3/14/20	3/16/20	3/24/20	3/16/20	3/24/20
Maine	4/2/20	3/16/20	4/2/20	3/16/20	4/2/20
Maryland	3/14/20	3/16/20	3/31/20	3/16/20	3/31/20
Massachusetts	3/17/20	3/16/20	3/24/20	3/16/20	3/24/20
Michigan	3/14/20	3/16/20	3/24/20	3/16/20	3/24/20
Minnesota	3/18/20	3/16/20	3/28/20	3/16/20	3/28/20
Mississippi	3/20/20	3/16/20	4/4/20	3/16/20	4/4/20
Missouri	3/19/20	3/16/20	4/6/20	3/16/20	4/6/20
Nevada	3/14/20	3/16/20	4/1/20	3/16/20	4/1/20
New_Jersey	3/18/20	3/16/20	3/22/20	3/16/20	3/22/20
New_Mexico	3/14/20	3/16/20	3/24/20	3/16/20	3/24/20
New_York	3/18/20	3/16/20	3/23/20	3/16/20	3/23/20
North_Carolina	3/14/20	3/16/20	3/31/20	3/16/20	3/31/20

Ohio	3/17/20	3/16/20	3/24/20	3/16/20	3/24/20
Oklahoma	3/17/20	3/16/20	3/29/20	3/16/20	3/29/20
Oregon	3/14/20	3/16/20	3/24/20	3/16/20	3/24/20
Pennsylvania	3/14/20	3/16/20	4/2/20	3/16/20	4/2/20
Puerto_Rico	3/14/20	3/16/20	3/16/20	3/16/20	3/16/20
South_Carolina	3/14/20	3/16/20	3/29/20	3/16/20	3/29/20
Tennessee	3/20/20	3/16/20	4/1/20	3/16/20	4/1/20
Texas	3/21/20	3/16/20	4/2/20	3/16/20	4/2/20
Virginia	3/14/20	3/16/20	3/30/20	3/16/20	3/30/20
Washington	3/17/20	3/16/20	3/23/20	3/16/20	3/23/20
Wisconsin	3/18/20	3/16/20	3/25/20	3/16/20	3/25/20

Table A1. Dates we assign for each suppression measure in each state.

DC

Reported Fatalities — Estimated Fatalities

— Estimated Infections Reported Infections

DC