Supplementary Information for

Using Machine Learning to Estimate the Effect of Racial Segregation on COVID-19 Mortality in the United States

This PDF includes the following sections:

- Section 1: Data description
- Section 2: Supplementary information results
- Section 3: Results using Theil's Information Theory Index to measure segregation
- Section 4: Sensitivity analysis
- Section 5: Results using equal samples for the racial/ethnic gap analysis
- References for the supplementary information

This PDF includes the following figures and tables:

- Figures S1 to S10
- Tables S1 to S4

Section 1 Data Description

Section 1.1 Descriptive Statistics

Table S1: Descriptive statistics for sample of 2,174 counties

	Mean	SD	Min	Median	Max
COVID-19 Outcomes					
Deaths per 100,000	49.20	53.25	0.00	32.71	503.81
Log deaths per 100,000	3.41	1.21	0.00	3.57	6.25
Cases per 100,000	2,095.73	1,487.96	50.48	1,814.28	17,507.57
Log cases per 100,000	7.40	0.76	3.98	7.50	9.77
Racial segregation					
Multi-group Relative Diversity Index	0.08	0.08	0.00	0.06	0.57
Black-White Relative Diversity Index	0.10	0.12	0.00	0.05	0.75
Hispanic-White Relative Diversity Index	0.07	0.08	0.00	0.04	0.68
Multi-group Theil Segregation Index	0.11	0.06	0.00	0.10	0.52
Black-White Theil Segregation Index	0.15	0.10	0.00	0.13	0.73
Hispanic-White Theil Segregation Index	0.10	0.07	0.00	0.09	0.57
Demographics					
% Asian	1.58	2.67	0.00	0.75	35.65
% Black	10.65	14.87	0.00	3.84	85.90
% Hispanic	9.41	13.28	0.00	4.46	99.07
% White	75.10	19.46	0.73	81.17	99.44
% no high school	13.51	5.94	2.02	12.35	48.52
% college or more	22.65	10.08	5.38	20.07	74.56
Median income (in 1,000s)	52.43	14.33	20.19	50.24	136.27
% in poverty	15.86	6.06	3.46	15.13	49.72
Income segregation (Theil Index)	0.04	0.02	0.00	0.04	0.11
% younger than 25	31.55	4.42	10.47	31.30	54.69
% older 65	17.53	4.18	3.80	17.28	55.60
Age segregation (Theil Index)	0.03	0.02	0.00	0.02	0.19
Density and public interaction					
Population density (log)	4.49	1.40	0.57	4.29	11.19
Average commute (in minutes)	24.47	5.08	11.00	24.00	45.00
% public transit	1.18	3.57	0.00	0.44	61.92
% working from home	4.34	2.21	0.13	4.01	32.60
% unemployed	3.43	1.11	0.20	3.32	10.18
% vacant units	16.52	9.57	3.01	14.32	70.37
% households with 6+ occupants	3.31	1.59	0.14	3.02	15.97
% units in 50+ unit buildings	1.65	2.86	0.00	0.85	55.87
% families with grandchildren present	2.62	1.31	0.24	2.38	12.69
Domestic airport passengers per 1,000 (log)	13.47	1.85	0.00	13.62	16.42
International airport passengers per 1,000 (log)	6.97	5.44	0.00	7.71	15.37
Social capital (per 100,000)					
Civic organizations	0.09	0.10	0.00	0.06	0.67
Religious organizations	0.86	0.36	0.00	0.82	3.22
Sports and bowling centers	0.02	0.06	0.00	0.00	0.36
Health risk factors					
Life expectancy	77.32	2.77	67.07	77.35	97.97
% premature deaths	0.41	0.11	0.13	0.40	0.83
% diabetic	11.75	2.66	3.30	11.70	20.90
% HIV positive	0.18	0.20	0.01	0.12	2.31
% obese	32.10	4.76	13.60	32.40	49.50
% smokers	18.07	3.45	6.74	17.97	33.17
% excessive drinking	17.44	3.29	9.27	17.38	29.44
% physically inactive	25.65	5.41	8.40	25.80	45.10
% sleep less 7h	33.95	3.76	23.03	33.99	46.71
Health system capacity					
Primary care physicians per 100,000	56.98	32.42	2.17	51.21	448.23
Primary care providers per 100,000	77.87	56.23	1.95	67.95	1,433.89
Hospital beds per 1,000	2.83	3.59	0.00	2.18	95.10

% insured	90.28	4.48	60.83	90.94	98.00
% flu vaccine	42.93	8.01	12.00	44.00	65.00
Air pollution					
PM2.5 daily average	9.54	1.72	3.00	9.80	19.70
Employment in essential businesses (in %)					
Food stores	0.01	0.00	0.00	0.01	0.03
Hospitals	0.00	0.01	0.00	0.00	0.11
Nursing homes	0.01	0.01	0.00	0.01	0.04
Pharmacies	0.00	0.00	0.00	0.00	0.01
Public sector	15.79	5.27	5.50	14.74	62.25
Construction	7.23	2.18	1.97	6.95	20.42
Political views					
% voted democrat in 2016	34.70	14.99	7.38	31.60	89.33
% thinks global warming is happening	60.92	6.38	45.61	60.10	82.96
% supports CO2 regulation	68.83	3.99	59.07	68.40	81.83

Section 1.2 Data Sources

Table S2: Data sources and variable construction

Variable	Description	Source	Year Measured
COVID-19 Outcomes			
Log death rate and log case rate	Log of total deaths and cases in the county per 100,000 residents	*USA Facts* and *The New York Times*	September 30, 2020
Log death rate for blacks, Hispanics, and whites	Log of total deaths among each group in the county per 100,000 individuals of the corresponding group	Centers for Disease Control and Prevention	September 30, 2020
Racial Segregation			
Relative Diversity Index and Theil's Information Theory Index	Multi-group indices computed using counts of individuals in the following groups: non-Hispanic Asian, non-Hispanic black, Hispanic, non-Hispanic white, and non-Hispanic of other race. The black-white (Hispanic-white) index is computed using counts of black (Hispanic) and non-Hispanic white individuals only.	Author's calculations from tract-level data from the American Community Survey, 5-Year Estimates (See Section 2.1 and Section 3 for the calculations of both indices)	2014-2018
Demographics			
% non-Hispanic Asian, % non-Hispanic black, % Hispanic, and % non-Hispanic white	Percentage of each racial group in the population.	American Community Survey, 5-Year Estimates	2014-2018
% no high-school	Percentage of adults 25 years old and older without a high-school diploma	American Community Survey, 5-Year Estimates	2014-2018
% college or more	Percentage of adults 25 years old and older with a college degree or higher	American Community Survey, 5-Year Estimates	2014-2018
Median income	Median household income in the county	American Community Survey, 5-Year Estimates	2014-2018
% in poverty	Percentage of all residents living below the poverty line	American Community Survey, 5-Year Estimates	2014-2018

Income segregation	Theil index of income segregation computed using counts of households in the following income bins: less than $10,000, $10,000 to $14,999, $15,000 to $19,999, $20,000 to $24,999, $25,000 to $29,999, $30,000 to $34,999, $35,000 to $39,999, $40,000 to $44,999, $45,000 to $49,999, $50,000 to $59,999, $60,000 to $74,999, $75,000 to $99,999, $100,000 to $124,999, $125,000 to $149,999, $150,000 to $199,999, and $200,000 or more	Author's calculations from tract-level data from American Community Survey, 5-Year Estimates	2014-2018
% younger than 25	Percentage of the population younger than 25	American Community Survey, 5-Year Estimates	2014-2018
% older than 65	Percentage of the population older than 65	American Community Survey, 5-Year Estimates	2014-2018
Age segregation	Theil index of age segregation computed using counts of individuals in the following age bins: 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 to 74, 75 to 84, and 85 and older	Author's calculations from tract-level data from American Community Survey, 5-Year Estimates	2014-2018
Density and Public Interaction			
Population density	Population per square mile	American Community Survey, 5-Year Estimates	2014-2018
Average commute	Mean travel time to work (in minutes)	American Community Survey, 5-Year Estimates	2014-2018
% public transit	Percentage of workers 16 years and over that take public transportation to commute	American Community Survey, 5-Year Estimates	2014-2018
% working from home	Percentage of workers 16 years and over that work from home	American Community Survey, 5-Year Estimates	2014-2018
% unemployed	Percentage of the civilian labor force that is unemployed	American Community Survey, 5-Year Estimates	2014-2018
% vacant units	Percentage of vacant housing units	American Community Survey, 5-Year Estimates	2014-2018
% households with 6+ occupants	Percentage of occupied housing units with 6 or more occupants	American Community Survey, 5-Year Estimates	2014-2018
% units in 50+ unit buildings	Percentage of housing units in structures with 50 or more housing units (this is a proxy for the share of apartments in the county that are in high-rise buildings).	American Community Survey, 5-Year Estimates	2014-2018
% living with grandchildren	Percentage of the population in family households where grandchildren are present	American Community Survey, 5-Year Estimates	2014-2018

Domestic airport traffic	Rate of passengers from domestic flights landing in any airport inside the county or within 100 miles of the county boundary in the first quarter of 2020. Rates are in passengers per 1,000 county residents.	Author's calculations from Air Carrier Statistics data, T-100 Domestic Market (All Carriers), US Bureau of Transportation Statistics	January, 2020 to March, 2020
International airport traffic	Rate of passengers from international flights landing in any airport inside the county or within 100 miles of the county boundary in the first quarter of 2020. Rates are in passengers per 1,000 county residents.	Author's calculations from Air Carrier Statistics data, T-100 International Market (All Carriers), US Bureau of Transportation Statistics	January, 2020 to March, 2020
Social Capital			
Civic organizations	Establishments per 100,000 residents with 8134– NAICS codes	County Business Patterns, US Census Bureau	2017
Religious organizations	Establishments per 100,000 residents with 8131– NAICS codes	County Business Patterns, US Census Bureau	2017
Sports and bowling centers	Establishments per 100,000 residents with 71394- and 71395- NAICS codes	County Business Patterns, US Census Bureau	2017
Health Risk Factors			
Life expectancy	Number of years that the average county resident is expected to live	Mortality Files, National Center of Health Statistics†	2016-2018
% premature deaths	Percentage of deaths among residents under age 75 (age-adjusted)	Mortality Files, National Center of Health Statistics†	2016-2018
% diabetic	Percentage of adults aged 20 and over diagnosed with diabetes	United States Diabetes Surveillance System†	2016
% HIV positive	Percentage of residents ages 13 and over diagnosed with HIV	National Center for HIV/AIDS, Viral Hepatitis, STD, and TB prevention†	2016
% smokers	Percentage of adults who are smokers	Behavioral Risk Factor Surveillance System†	2016
% excessive drinking	Percentage of adults reporting binge or heavy drinking	Behavioral Risk Factor Surveillance System†	2016
% sleep less than 7h	Percentage of adults reporting less than 7 hours of sleep on average	Behavioral Risk Factor Surveillance System†	2016
% physically inactive	Percentage of adults reporting no physical activity during leisure time	United States Diabetes Surveillance System†	2016
Health System Capacity			
Primary care physicians	Primary care physicians per 100,000 residents	Area Health File, American Medical Association†	2016

Primary care providers	Primary care providers per 100,000 residents	National Provider Identification, Centers for Medicare and Medical Services†	2016
Hospital beds	Hospital beds per 1,000 residents	Homeland Infrastructure Foundation-Level Data, US Department of Homeland Security	2018
% insured	Percentage of the population with health insurance coverage	American Community Survey, 5-Year Estimates	2014-2018
% flu vaccine	Percentage of medicare enrollees who received a flu vaccine	Mapping Medicare Disparities Tool Centers for Medicare and Medical Services†	2016
Air Pollution			
PM2.5 average	Average daily density of fine particulate matter in micrograms per cubic meter	National Environmental Public Health Tracking Network†	2014
Employment in Essential Businesses			
Food stores	Percentage of the population employed in establishments with NAICS codes 445—	County Business Patterns, US Census Bureau	2017
Hospitals	Percentage of the population employed in establishments with NAICS codes 622—	County Business Patterns, US Census Bureau	2017
Nursing homes	Percentage of the population employed in establishments with NAICS codes 6231– and 6233–	County Business Patterns, US Census Bureau	2017
Pharmacies	Percentage of the population employed in establishments with NAICS codes 44611-	County Business Patterns, US Census Bureau	2017
Public sector	Percentage of employed civilian population 16 years and over employed in the public sector	American Community Survey, 5-Year Estimates	2014-2018
Construction	Percentage of employed civilian population 16 years and over employed in construction	American Community Survey, 5-Year Estimates	2014-2018
Political Views			
% voted democrat in 2016	Percentage of voters who voted for Hillary Clinton in the 2016 Presidential Election	MIT Election Data	2016
% thinks global warming is happening	Percentage of the population who thinks that global warming is happening	Yale Climate Opinion Maps	2019
% supports CO2 regulation	Percentage of the population who is in favor of CO2 emissions regulation	Yale Climate Opinion Maps	2019

†These data were collected by the Robert Wood Johnson's County Health Rankings & Roadmaps program.

Section 2 Supplementary Information Results

Section 2.1 Measuring Segregation with the Relative Diversity Index

To measure racial residential segregation in the county, I use data from the 2014-2018 American Community Survey on census tract counts of non-Hispanic Asians, non-Hispanic blacks, non-Hispanic whites, non-Hispanics of other racial groups, and Hispanics to estimate the Relative Diversity Index, a metric capturing the ratio of within-tract diversity to total diversity in the county (1).

The Relative Diversity Index is computed in two steps. In the first step we obtain the Simpson's Interaction Index, and in the second step we use that index to calculate the Relative Diversity Index. The Simpson's Interaction Index is calculated for each of the small units that we choose as our definition of neighborhoods—census tracts in this analysis—and for the larger area for which we want to measure segregation—counties. With M racial and ethnic groups, the Simpson's Interaction Index for tract i, I_i, is calculated as follows:[1]

$$I_i = \sum_{m=1}^{M} p_{mi}(1 - p_{mi}),$$

where m indexes the racial and ethnic groups into which the population is divided, and p_{mi} is the proportion of members of group m in tract i. Similarly, the Simpson's Interaction Index for county c, I_c, is computed by using p_{mc} as the proportion of members of group m in the county:

$$I_c = \sum_{m=1}^{M} p_{mc}(1 - p_{mc}),$$

The Relative Diversity Index of segregation for county c, R_c, is a weighted average deviation of each tract's Simpson's Interaction Index from the county's Simpson's Interaction Index, and it is computed as follows:

$$R_c = \frac{1}{T_c I_c} \sum_{i=1}^{n} t_i(I_c - I_i), \tag{1}$$

where T_c is the total population in county c, I_c is the Simpson's Interaction Index of county c, I_i is the Simpson's Interaction Index of tract i, and t_i is the total population in tract i.

The Relative Diversity Index can be interpreted as one minus the ratio of the probability that two individuals from the same tract are members of different racial/ethnic groups to the probability that any two individuals are members of different groups (1). The index can take values from 0 to 1, with 0 indicating that all tracts have the exact same diversity than the county as a whole (i.e., the shares of each racial group are the same across all tracts in the county), and

[1] For the multi-group Relative Diversity Index, the M groups are non-Hispanic Asian, non-Hispanic black, Hispanic, non-Hispanic white, and non-Hispanic of other race. For the black-white (Hispanic-white) Relative Diversity Index, the M groups are black and non-Hispanic white (Hispanic and non-Hispanic white).

1 representing a county where tracts have no diversity (e.g., one set of tracts includes all black residents and no one else, another set of tracts includes all Hispanic residents and no one else, and so on).

Fig S1 shows the bivariate associations between the 50 controls and the Relative Diversity Index, net of state fixed effects. Section 3 in this Supplementary Information appendix shows results using the Theil Information Theory Index to measure racial segregation.

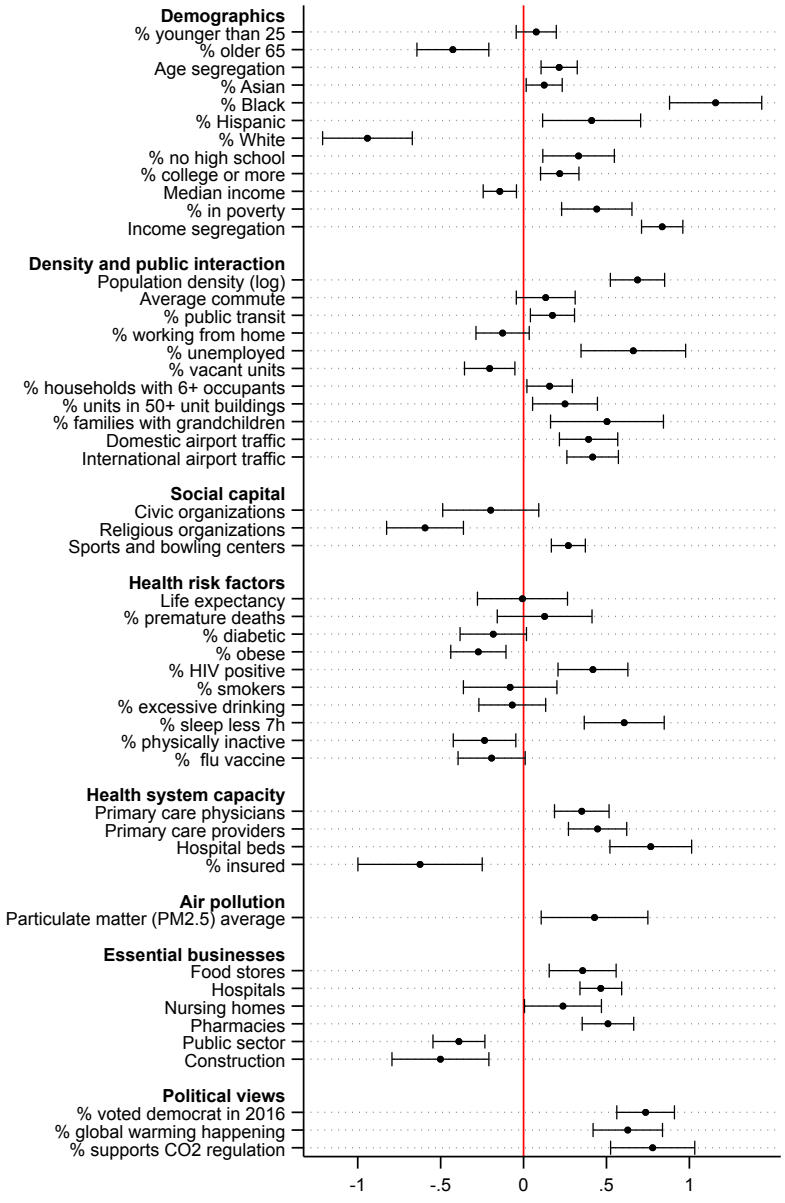

Figure S1 Standardized bivariate associations between the multi-group Relative Diversity Index and county attributes
Racial segregation is measured as the multi-group Relative Diversity Index using census tract data from the 2014-2018 American Community Survey. The associations are estimated via OLS regression with state fixed effects, population weights, and standard errors clustered by state. Bars around estimated bivariate associations reflect 95% confidence intervals. The outcome and covariates have been standardized to have mean 0 and SD 1.

Section 2.2 Full Lasso Regression Output

Table S3: Double-lasso regression output from Fig. 3 and when adding additional controls

	Log death rate			Log case rate		
	(1) Fig. 3	(2)	(3)	(4) Fig. 3	(5)	(6)
Multi-Group Relative Diversity Index	0.082***	0.065**	0.085**	0.054**	0.043**	0.051*
	(0.030)	(0.031)	(0.037)	(0.021)	(0.021)	(0.026)
% Asian		-0.104***	-0.060*		-0.063***	-0.047**
		(0.025)	(0.031)		(0.015)	(0.021)
% Black		-0.142	-0.044		-0.094	-0.053
		(0.115)	(0.152)		(0.062)	(0.082)
% Hispanic	-0.014	-0.202*	-0.110	0.054*	-0.064	-0.011
	(0.067)	(0.111)	(0.152)	(0.029)	(0.054)	(0.064)
% White	-0.303***	-0.602***	-0.368*	-0.059*	-0.247***	-0.140
	(0.044)	(0.140)	(0.201)	(0.034)	(0.076)	(0.109)
Sports and bowling centers	0.067***	0.068***	0.050***	0.015	0.016*	0.007
	(0.018)	(0.018)	(0.017)	(0.010)	(0.009)	(0.008)
Hospitals	-0.012	-0.011	-0.005	0.011	0.011	0.010
	(0.021)	(0.022)	(0.021)	(0.013)	(0.012)	(0.014)
Nursing homes	0.160***	0.165***	0.208***	0.042**	0.044***	0.067***
	(0.029)	(0.029)	(0.036)	(0.016)	(0.016)	(0.018)
% working from home	-0.070	-0.084*	-0.031	-0.052	-0.060**	-0.053
	(0.053)	(0.045)	(0.059)	(0.034)	(0.029)	(0.034)
% households with 6+ occupants	0.105**	0.088*	0.078	0.032	0.021	0.040
	(0.048)	(0.047)	(0.051)	(0.026)	(0.027)	(0.031)
% no high school	0.309***	0.343***	0.391***	0.212***	0.232***	0.209***
	(0.094)	(0.085)	(0.109)	(0.033)	(0.029)	(0.030)
% HIV positive	-0.059***	-0.093***	-0.069***	0.004	-0.016	-0.003
	(0.021)	(0.023)	(0.019)	(0.014)	(0.017)	(0.011)
Income Segregation	0.005	0.014	0.027	0.037*	0.043**	0.053**
	(0.042)	(0.040)	(0.056)	(0.019)	(0.018)	(0.022)
% younger than 25	-0.061	-0.092**	-0.106**	0.102***	0.083***	0.063**
	(0.038)	(0.045)	(0.043)	(0.021)	(0.025)	(0.029)
% older 65	0.128**	0.105**	0.099**	0.027	0.014	0.002
	(0.048)	(0.042)	(0.042)	(0.042)	(0.034)	(0.035)
% units in 50+ unit buildings	0.011	0.023	0.032	-0.009	-0.002	-0.003
	(0.027)	(0.029)	(0.021)	(0.020)	(0.020)	(0.013)
% public transit	0.056**	0.075**	0.051**	0.023	0.035**	0.026**
	(0.027)	(0.029)	(0.025)	(0.015)	(0.016)	(0.011)
PM2.5 daily average	0.122***	0.104***	0.088**	0.070***	0.060***	0.051***
	(0.027)	(0.028)	(0.037)	(0.017)	(0.017)	(0.018)
% flu vaccine	0.045	0.068	-0.043	0.011	0.024	-0.006
	(0.048)	(0.045)	(0.060)	(0.031)	(0.028)	(0.031)
% insured	0.052	0.062	0.034	0.006	0.013	-0.021
	(0.056)	(0.051)	(0.049)	(0.041)	(0.042)	(0.058)
Median income (in 1,000s)	0.099	0.158**	0.160*	0.119*	0.154***	0.140**
	(0.081)	(0.067)	(0.088)	(0.060)	(0.051)	(0.059)
White-Black poverty rate gap			0.132			0.005
			(0.083)			(0.049)
White-Hispanic poverty rate gap			-0.006			0.035
			(0.060)			(0.030)
White-Black median household income gap			0.017			0.016
			(0.035)			(0.023)
White-Hispanic median household income gap			-0.104**			-0.034
			(0.047)			(0.035)
White-Black unemployment rate gap			0.100			0.098*
			(0.061)			(0.052)
White-Hispanic unemployment rate gap			0.008			0.067
			(0.049)			(0.042)
White-Black life expectancy gap			-0.003			0.018

			(0.036)			(0.027)
White-Hispanic life expectancy gap			0.026			-0.012
			(0.030)			(0.025)
Counties	2,174	2,174	830	2,174	2,174	830
Adj. R^2	0.668	0.676	0.742	0.750	0.757	0.760
R^2	0.678	0.686	0.765	0.757	0.764	0.781

Standard errors are clustered by state. (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). All models include state fixed effects and population weights. Death and case rates are log transformed. Segregation and covariates have been standardized to have mean 0 and SD 1. COVID-19 outcomes are measured as of September 30, 2020. The sample in models 3 and 6 includes the 830 counties for which race-specific data on demographics and life expectancy could be obtained or reliably estimated. COVID-19 outcomes are measured as of September 30, 2020.

Table S4: Double-lasso regression output from Fig. 4 and when adding additional controls

	Black-White gap			Hispanic-White gap		
	(1) Fig. 4	(2)	(3)	(4) Fig. 4	(5)	(6)
Black-White Relative Diversity Index	0.075***	0.076***	0.068**			
	(0.024)	(0.023)	(0.028)			
Hispanic-White Relative Diversity Index				-0.013	-0.011	0.026
				(0.050)	(0.050)	(0.035)
% Asian		0.134	0.104		0.114	0.340***
		(0.084)	(0.089)		(0.072)	(0.124)
% Black		0.621	0.495		0.206	1.457**
		(0.446)	(0.466)		(0.384)	(0.685)
% Hispanic	-0.039	0.524	0.385	0.238***	0.504	1.587***
	(0.048)	(0.392)	(0.417)	(0.079)	(0.339)	(0.603)
% White	0.297***	1.159*	0.938	0.322***	0.785	2.320***
	(0.072)	(0.600)	(0.635)	(0.116)	(0.510)	(0.863)
Sports and bowling centers	-0.012	-0.007	-0.004	-0.010	-0.001	0.006
	(0.023)	(0.023)	(0.025)	(0.025)	(0.025)	(0.024)
Hospitals	0.017	0.017	0.011	0.017	0.016	-0.016
	(0.019)	(0.019)	(0.018)	(0.031)	(0.028)	(0.027)
Nursing homes	-0.098	-0.109	-0.110	-0.184*	-0.193*	-0.242**
	(0.098)	(0.101)	(0.101)	(0.110)	(0.111)	(0.119)
% working from home	-0.001	0.004	0.014	0.077	0.092	0.107
	(0.058)	(0.059)	(0.063)	(0.085)	(0.084)	(0.091)
% households with 6+ occupants	0.040	0.040	0.050	0.173	0.183*	0.133
	(0.054)	(0.053)	(0.055)	(0.107)	(0.107)	(0.082)
% no high school	0.182*	0.160	0.173*	-0.109	-0.156	-0.239*
	(0.098)	(0.097)	(0.103)	(0.166)	(0.170)	(0.137)
% HIV positive	0.013	0.019	0.017	0.039	0.084**	0.071
	(0.024)	(0.027)	(0.029)	(0.040)	(0.041)	(0.044)
Income Segregation	-0.040	-0.037	-0.044	0.096	0.117	-0.027
	(0.048)	(0.048)	(0.057)	(0.104)	(0.102)	(0.072)
% younger than 25	-0.025	-0.003	0.007	-0.159	-0.075	-0.065
	(0.074)	(0.073)	(0.070)	(0.112)	(0.109)	(0.096)
% older 65	-0.052	-0.053	-0.048	-0.020	0.031	-0.010
	(0.057)	(0.061)	(0.062)	(0.069)	(0.066)	(0.063)
% units in 50+ unit buildings	0.036**	0.032*	0.029	0.013	0.003	-0.004
	(0.017)	(0.017)	(0.017)	(0.022)	(0.021)	(0.019)
% public transit	-0.002	-0.002	-0.003	0.027	0.012	0.010
	(0.013)	(0.014)	(0.016)	(0.019)	(0.016)	(0.016)
PM2.5 daily average	-0.029	-0.036	-0.041	-0.125**	-0.108*	-0.102**
	(0.030)	(0.033)	(0.035)	(0.058)	(0.058)	(0.049)
% flu vaccine	-0.074	-0.091	-0.076	-0.040	-0.097	-0.034
	(0.061)	(0.063)	(0.071)	(0.089)	(0.091)	(0.098)
% insured	0.108	0.097	0.098	0.053	0.021	0.026
	(0.073)	(0.072)	(0.069)	(0.135)	(0.141)	(0.123)
Median income (in 1,000s)	0.158***	0.128***	0.115	-0.015	-0.076	-0.136
	(0.041)	(0.046)	(0.073)	(0.052)	(0.054)	(0.093)
White-Black poverty rate gap			-0.073			-0.002
			(0.165)			(0.152)
White-Hispanic poverty rate gap			0.130			0.171
			(0.100)			(0.160)
White-Black median household income gap			-0.020			0.101
			(0.107)			(0.127)
White-Hispanic median household income gap			0.026			-0.049
			(0.099)			(0.121)
White-Black unemployment rate gap			0.230			-0.137
			(0.144)			(0.235)
White-Hispanic unemployment rate gap			-0.019			-0.460
			(0.129)			(0.307)
White-Black life expectancy gap			0.097			-0.017
			(0.061)			(0.057)
White-Hispanic life expectancy gap			0.001			-0.003

			(0.042)			(0.109)
Counties	243	243	239	218	218	215
Adj. R^2	0.435	0.438	0.437	0.441	0.458	0.529
R^2	0.561	0.568	0.588	0.588	0.605	0.677

Standard errors are clustered by state. (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). All models include state fixed effects and population weights. The black-white (Hispanic-white) death rate gap is the difference between the log death rate for blacks (Hispanics) and the log death rate for whites. Segregation indices and covariates have been standardized to have mean 0 and SD 1. COVID-19 outcomes are measured as of September 30, 2020.

Section 3 Results Using Theil's Information Theory Index

This section shows the counterparts to Figs. 2 to 4 when using Theil's Information Theory Index. This index, also known as the entropy index has the most appealing mathematical properties among all indices of multi-group segregation (1). The index takes into account the spatial distribution of all racial groups within the county, rather than just focusing on the extent to which a specific racial group is segregated from the rest.

The Theil index is computed in two steps. In the first step we obtain the entropy score of diversity, and in the second step we use the entropy score to calculate the Theil index. The entropy score of diversity is calculated for each of the small units that we choose as our definition of neighborhoods—census tracts in this analysis—and for the larger area for which we compute the segregation measure—counties. With M racial and ethnic groups, the entropy score of diversity for tract i, e_i, is calculated as follows:[2]

$$e_i = \sum_{m=1}^{M} p_{mi}\, ln\left(\frac{1}{p_{mi}}\right),$$

where m indexes the racial and ethnic groups into which the population is divided, and p_{mi} is the proportion of members of group m in tract i. The entropy score of for county c, E_c, is computed analogously by using p_{mc} as the proportion of members of group m in the county:

$$E_c = \sum_{m=1}^{M} p_{mc}\, ln\left(\frac{1}{p_{mc}}\right).$$

The Theil index for county c, H_c, is a weighted average deviation of each tract's entropy score from the entropy score of the county, and it is computed as follows:

$$H_c = \frac{1}{T_c E_c} \sum_{i=1}^{n} t_i (E_c - e_i), \tag{2}$$

[2]For the multi-group Theil Index, the M groups are non-Hispanic Asian, non-Hispanic black, Hispanic, non-Hispanic white, and non-Hispanic of other race. For the black-white (Hispanic-white) Theil Index, the M groups are black and non-Hispanic white (Hispanic and non-Hispanic white).

where T_c is the total population in county c, E_c is the entropy score of county c, e_i is the entropy score of tract i, and t_i is the total population in tract i.

The index can take values from 0 to 1, with 0 indicating that the shares of each racial group are the same across all tracts in the county, and 1 representing a county where all blacks live in a small set of tracts, Hispanics in another set of tracts, and so on.

By looking at equations (1) and (2), one can note the similarities between the Theil Index and the Relative Diversity Index: the Theil Index formula replaces the Simpson's Interaction Index by the entropy score. It is then not surprising to see that the two indices exhibit a correlation of .96 and an R^2 of .94, as shown in Fig. S2.

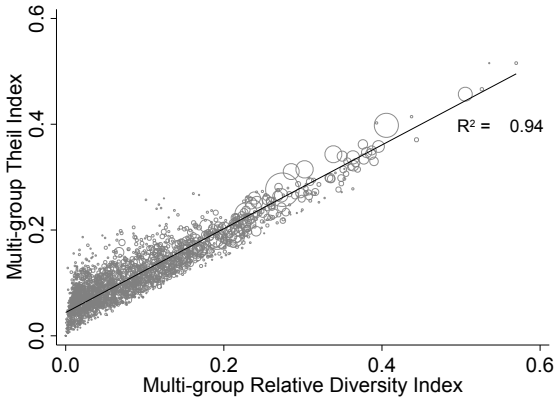

Figure S2 Correlation between the multi-group Theil and Relative Diversity Indices of segregation
Both segregation indices are generated using Census tract data from the 2014-2018 American Community Survey. The size of the dots is proportional to the county population.

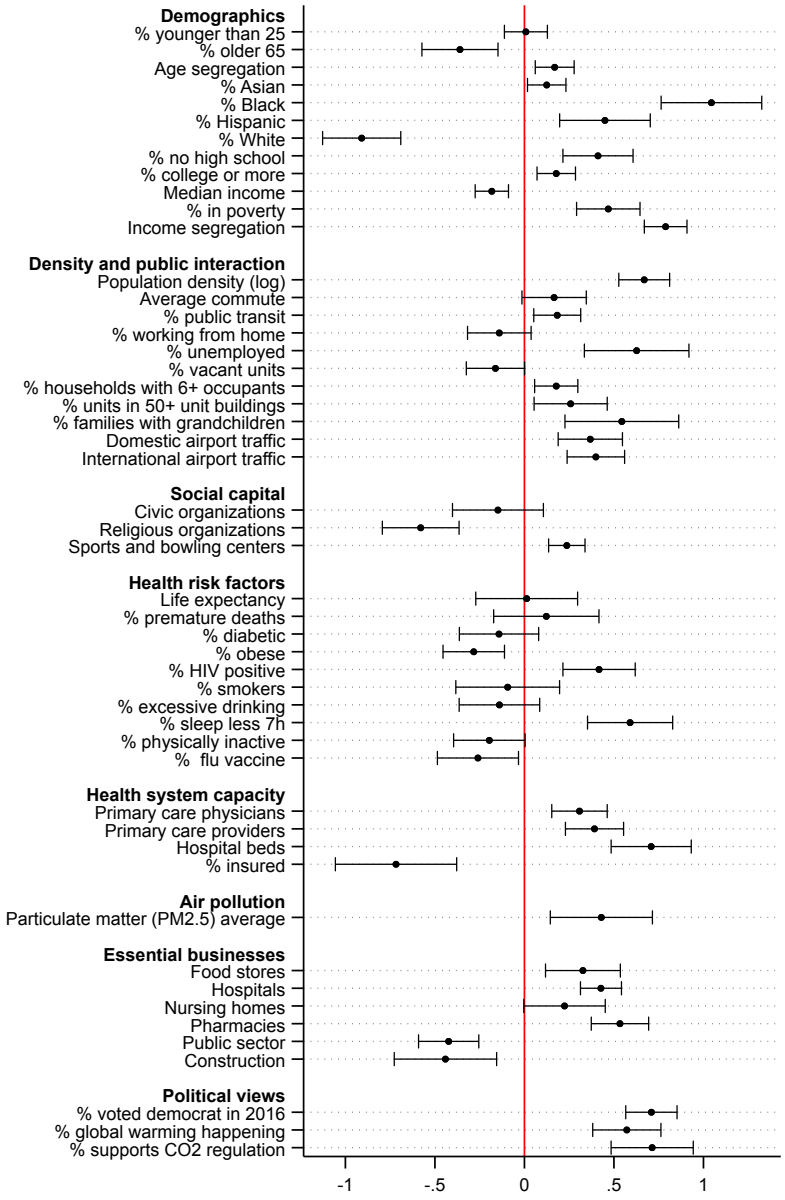

Figure S3 Standardized bivariate associations between county attributes and segregation using the multi-group Theil Information Theory Index

Racial segregation is measured as Theil's Information Theory Index using census tract data from the 2014-2018 American Community Survey. The associations are estimated via OLS regression with state fixed effects, population weights, and standard errors clustered by state. Bars around estimated bivariate associations reflect 95% confidence intervals. The outcome and covariates have been standardized to have mean 0 and SD 1.

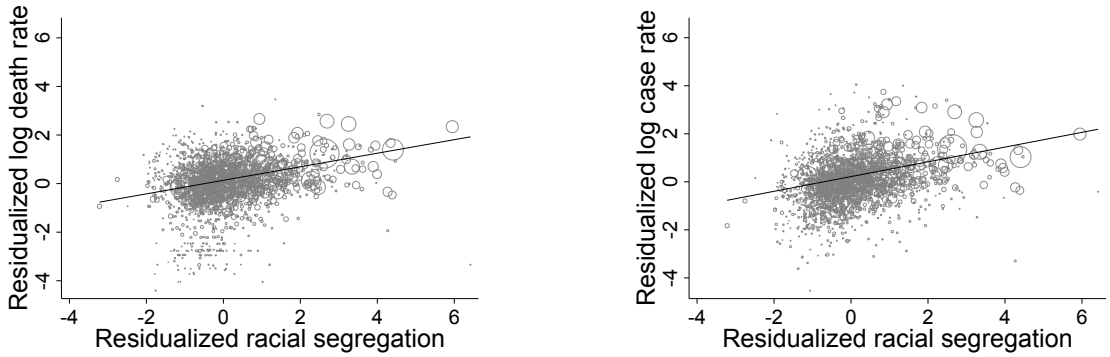

Figure S4 Relationship between racial segregation and COVID-19 mortality and infection rates net of state fixed effects using the multi-group Theil Information Theory Index

The x-axis represents the residuals from a regression of the multi-group Theil Information Theory Index on the set of state dummies. The y-axis represents the residuals from a regression of the corresponding COVID-19 outcome on the set of state dummies. The size of the dots is proportional to the county population. COVID-19 outcomes are measured as of September 30, 2020.

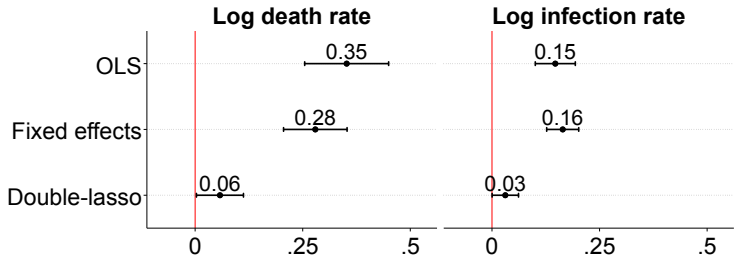

Figure S5 OLS and double-lasso regression estimates of the relationship between the multi-group Theil Information Theory Index and COVID-19 death and infection rates

OLS models include no controls. Fixed effects models include state fixed effects. Double-lasso models include the 18 controls selected by the lasso procedure (shown in Table S3) and state fixed effects. Standard errors are clustered by state. All regressions include population weights. Bars around estimated coefficients reflect 95% confidence intervals. The segregation index and covariates have been standardized to have mean 0 and SD 1. Outcomes are in log rates. The sample includes 2,174 counties. COVID-19 outcomes are measured as of September 30, 2020.

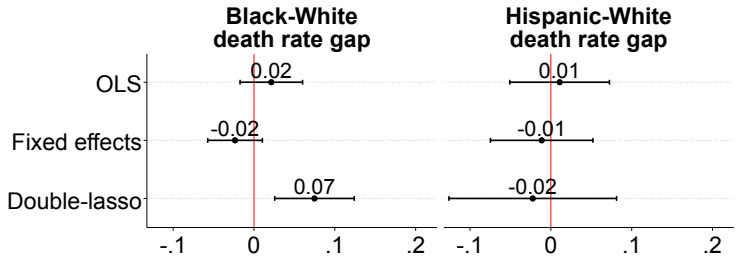

Figure S6 OLS and double-lasso regression estimates of the relationship between the black-white and Hispanic-white Theil Information Theory Indices and racial gaps in COVID-19 death rates

OLS models include no controls. Fixed effects models include state fixed effects. Double-lasso models include the 18 controls selected by the lasso procedure (shown in Table S4) and state fixed effects. Standard errors are clustered by state. All regressions include population weights. Standard errors are clustered by state. Bars around estimated coefficients reflect 95% confidence intervals. Segregation indices and covariates have been standardized to have mean 0 and SD 1. The black-white (Hispanic-white) death rate gap is the difference between the log death rate for blacks (Hispanics) and the log death rate for whites. The sample includes 243 counties for the black-white model and 218 for the Hispanic-white model. Figures S9 and S10 show results when using equal samples (N=180) across both models. COVID-19 outcomes are measured as of September 30, 2020.

Section 4 Sensitivity Analysis

In this section, I present the results of two sensitivity tests due to Frank (2) and Oster (3). Frank's test estimates the correlations that an omitted variable would have to exhibit with the outcome and the independent variable of interest such that when adding that confounder to the model, the inference would be invalidated at the 5% level (i.e., the 95% confidence interval around the point estimate of interest would include 0). I use such test to characterize a confounder's correlation with the black-white Relative Diversity Index and with the black-white mortality gap that would make the estimate from the black-white model in Fig. 4 statistically non-significant at the 5% level if such confounder was added to the set of 18 controls selected by the double-lasso procedure.

Oster's test follows a similar logic but focuses on what it would take for the magnitude of the point estimate to be zero. Given the set of 18 controls selected by the double-lasso procedure, the test will allow for an unobserved covariate that could make the results in Fig. 4 go away (i.e., the true relationship between the black-white Relative Diversity Index and the black-white COVID-19 mortality gap would be zero). The idea is to estimate two characteristics of the unobserved covariate that is potentially biasing the estimate of the segregation coefficient: the predictive power that this covariate would have on the death rate and its importance in predicting segregation, relative to the full set of covariates already included in the model.

Results from Frank's test are shown in Fig. S7. Assuming that the estimate from the black-white model in Fig. 4 is upwardly biased, the curved solid lines in the first and third quadrants represent the combinations of partial correlations of a confounder with the black-white mortality gap and the black-white Relative Diversity Index that would invalidate the inference. Any omitted variable whose partial correlations sit on the curves would make the 95% confidence interval around the point estimate in Fig. 4 cross zero. For example, the figure plots an omitted variable that has a correlation of .21 with the black-white mortality gap and a correlation of .32 with the black-white Relative Diversity Index. If such variable existed and was added to the model, the estimate from the black-white model in Fig. 4 would become statistically non-significant at the 5% level.[3]

[3]The pairs of partial correlations are derived using Frank's (2) impact threshold for a confounding variable (ITCV) formula for the multivariate regression case:

$$ITCV = k = \left(\sqrt{(1 - r^2_{x\cdot\mathbf{z}})(1 - r^2_{y\cdot\mathbf{z}})}\right)\left(\frac{t^2 + t\sqrt{d}}{-(n-q-1)} + \left(\frac{-d - t\sqrt{d}}{-(n-q-1)}\right) r_{y\cdot x|\mathbf{z}}\right), \quad (3)$$

where $r_{x\cdot\mathbf{z}}$ is the correlation between the black-white Relative Diversity Index and all other covariates in the model; $r_{y\cdot\mathbf{z}}$ is the correlation between the black-white mortality gap and the other covariates in the model (excluding the black-white Relative Diversity Index); t is the t-statistic for the coefficient on the black-white Relative Diversity Index from the double-lasso model in Fig. 4; $d = t^2 + (n - q - 1)$; n is the sample size; q is the number of covariates in the model in Fig. 4; and $r_{y\cdot x|\mathbf{z}}$ is the partial correlation between the black-white mortality gap and the black-white Relative Diversity Index. Denoting the correlation between the confounding variable and the black-white mortality gap as $r_{y\cdot cv}$ and the correlation between the confounding variable and the black-white Relative Diversity Index as $r_{x\cdot cv}$, the curved lines in Fig. S7 are defined by the correlations pairs $(r_{y\cdot cv}, r_{x\cdot cv})$ that satisfy $k = r_{y\cdot cv} \cdot r_{x\cdot cv}$.

To assess the extent to which any of the pairs of correlations that would invalidate the inference are plausible, it is useful to compare them to the partial correlations that each of the 50 covariates listed in Fig. 1 would exhibit if they were added to the model; the circles in Fig. S7 represent that.[4] None of the 50 covariates comes close to the curve that characterizes the partial correlations of a confounder that would overturn the finding from Fig. 4. In other words, an omitted variable that would invalidate the inference would have to be more strongly correlated with the black-white Relative Diversity Index and the black-white mortality gap than, for example, income segregation, poverty, or population density, which is highly implausible.

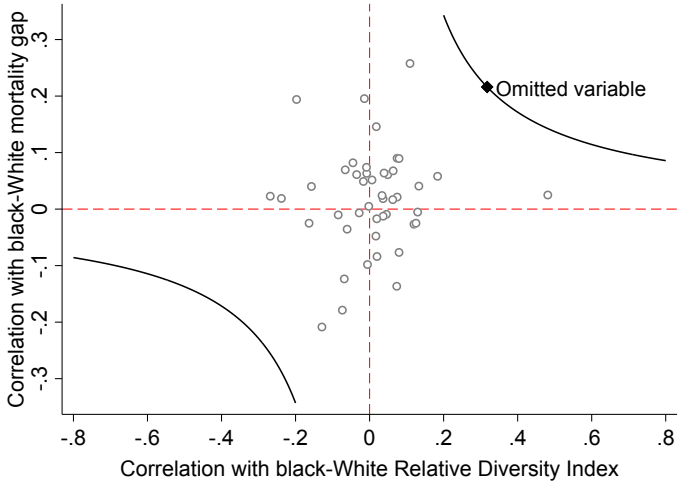

Figure S7 Frank's sensitivity analysis for the black-white mortality gap model
Each dot represents one of the 50 covariates listed in Fig. 1. The y-axis (x-axis) represents the partial correlation between the covariate and the black-white mortality gap (the black-white Relative Diversity Index) if such covariate was added to the set of 18 controls selected by the double lasso. The black curves represent the pairs of partial correlations between an omitted variable and the black-white mortality gap and the black-white Relative Diversity Index that would make the point estimate in Fig. 4 statistically non-significant at the 5% level.

Figure S8 reports results from Oster's test. The x-axis shows the R^2 from a hypothetical regression that adds an unobserved covariate to the black-white regression in Fig. 4. The y-axis shows the importance that the unobserved covariate would have in predicting the black-white Relative Diversity Index, relative to all 18 controls that are already in the model. The black curve represents the pairs of x and y values that would yield a zero coefficient on the black-white Relative Diversity Index if the unobserved covariate was added to the models. For example, for the true association between the black-white Relative Diversity Index and the black-white mortality gap to be zero, there should exist an unobserved covariate that when added to the regression increases the R^2 from .56 (as shown in Table S4) to .80 *and* is six times more predictive of the black-white Relative Diversity Index than the 18 controls selected by the lasso. Or, moving to the far right of the curve, an unobserved covariate that makes the black-white results in Fig. 4 go away would have to increase the R^2 from .56 to 1 and be almost four times more predictive of segrega-

[4]Note that 18 of the 50 covariates are already in the model. For those covariates, the partial correlations plotted in Fig. S7 are their partial correlations with the black-white mortality gap and the black-white Relative Diversity Index (i.e., net of other controls in the model).

tion than the 18 controls selected by the lasso. Such scenarios appear to be highly implausible.

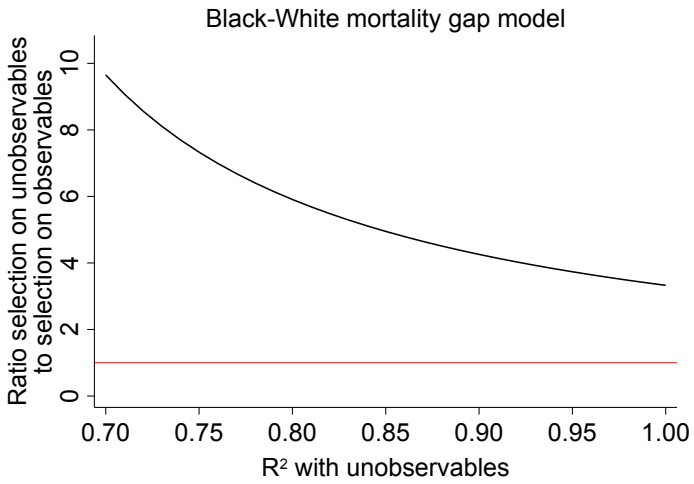

Figure S8 Oster's sensitivity analysis for the black-white mortality gap model
The y-axis represents the strength in predicting the black-white Relative Diversity Index of an unmeasured confounder, relative to the 18 covariates already included in the model (a value of 1 means that the unmeasured confounder is as predictive of the black-white Relative Diversity Index as the 18 covariates). The x-axis represents the R^2 from a hypothetical regression including the 18 covariates, state fixed effects, and the unmeasured confounder. The black curve represents the pairs of x and y values that would make the association between the black-white Relative Diversity Index and the black-white COVID-19 mortality gap equal to zero. The horizontal line indicates the threshold below which the unmeasured confounder is assumed to be less predictive of the black-white Relative Diversity Index than the 18 covariates already in the model.

Section 5 Results Using Equal Samples for the Racial/Ethnic Gap Analysis

Figures S9 and S10 show racial/ethnic mortality gaps results using a consistent sample (N=180) across the black-white and Hispanic-white models.

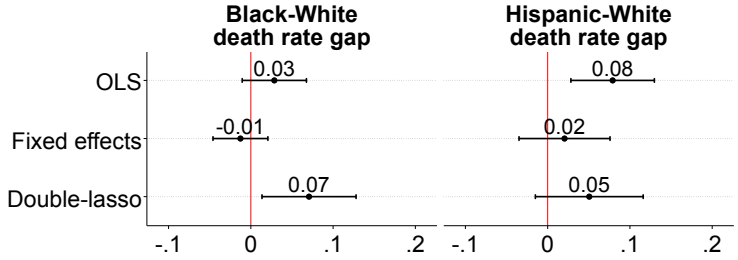

Figure S9 OLS and double-lasso regression estimates of the relationship between the black-white and Hispanic-white Relative Diversity Indices and racial gaps in COVID-19 death rates
OLS models include no controls. Fixed effects models include state fixed effects. Double-lasso models include the 18 controls selected by the lasso procedure (shown in Table S4) and state fixed effects. All regressions include population weights. Standard errors are clustered by state. Bars around estimated coefficients reflect 95% confidence intervals. Segregation indices and covariates have been standardized to have mean 0 and SD 1. The black-white (Hispanic-white) death rate gap is the difference between the log death rate for blacks (Hispanics) and the log death rate for whites. The sample includes 180 counties that reported deaths among blacks, Hispanics, and whites. COVID-19 outcomes are measured as of September 30, 2020.

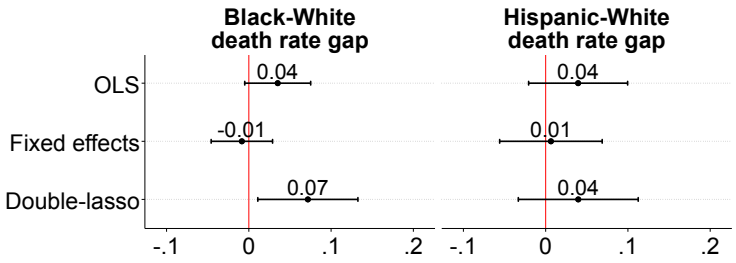

Figure S10 OLS and double-lasso regression estimates of the relationship between the black-white and Hispanic-white Theil Information Theory Indices and racial gaps in COVID-19 death rates

OLS models include no controls. Fixed effects models include state fixed effects. Double-lasso models include the 18 controls selected by the lasso procedure (shown in Table S4) and state fixed effects. All regressions include population weights. Standard errors are clustered by state. Bars around estimated coefficients reflect 95% confidence intervals. Segregation indices and covariates have been standardized to have mean 0 and SD 1. The black-white (Hispanic-white) death rate gap is the difference between the log death rate for blacks (Hispanics) and the log death rate for whites. The sample includes 180 counties that reported deaths among blacks, Hispanics, and whites. COVID-19 outcomes are measured as of September 30, 2020.

References

[1] S. F. Reardon and G. Firebaugh, "Measures of multigroup segregation," *Sociological Methodology*, vol. 32, no. 1, pp. 33–67, 2002.

[2] K. A. Frank, "Impact of a confounding variable on a regression coefficient," *Sociological Methods & Research*, vol. 29, no. 2, pp. 147–194, 2000.

[3] E. Oster, "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, vol. 37, no. 2, pp. 187–204, 2019.