Supplementary materials for "Early warning of vulnerable counties in a pandemic using socio-economic variables"

Damian J. Ruck,[1,2] R. Alexander Bentley,[1] and Joshua Borycz[3]

[1]Anthropology Dept., University Tennessee, Knoxville, TN, 37996 USA

[2]Network Science Institute, Northeastern University, Boston, MA, 02115 USA

[3]Sarah Shannon Stevenson Science and Engineering Library,
Vanderbilt University, Nashville, TN 37203 USA

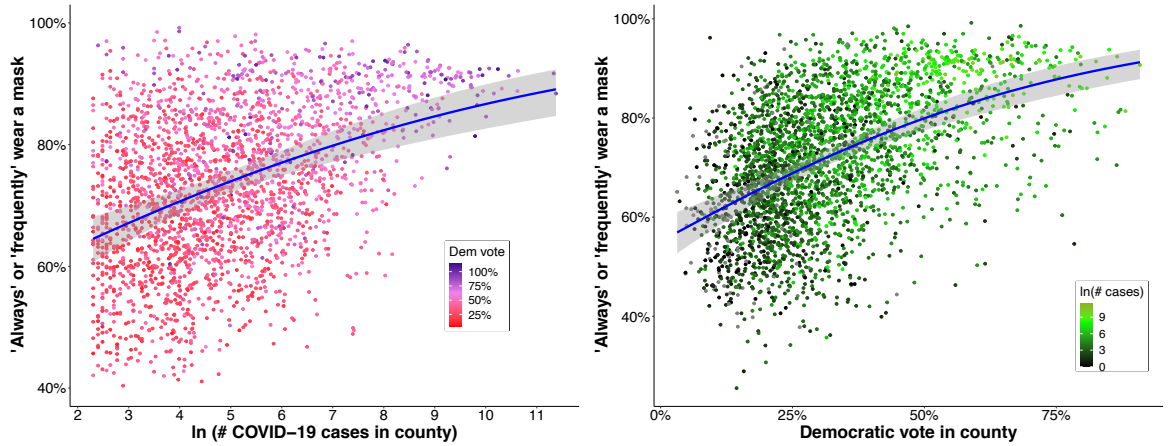

Figure S1: Fraction of survey respondents ($n = 250,000$) in U.S. counties who replied 'Always' or 'Frequently', when asked how often they wear a mask in public when within six feet of another person (Katz et al. 2020), versus (left) logarithm of number of COVID-19 cases reported in the county in first week of July 2020, and (right) Democratic voteshare in the 2016 election. Each datum point represents a U.S. county; each plot is fit with a binomial logistic regression (blue curve).

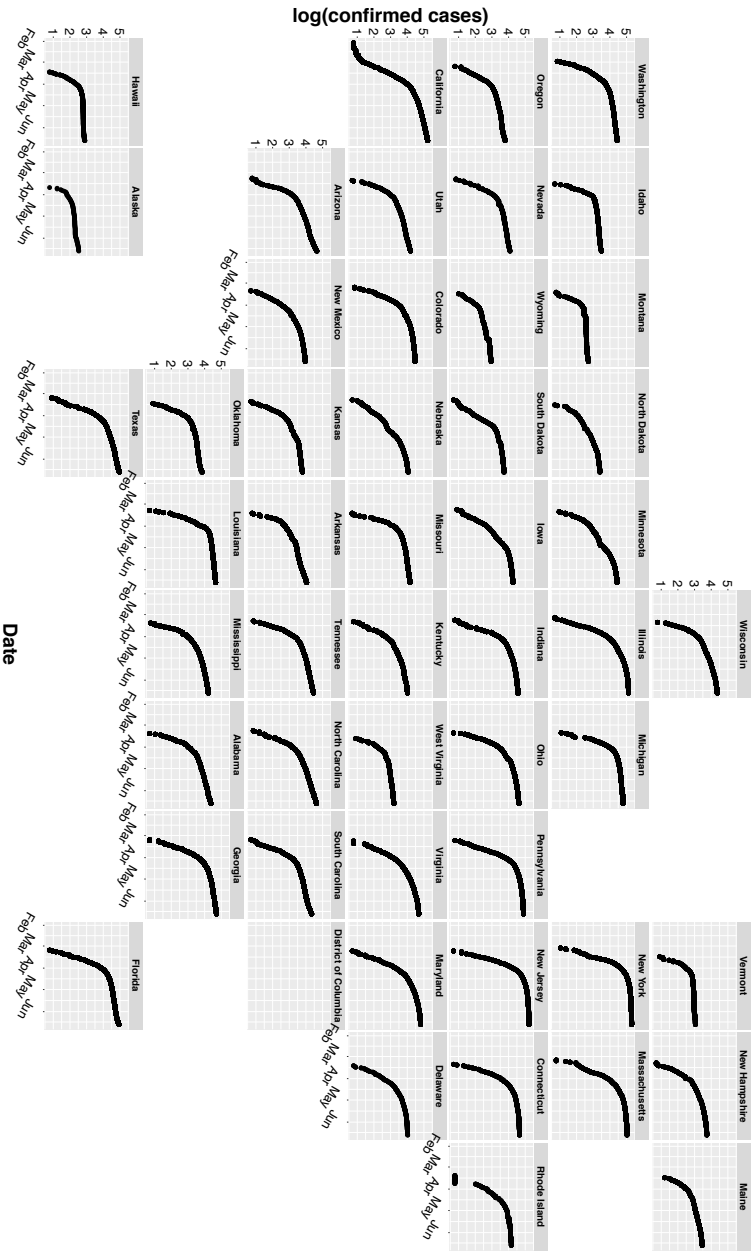

Figure S2: Plots of number of confirmed cases in 50 U.S. states. Each timeline ends on April 20, 2020, and begins on the day when there were at least five cases reported in the state. The y-axes show log-transformed number of cases. The timeline for each state is February 3 to April 20. Counties with fewer than 5 cases were excluded.

S1 Data sources

S1.1 COVID-19 data

County-scale data for confirmed cases and death count were obtained from usafacts.org (USAFacts 2020), which aggregates data of the Centers for Disease Control (CDC) and local public health agencies.

S1.2 Covariate data

S1.2.1 Population size p, population density ρ, mean age a

Estimates of the population size, p, and mean age, a in each county are from the U.S. Census Bureau's Population Estimates Program (U.S. Census 2020a). We use estimates of population size, p for 2018, which the Census estimates using annual information on births, deaths and migration in each county to update the baseline population measurement from the 2010 U.S. census. Population density, ρ, is estimated using population estimates for 2018 and known fixed land areas for each county. To calculate mean age, a, we used total population to calculate the proportion of the population in each designated age bracket at the county-level. Census age data are in 5-year age brackets (0-5, 5-10, 10-15 up to 85+), and calculated the mean age using the weighted sum of all age brackets.

S1.2.2 Voteshare index

Mass attitudes and values in a society and institutions can affect social distancing practices and other protective behaviours. As a proxy for cultural attitudes by U.S county, we used data for the differences between Democrat and Republican voting provided by the MIT Election Lab (https://electionlab.mit.edu/data). We used data from the November 2016 presidential election because of the high turnout (55.7%). We define voteshare, v, in county i, as:

$$v_i = (D_i - R_i)/p_i \tag{1}$$

where D_i and R_i are the numbers of people who voted Democratic and Republican, respectively, in county i with population p_i.

S1.2.3 Number of hospitals and ICU beds per capita

County-level data on numbers of hospitals and intensive care beds per capita (ICU) were obtained from Kaiser Health News (Schulte et al. 2020), which collates data reported annually by hospitals to the Centers for Medicare and Medicaid (except Veterans Affairs hospitals, which do not report). The data include ICU beds in: intensive care units, surgical intensive care units, coronary care unit and burn intensive care units. We use ICU beds per capita, as the range runs from zero ICU beds in some low-population counties to 2,126 ICU beds in Los Angeles County.

S1.2.4 Obesity rate

Age-adjusted data on obesity rates at the county level for the year 2015 (most recent data) are available from the CDC Behavioral Risk Factor Surveillance System (Centers for Disease Control and Prevention 2020a). We exclude Alaska from the data because we do not have county-level data for that state.

S1.2.5 Fraction of population without heath insurance

Fraction of uninsured population for U.S. counties, U, is from the Small Area Health Insurance Estimates (SAHIE) from the U.S. Census Bureau (U.S. Census Bureau 2020b). The data are estimated from the American Community Survey, which allows for measurement in counties smaller than 65,000 people. The SAHIE measurements use of supplemental information from census and local administrative records. We used SAHIE data from 3,108 counties in the year 2017. The median county was 10.6% uninsured population, with a range from 2.3% to 33.7% across all the counties.

S1.2.6 Median household income

Median household income data per county are published by the U.S. Census Bureau for 2017 from the Small Area Income and Poverty Estimates (U.S. Census Bureau 2020c). The SAIPE county-level household income estimates are derived from combining data from all these sources including tax returns, the American Community Survey and Current Population Survey.

S1.2.7 Black, Hispanic and Native ethnic populations

We measure the percentage of each county's population that self-identify as Black, Hispanic or Native American from the U.S. Census Bureau's Population Estimates Program (U.S. Census 2020a).

S1.2.8 Public transportation

We use data from the U.S. Census Bureau (U.S. Census 2020a) to measure the proportion of thew working population in each county that use public transportation to get to work. Specifically, the percentage of workers over 16 that self-report regularly taking any form of public transportation (excluding taxis) to work.

S1.2.9 Residential overcrowding

We use data from the U.S Census Bureau (U.S. Census 2020a) to derive three measures of residential overcrowding by county. We use self-report measures for the number of people per house, people per room (a control for average size of house) and people per bedroom.

S1.2.10 Employment by category

We use data from the U.S Census Bureau (U.S. Census 2020a) to measure the fraction of each county's adult population (16 years and over) working in five high level employment categories. These categories are defined by the U.S census Bureau as: management, business, science, and arts occupations (J_p); service occupations (J_s); sales and office occupations (J_o); natural resources, construction, and maintenance occupations (J_t) and production, transportation, and material moving occupations (J_r).

S1.2.11 Prison population

We measured the size of the incarcerated population in each U.S county using the 'Incarceration Trends' dataset from the Vera Institute of Justice (Vera Institute 2020). This represents the sum of the prison and jail populations for an average day in 2016. This is the latest date where both local prison and state prison data are both available.

S1.2.12 Education

We measure educational attainment using data from the U.S Census Bureau (U.S. Census 2020a). To measure the proportion of the population at the opposite poles of the education continuum, we use the fraction of the population with 'no high school education' and with 'a bachelors degree or higher'.

S1.2.13 Heath: diabetes, hypertension, air pollution

To measure underlying heath conditions in a county, we collected data on diabetes,m hypertension and exposure to particulate pollution. We measure the fraction of the population diagnosed with diabetes using data from 2016 collected by the Center for Disease and Control (Centers for Disease Control and Prevention 2020c). Our measure of the population fraction living with hypertension was taken from 2009 data collected by the Institute for Heath Metrics and Evaluation (**?**). Finally, we measure air pollution as exposure to particulates smaller than 2.5 micro meters using raw data from (Van Donkelaar et al. 2019), with county-level exposure percentages calculated by (Wu et al. 2020).

S1.2.14 Public transportation

We measured the fraction of the population that regularly take public transport to work using data from the U.S. Census Bureau (U.S. Census 2020a). Specifically, the fraction of workers over the age of 16 years that mainly take public transport (excluding taxis) to work.

S1.2.15 Facebook Connectedness Index

The Social Connectedness Index supplied by Facebook (Bailey et al. 2018) provides a matrix of social connectedness between all U.S. counties, as well as between all U.S. counties and other nations. We use the $\boldsymbol{SCI}$ matrix representing pairwise connectivity between all 3,136 U.S. counties, in terms of Facebook friendships. The Facebook data also include links between each U.S. county and other countries. Among the countries we chose three that were epicentres of the global COVID-19 outbreak: China, Italy, and Iran. The vector $\vec{\kappa}_1$ quantifies the importance of Facebook connectivity in predicting number of COVID-19 deaths/cases.

To derive this index, Facebook maps its users to their respective U.S. counties and other countries by counting the number of friendship links between the individuals in each county pair and between each county and other countries (Bailey et al. 2018). The location of an individual is assigned based on information provided by users and on user activity. Only active friendship ties are included in the data, defined as having an interaction in the last 30 days. Derived using Facebook data from April 2016, the SCI is normalized such that it has a maximal value of one million (assigned to Los Angeles County). Facebook ties tend to represent real-world personal relationships (Jones et al. 2013), and over half the U.S. adult population use Facebook across income, education and racial categories (Duggan et al. 2015).

S2 Negative binomial regression

If the dependent variable $D_t(T)$ counts the number of deaths during a specified number of days, T, then the observed rate D_t/T can be modeled by using a negative binomial model for rate data (Zwilling 2013). In this model

$$\ln D_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \ln t \tag{2}$$

$$ln(\vec{D_t}) = \vec{\beta}\boldsymbol{X} + \ln t \tag{3}$$

where the predictor variables $x_1, x_2, ...x_p$ the last term, are given, and the population regression coefficients $\beta_1, \beta_2, ...\beta_p$ are to be estimated. The last term, $\ln t$, acts as an offset. Negative binomial regression with this offset (i.e. using rate data) predicts a log-likelihood function $L(\alpha, \beta)$ as follows:

$$\ln L(\alpha, \beta = \sum_{i=1}^{n} \left(y_i \ln \alpha + y_i(\beta x_i + \ln t_i) - (y_i + \frac{1}{\alpha}) \ln \left(1 + \alpha e^{\beta x_i + \ln t_i}\right) - \ln \Gamma(y_i + 1) - \ln \Gamma \left(\frac{1}{\alpha}\right)\right) \tag{4}$$

where $\mu > 0$ is the mean of Y and $\alpha > 0$ is the heterogeneity parameter (Hilbe 2011; Zwilling 2013). In order to estimate the parameters, the regression maximizes this log-likelihood function.

S2.1 Full LASSO regressions results

In this section, we present the full LASSO regression results. This includes the coefficient estimates at various values for the regularization parameter (λ = 0.03, 0.05, 0.1, 0.2, 0.3). This regression only includes the variable for COVID-19 deaths, not COVID-19 cases. This means, by definition, these results are the same regardless of the time delay because all dependent variables are constant.

Table S1: Dependent variable: COVID-19 deaths on April 17th

	0.03	0.05	0.1	0.2	0.3
voteshare_diff	0.00	0.00	0.18	0.00	0.00
log_population	0.94	0.92	0.87	0.99	0.96
log_ICU_beds	0.00	0.00	0.00	0.00	0.00
log_Hospital	0.00	0.00	0.00	0.00	0.00
prop_uninsured	-0.00	0.00	0.00	-0.08	-0.07
obesity	0.00	0.00	0.00	0.00	0.00
log_income	0.26	0.03	0.00	0.42	0.34
mean_age	0.00	0.00	0.00	0.07	0.06
log_density_pop	0.14	0.15	0.14	0.31	0.31
black	2.53	1.97	1.07	2.12	1.89
hispanic	0.00	0.00	0.00	-0.02	0.00
native	0.00	0.00	0.00	0.00	0.00
log_SCI_CN	0.00	0.00	0.00	-0.31	-0.25
log_SCI_IT	0.00	0.00	0.00	0.78	0.71
log_SCI_IR	0.00	0.00	0.00	-0.43	-0.42
public_transport	4.77	4.77	5.37	2.01	1.86
persons_per_house	0.00	0.00	0.00	0.87	0.84
persons_per_bedroom	0.00	0.00	0.00	0.00	0.00
persons_per_room	0.00	0.00	0.00	0.00	0.00
job_proffesion	0.00	0.00	0.00	0.00	0.00
job_service	0.00	0.00	0.00	0.00	0.00
job_office	0.00	0.00	0.00	0.00	0.00
job_trade	0.00	0.00	0.00	0.00	0.00
job_transport	0.00	0.00	0.00	1.13	0.00
hypertension	0.00	0.00	0.00	0.03	0.00
pm25	0.00	0.00	0.00	0.00	0.00
diabetes	0.00	0.00	0.00	0.00	0.00
age_over65	0.00	0.00	0.00	0.00	0.00
bachelors	0.00	0.00	0.00	0.00	0.00
no_highschool	0.00	0.00	0.00	0.00	0.00
incarce	0.00	0.00	0.00	-9.42	-5.60

Table S2: Dependent variable: COVID-19 deaths on July 1st

	0.03	0.05	0.1	0.2	0.3
voteshare_diff	0.00	0.07	0.05	0.80	-0.10
log_population	0.99	0.97	0.93	0.89	1.05
log_ICU_beds	0.00	0.00	0.00	0.00	-0.02
log_Hospital	0.00	0.00	0.00	0.00	0.00
prop_uninsured	-0.00	0.00	0.00	0.00	-0.08
obesity	0.00	0.00	0.00	0.00	0.02
log_income	0.14	0.00	0.00	0.00	0.15
mean_age	0.00	0.00	0.00	0.00	0.07
log_density_pop	0.12	0.13	0.13	0.11	0.24
black	2.68	2.36	1.96	0.60	2.08
hispanic	0.00	0.00	0.00	0.00	0.00
native	1.38	1.74	0.00	0.00	4.48
log_SCI_CN	0.00	0.00	0.00	0.00	-0.23
log_SCI_IT	0.00	0.00	0.00	0.00	0.69
log_SCI_IR	0.00	0.00	0.00	0.00	-0.46
public_transport	3.08	2.52	1.36	8.10	0.81
persons_per_house	0.05	0.10	0.00	0.00	0.75
persons_per_bedroom	0.00	0.04	0.00	0.00	0.34
persons_per_room	0.00	0.00	0.00	0.00	-3.25
job_proffesion	0.00	0.00	0.00	0.00	0.00
job_service	0.00	0.00	0.00	0.00	0.00
job_office	0.00	0.00	0.00	0.00	0.09
job_trade	0.00	0.00	0.00	0.00	-1.10
job_transport	1.28	0.00	0.00	0.00	5.10
hypertension	0.00	0.00	0.00	0.00	0.00
pm25	0.06	0.05	0.00	0.00	0.00
diabetes	0.00	0.00	0.00	0.00	-2.91
age_over65	0.00	0.00	0.00	0.00	0.00
bachelors	0.00	0.00	0.00	0.00	0.03
no_highschool	0.00	0.00	0.00	0.00	0.07
incarce	0.00	0.00	0.00	0.00	-13.26

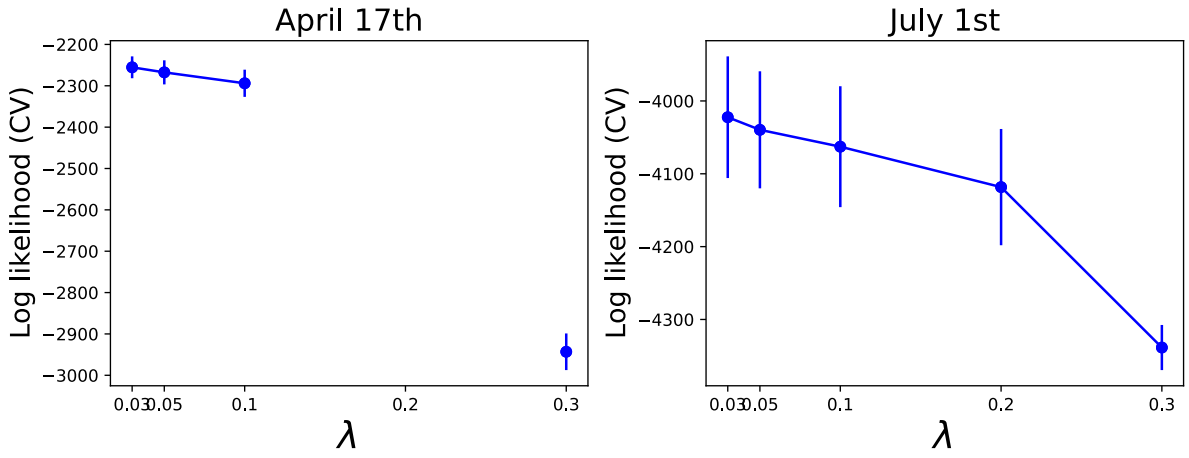

Figure S3: Cross validated (2 fold) log likelihood of Negative BinomialLASSO regression for different regularizing parameter (λ). The right hand side has COVID-19 deaths on April 17th as the depdenent variable and the right hand side is COVID-19 deaths on July 1st. The error bars are standard errors. We evaluate $\lambda = 0.01, 0.05, 0.1, 0.2$ and 0.3 in each case, but could not get an estimate for $\lambda = 0.2$ for April 17th.

S2.2 Effect of using 0, 7 and 14 day time lags between confirmed COVID-19 case and death

When we calculate the proposed under reporting of COVID-19 cases, we need to compare the number of reported cases against reported deaths. We need to factor into our calculation that there is a delay between when an individual is confirmed to have COVID-19 and subsequent death. There is no consensus on the size of this delay, therefore we run a sensitivity analysis on our results using time delays of 0, 7, 10 and 14 days. The negative binomial regression results in tables S3—S6 shows that more under reporting in April predicted more COVID-19 deaths in July. This was true in counties that had 0, 1, and 2 confirmed cases in April and regardless of whether we assume a delay of 0, 7, 10, and 14 days between infection and death. These significant relationships are illustrated in figures S4—S6 and figure 6b in the main text.

Table S3: Negative binomial regression results testing if under reporting of COVID-19 in April predicted higher COVID-19 deaths in July (assuming a delay of 0 days between detection and death). We report the results for U.S counties that had 0, 1 and 2 confirmed COVID-19 cases in April. Each entry represents the regression coefficient estimate with standard errors in brackets.

	0 cases in April	1 case in April	2 cases in April
intercept	-5.11 (0.64)	-2.75 (0.73)	-3.42 (0.86)
slope	2.68 (0.45)	1.14 (0.48)	2.43 (0.64)

Table S4: Negative binomial regression results testing if under reporting of COVID-19 in April predicted higher COVID-19 deaths in July (assuming a delay of 7 days between detection and death). We report the results for U.S counties that had 0, 1 and 2 confirmed COVID-19 cases in April. Each entry represents the regression coefficient estimate with standard errors in brackets.

	0 cases in April	1 case in April	2 cases in April
intercept	-6.29 (0.67)	-3.36 (0.61)	-4.22 (0.75)
slope	3.68 (0.45)	1.85 (0.38)	2.95 (0.52)

Table S5: Negative binomial regression results testing if under reporting of COVID-19 in April predicted higher COVID-19 deaths in July (assuming a delay of 10 days between detection and death). We report the results for U.S counties that had 0, 1 and 2 confirmed COVID-19 cases in April. Each entry represents the regression coefficient estimate with standard errors in brackets.

	0 cases in April	1 case in April	2 cases in April
intercept	-6.12 (0.61)	-3.77 (0.54)	-3.4 (0.66)
slope	3.56 (0.4)	2.23 (0.32)	2.67 (0.44)

Table S6: Negative binomial regression results testing if under reporting of COVID-19 in April predicted higher COVID-19 deaths in July (assuming a delay of 14 days between detection and death). We report the results for U.S counties that had 0, 1 and 2 confirmed COVID-19 cases in April. Each entry represents the regression coefficient estimate with standard errors in brackets.

	0 cases in April	1 case in April	2 cases in April
intercept	-5.43 (0.46)	-3.78 (0.48)	-2.66 (0.52)
slope	3.22 (0.29)	2.39 (0.27)	2.26 (0.33)

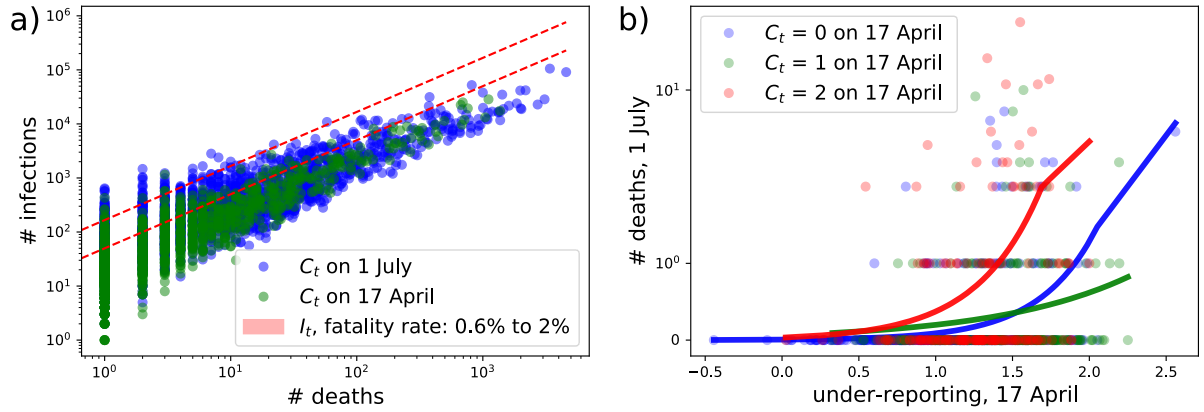

Figure S4: Robustness of under-reporting statistic (0 day time delay between confirmed case and death). Under-reporting on 17 April, 2020 vs deaths on July 1, 2020; where green circles are counties where $C_t = 0$ on 17 April (n= 414), blue circles where $C_t = 1$ (n= 254) and red circles where $C_t = 2$ (n= 173). Solid lines are best fits from a negative binomial regression.

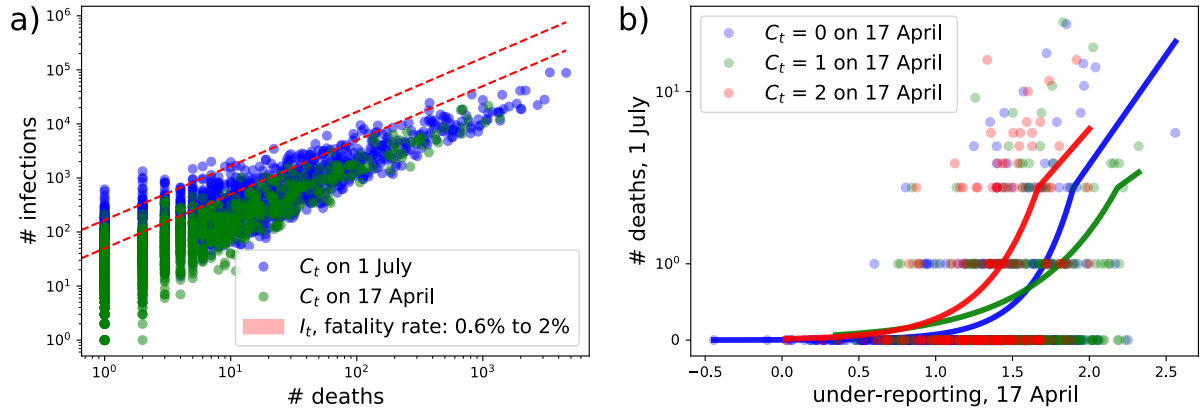

Figure S5: Robustness of under-reporting statistic (7 day time delay between confirmed case and death). Under-reporting on 17 April, 2020 vs deaths on July 1, 2020; where green circles are counties where $C_t = 0$ on 17 April (n= 526), blue circles where $C_t = 1$ (n= 309) and red circles where $C_t = 2$ (n= 185). Solid lines are best fits from a negative binomial regression.

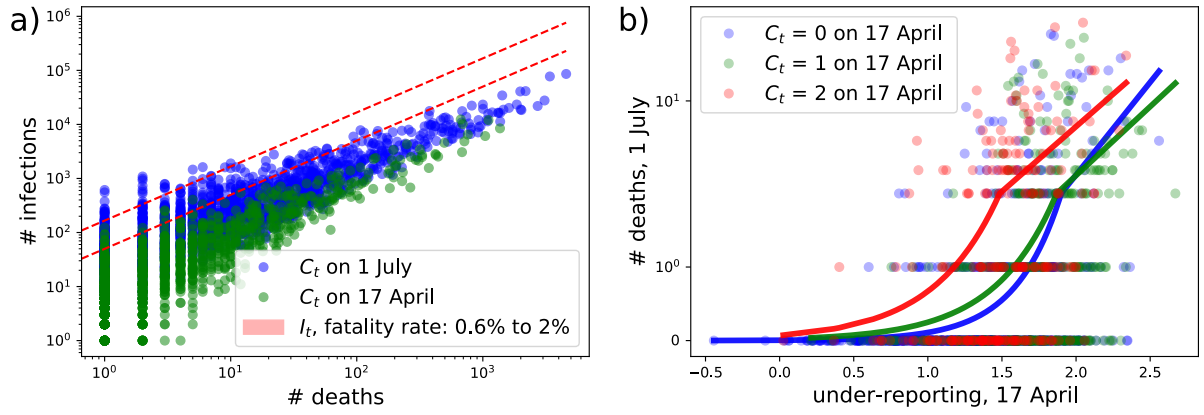

Figure S6: Robustness of under-reporting statistic (14 day time delay between confirmed case and death). Under-reporting on 17 April, 2020 vs deaths on July 1, 2020; where green circles are counties where $C_t = 0$ on 17 April (n= 784), blue circles where $C_t = 1$ (n= 402) and red circles where $C_t = 2$ (n= 237). Solid lines are best fits from a negative binomial regression.

References

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., Wong, A., 2018. Social connectedness: Measurement, determinants, and effects. J. Econ. Persp. 32(3), 259–280.

Centers for Disease Control & Prevention, 2020. Data and statistics, overweight and obesity. http://www.cdc.gov/obesity.

Centers for Disease Control and Prevention (2020c). Diabetes Data and statistics. http://www.cdc.gov/diabetes/data.

Duggan, M., Ellison, N.B., Lampe, C., Lenhart, A., Madden, M., 2015. Frequency of social media use. PEW Research Center, January 9 (2015). https://www.pewresearch.org/internet/2015/01/09/frequency-of-social-media-use-2/

Hilbe, J., 2011. Negative Binomial Regression. Cambridge University Press.

Jones, J.J., Settle, J.E., Bond, R.M., Fariss, C.J., Marlow, C., Fowler, J.H., 2013. Inferring tie strength from online directed behavior. PLoS ONE 8(1), e52168.

Katz, J., Sanger-Katz, M., Quealy, K., 2020. A detailed map of who is wearing masks in the U.S. The New York Times, July 17, 2020.

Schulte, F., Lucas, E., Rau, J., Szabo, L., Hancock, J., 2020. Millions of older Americans Live in counties with no ICU beds as pandemic intensifies. Kaiser Health News, March 20 (2020). Data: https://khn.org/wp-content/uploads/sites/2/2020/03/KHN-ICU-bed-county-analysis_2.zip

U.S. Census Bureau. Small Area Health Insurance Estimates. https://www.census.gov/content/dam/Census/library/publications/2019/demo/p30-05.pdf

U.S. Census Bureau. Small Area Income and Poverty Estimates. https://www.census.gov/programs-surveys/saipe/about.html

U.S. Census Bureau. Population Estimates Program. https://www.census.gov/programs-surveys/popest/about.html

USAFacts.org. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map(2020).

Van Donkelaar, A., Martin, R.V., Li, C., Burnett, R.T., 2019. Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. Enviro. Sci. Tech. 53(5), 2595–2611.

Vera Institute of Justice, 2020. Incarceration Trends Dataset. https://github.com/vera-institute/incarceration_trends.

Wu, X., Nethery, R. C., Sabath, B. M., Braun, D., Dominici, F., 2020. Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study. medRxiv: 2020.04.05.20054502.

Zwilling, M.L., 2013. Negative binomial Regression. The Mathematica Journal. dx.doi.org/10.3888/tmj.15-6.