Reproduction of Chakraborty 2021: An intracategorical analysis of COVID-19 and people with disabilities

Joseph Holler, Junyi Zhou, Peter Kedron, Drew An-Pham, Derrick Burt

2023-05-29

Version 1.4 | First Created Jul 7, 2021

Abstract

Chakraborty (2021) investigates the relationships between COVID-19 rates and demographic characteristics of people with disabilities by county in the lower 48 states. The aim of the study is to investigate whether people with disabilities (PwDs) face disproportionate challenges due to COVID-19. To do so, Chakraborty examines the statistical relationship between county incidence rates of COVID-19 cases and county-level percentages of people with disabilities and different socio-demographic characteristics. Specifically, Chakraborty tests county-level bivariate correlations between COVID-19 incidence against the percentage of disability as one hypothesis, and tests correlation between COVID-19 incidence and percentage of people with disabilities in 18 different socio-demographic categories of race, ethnicity, poverty status, age, and biological sex. Chakraborty then re-tests for the same county-level associations while controlling for spatial dependence. Spatial dependence is controlled by constructing generalized estimating equation (GEE) models using a combination of state and spatial clusters of COVID-19 incidence as to define the GEE clusters. One GEE model is constructed for each of the four types of socio-demographic category: race, ethnicity, age, and biological sex. Chakraborty (2021) finds significant positive relationships between COVID-19 rates and socially vulnerable demographic categories of race, ethnicity, poverty status, age, and biological sex.

This reproduction study is motivated by expanding the potential impact of Chakraborty's study for policy, research, and teaching purposes. Measuring the relationship between COVID-19 incidence and socio-demographic and disability characteristics can provide important information for public health policy-making and resource allocation. A fully reproducible study will increase the accessibility, transparency, and potential impact of Chakraborty's (2021) study by publishing a compendium complete with metadata, data, and code. This will allow other researchers to review, extend, and modify the study and will allow students of geography and spatial epidemiology to learn from the study design and methods.

In this reproduction, we will attempt to identically reproduce all of the results from the original study. This will include the map of county level distribution of COVID-19 incidence rates (Fig. 1), the summary statistics for disability and sociodemographic variables and bivariate correlations with county-level COVID-19 incidence rate (Table 1), and the GEE models for predicting COVID-19 county-level incidence rate (Table 2). A successful reproduction should be able to generate identical results as published by Chakraborty (2021).

The replication study data and code will be made available in a GitHub repository to the greatest extent that licensing and file sizes permit. The repository will be made public at github.com/HEGSRR/RPr-Chakraborty2021.

Chakraborty, J. 2021. Social inequities in the distribution of COVID-19: An intra-categorical analysis of people with disabilities in the U.S. *Disability and Health Journal* 14:1-5. DOI:[10.1016/j.dhjo.2020.101007](DOI:%5B10.1016/j.dhjo.2020

Keywords

COVID-19; Disability; Intersectionality; Race/ethnicity; Poverty; Reproducibility

Study design

The aim of this reproduction study is to implement the original study as closely as possible to reproduce the map of county level distribution of COVID-19 incidence rate, the summary statistics and bivariate correlation for disability characteristics and COVID-19 incidence, and the generalized estimating equations. Our two confirmatory hypotheses are that we will be able to exactly reproduce Chakraborty's results as presented in table 1 and table 2. Stated as null hypotheses:

> H1: There is a less than perfect match between Chakraborty's bivariate correlation coefficient for each disability/sociodemographic variable and COVID-19 incidence rate and our bivariate correlation coefficient for each disability/sociodemographic variable and COVID-19 incidence rate.
>
> H2: There is a less than perfect match between Chakraborty's beta coefficient for the GEE of each disability/sociodemographic variable and our beta coefficient for the GEE of each disability/sociodemographic variable.

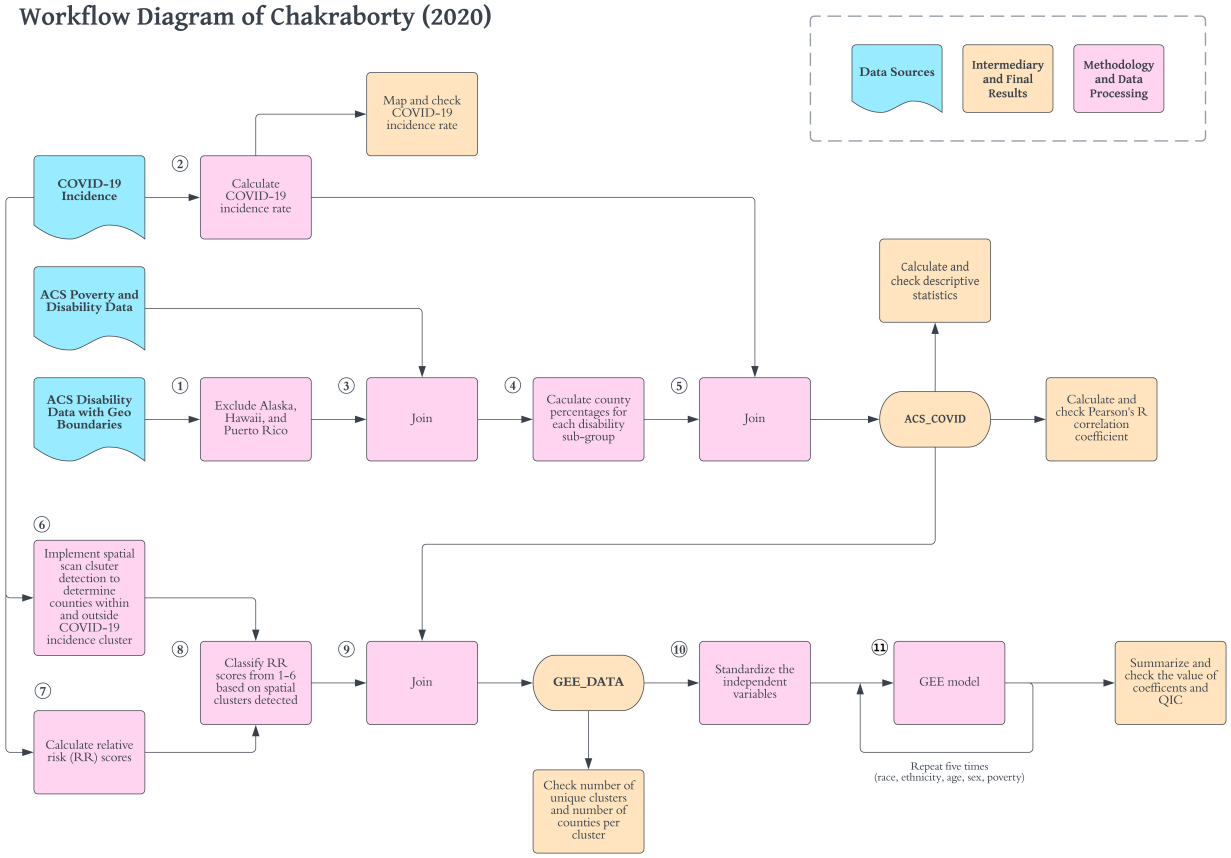
There are multiple models being tested within each of the two hypotheses. That is, H1 and H2 both encompass five models, including one for each dimension of socio-demographics: race, ethnicity, poverty status, age, and biological sex.

Original study design

The original study is **observational**, with the **exploratory** objective of determining "whether COVID-19 incidence is significantly greater in counties containing higher percentages of socio-demographically disadvantaged [people with disabilities], based on their race, ethnicity, poverty status, age, and biological sex" (Chakraborty 2021). This exploratory objective is broken down into five implicit hypotheses that each of the demographic characteristics of people with disabilities is associated with higher COVID-19 incidence rates.

The **spatial extent** of the study are the 49 contiguous states in the U.S. The **spatial scale** of the analysis is at the county level. Both COVID-19 incidence rates and demographic variables are all measured at the county level. The **temporal extent** of the COVID-19 data ranges from 1/22/2020 (when John Hopkins began collecting the data) to 8/1/2020 (when the data was retrieved for the original study). The data on disability and sociodemographic characteristics come from the U.S. Census American Community Survey (ACS) five-year estimates for 2018 (2014-2018).

There is no **randomization** in the original study.



Computational environment

The study was originally conducted using SaTScan software to implement the Kulldorff spatial scan statistic. Other software are not specified in the publication; however data files suggest and communication with the author verifies that spatial analysis and mapping was conducted in ArcGIS, generalized estimating equation (GEE) models were calculated in SPSS, and the SaTScan software version was `9.6`.

This reproduction study uses R, including the SpatialEpi package for the Kulldorff spatial scan statistics and the geepack package for GEE models.

Data

American Community Survey

American Community Survey (ACS) data for sociodemographic subcategories of people with disabilities can be accessed by using the `tidycensus` package to query the Census API. This requires an API key which can be acquired at api.census.gov/data/key_signup.html

The original study extent is the lower 48 states and Washington D.C. Therefore, Alaska, Hawai'i and Puerto Rico are removed from the data (workflow step 1). Data on people with disabilities in poverty is derived from a different census table (C18130) than data on people with disabilities and age, race, ethnicity, age, and biological sex (S1810). Therefore, join the poverty data to the other data using the GEOID (workflow step 3). Also transform the ACS geographic data into Contiguous USA Albers Equal Area projection and fix geometry errors.

Optionally, save the raw ACS data to `data/raw/public/acs.gpkg` for use in GIS software.

Calculate independent socio-demographic variables of people with disabilities as percentages for each subcategory of disability (race, ethnicity, poverty, age, and biological sex) and remove raw census data from the data frame (workflow step 4). Reproject the data into an Albers equal area conic projection.

```
## Simple feature collection with 6 features and 22 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 600219.1 ymin: 1563517 xmax: 1082567 ymax: 1791009
## Projected CRS: NAD83 / Conus Albers
##    fips statefp  county               county_st  dis_pct white_pct black_pct
## 1 21007      21 Ballard Ballard County, Kentucky 16.70848  15.21576 0.8529854
## 2 21017      21 Bourbon Bourbon County, Kentucky 16.04442  14.17517 1.7687553
## 3 21031      21  Butler  Butler County, Kentucky 21.85330  21.24671 0.0478889
## 4 21065      21  Estill  Estill County, Kentucky 26.30170  25.25188 0.0000000
## 5 21069      21 Fleming Fleming County, Kentucky 24.46302  23.88286 0.1795704
## 6 21093      21  Hardin  Hardin County, Kentucky 17.57379  13.87288 2.2477768
##   native_pct asian_pct other_pct non_hisp_white_pct   hisp_pct
## 1 0.37631711 0.0125439 0.2508781           15.10286 0.37631711
## 2 0.00000000 0.0000000 0.1004975           14.14502 0.13064670
## 3 0.20751856 0.1755926 0.1755926           21.19882 0.04788890
## 4 0.38046925 0.0000000 0.6693440           25.18847 0.06341154
## 5 0.06215899 0.0000000 0.3384212           23.88286 0.00000000
## 6 0.08207089 0.4277342 0.9433325           13.44321 0.68553332
##   non_hisp_non_white_pct  bpov_pct apov_pct  pct_5_17 pct_18_34 pct_35_64
## 1              1.2293026  4.071765 12.87696 1.8941295 0.8655294  6.460110
## 2              1.7687553  4.200364 11.83785 1.1607457 1.2160193  5.989649
## 3              0.6065927  5.696814 16.33408 1.3329077 2.3066486  8.707798
## 4              1.0498133 11.244357 15.08888 0.7257099 2.4660044 14.359191
## 5              0.5801506  6.129524 18.38164 2.4587333 2.7419021  9.841840
## 6              3.4450463  3.795861 13.82833 1.4695517 2.3790903  7.658662
##   pct_65_74   pct_75  male_pct female_pct                       geometry
## 1  2.696939 4.703964  8.454591   8.253889 POLYGON ((600246.8 1577488,...
## 2  3.914376 3.643033  8.637757   7.406663 POLYGON ((998945.1 1755389,...
## 3  4.214223 4.996408 11.293798  10.559502 POLYGON ((796316.5 1596948,...
## 4  4.692454 3.755372 12.534348  13.767350 POLYGON ((1035030 1687824, ...
## 5  4.530700 4.889840 11.658264  12.804752 POLYGON ((1036064 1778344, ...
## 6  2.830963 3.168902  8.652203   8.921589 POLYGON ((848902.9 1660201,...
```

COVID-19 rates

Data on COVID-19 rates from the Johns Hopkins University dashboard have been provided directly with the research compendium because the data is no longer available online in the state in which it was downloaded on August 1, 2020. The dashboard and cumulative counts of COVID-19 cases and deaths were continually updated, so an exact reproduction required communication with the original author, Jayajit Chakraborty, for assistance with provision of data from August 1, 2020.

Calculate the COVID incidence rate as the cases per 100,000 people (workflow step 2). Convert the COVID data to a non-geographic data frame.

```
## # A tibble: 6 x 6
##   fips     pop cases     x     y covid_rate
##   <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
## 1 01001  55601   972 -86.6  32.5      1748.
## 2 01003 218022  3056 -87.7  30.7      1402.
```

```
## 3 01005  24881   550 -85.4  31.9      2211.
## 4 01007  22400   355 -87.1  33.0      1585.
## 5 01009  57840   685 -86.6  34.0      1184.
## 6 01011  10138   430 -85.7  32.1      4241.
```

Join dependent COVID data to independent ACS sociodemographic data.

```
##    fips statefp  county               county_st covid_rate  dis_pct white_pct
## 1 21007      21 Ballard Ballard County, Kentucky     375.99 16.70848  15.21576
## 2 21017      21 Bourbon Bourbon County, Kentucky     336.90 16.04442  14.17517
## 3 21031      21  Butler  Butler County, Kentucky    2278.42 21.85330  21.24671
## 4 21065      21  Estill  Estill County, Kentucky      84.52 26.30170  25.25188
## 5 21069      21 Fleming Fleming County, Kentucky     367.24 24.46302  23.88286
## 6 21093      21  Hardin  Hardin County, Kentucky     463.05 17.57379  13.87288
##   black_pct native_pct asian_pct other_pct non_hisp_white_pct   hisp_pct
## 1 0.8529854 0.37631711 0.0125439 0.2508781           15.10286 0.37631711
## 2 1.7687553 0.00000000 0.0000000 0.1004975           14.14502 0.13064670
## 3 0.0478889 0.20751856 0.1755926 0.1755926           21.19882 0.04788890
## 4 0.0000000 0.38046925 0.0000000 0.6693440           25.18847 0.06341154
## 5 0.1795704 0.06215899 0.0000000 0.3384212           23.88286 0.00000000
## 6 2.2477768 0.08207089 0.4277342 0.9433325           13.44321 0.68553332
##   non_hisp_non_white_pct  bpov_pct apov_pct  pct_5_17 pct_18_34 pct_35_64
## 1              1.2293026  4.071765 12.87696 1.8941295 0.8655294  6.460110
## 2              1.7687553  4.200364 11.83785 1.1607457 1.2160193  5.989649
## 3              0.6065927  5.696814 16.33408 1.3329077 2.3066486  8.707798
## 4              1.0498133 11.244357 15.08888 0.7257099 2.4660044 14.359191
## 5              0.5801506  6.129524 18.38164 2.4587333 2.7419021  9.841840
## 6              3.4450463  3.795861 13.82833 1.4695517 2.3790903  7.658662
##   pct_65_74   pct_75  male_pct female_pct    pop cases         x        y
## 1  2.696939 4.703964  8.454591   8.253889   7979    30 -88.99934 37.05844
## 2  3.914376 3.643033  8.637757   7.406663  20184    68 -84.21713 38.20670
## 3  4.214223 4.996408 11.293798  10.559502  12772   291 -86.68173 37.20720
## 4  4.692454 3.755372 12.534348  13.767350  14198    12 -83.96427 37.69245
## 5  4.530700 4.889840 11.658264  12.804752  14432    53 -83.69666 38.37013
## 6  2.830963 3.168902  8.652203   8.921589 110356   511 -85.96334 37.69795
```

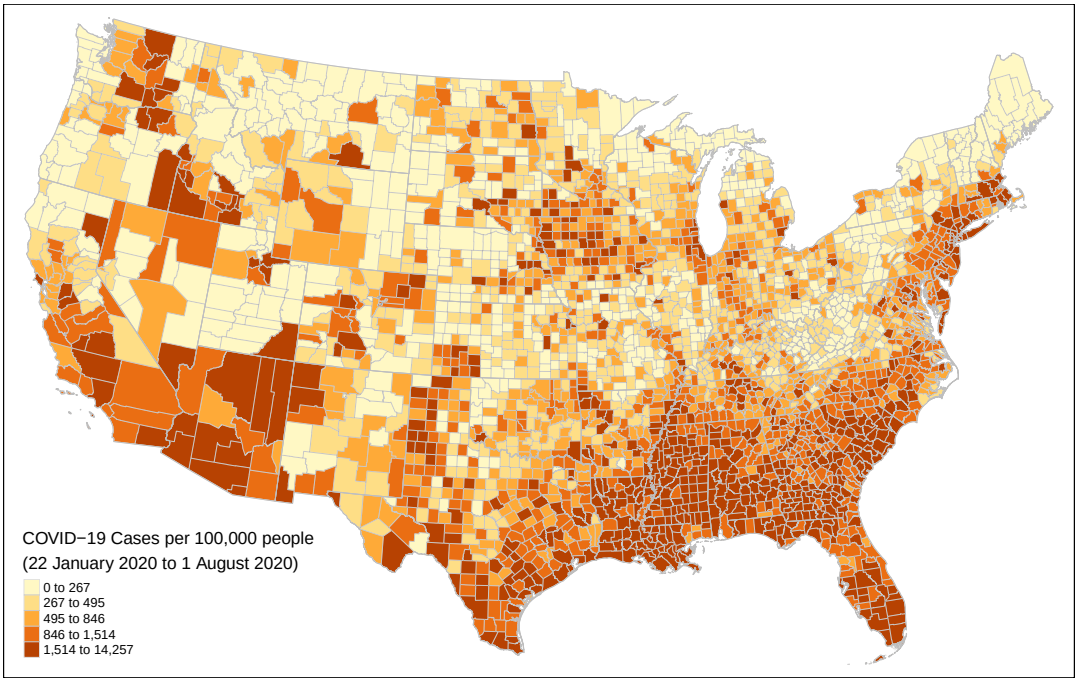
Missing data

There is one county with missing disability and poverty data. This was not mentioned in the original study or our pre-analyis plan. However, we replace the missing data with zeros, producing results identical to Chakraborty's.

```
##    fips statefp     county                     county_st covid_rate  dis_pct
## 1 35039      35 Rio Arriba Rio Arriba County, New Mexico     751.17 16.06467
##   white_pct  black_pct native_pct  asian_pct other_pct non_hisp_white_pct
## 1  10.77458 0.03837102   2.744807 0.03837102  2.468536           2.355981
##   hisp_pct non_hisp_non_white_pct bpov_pct apov_pct  pct_5_17 pct_18_34
## 1 11.39619               2.312494       NA       NA 0.3069682   1.25857
##   pct_35_64 pct_65_74   pct_75 male_pct female_pct   pop cases         x
## 1  6.781439  3.391998 4.279648 8.556738    7.50793 39006   293 -106.6932
##          y
## 1 36.50962
```

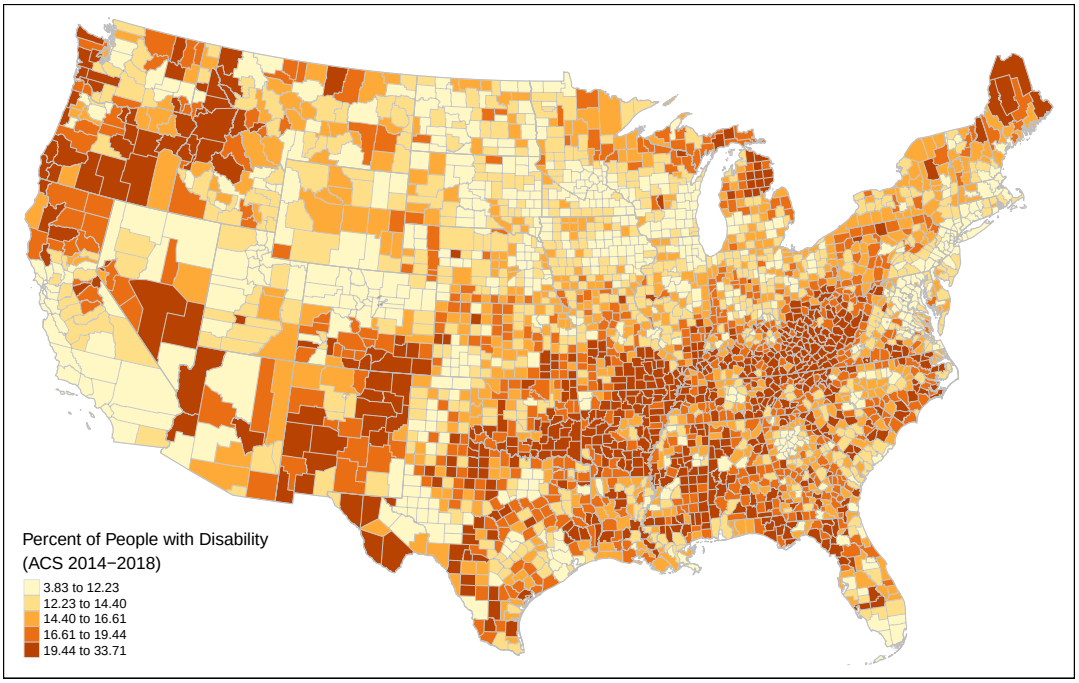
Map COVID-19 incidence

Map the county level distribution of COVID-19 incidence rates, comparing to Figure 1 of the original study.



Map disability rates

Unplanned deviation for reproduction: We also map the spatial distribution of the percent of people with any disability to improve our understanding of the geographic patterns and relationships of between the overarching independent variable (percentage of people with disability) and the dependent variable (COVID-19 incidence rate).



	min	max	mean	SD	ShapiroWilk	p
covid_rate	0.00	14257.17	966.90	1003.96	0.74	0
dis_pct	3.83	33.71	15.95	4.40	0.98	0
white_pct	0.85	33.26	13.55	4.63	0.98	0
black_pct	0.00	20.70	1.48	2.66	0.61	0
native_pct	0.00	13.74	0.28	0.94	0.28	0
asian_pct	0.00	3.45	0.09	0.18	0.51	0
other_pct	0.00	15.24	0.55	0.65	0.57	0
non_hisp_white_pct	0.10	33.16	12.84	4.81	0.99	0
hisp_pct	0.00	25.26	0.99	2.15	0.42	0
non_hisp_non_white_pct	0.00	20.93	2.13	2.75	0.70	0
bpov_pct	0.00	14.97	3.57	1.85	0.93	0
apov_pct	0.00	27.30	12.48	3.06	0.99	0
pct_5_17	0.00	5.08	1.03	0.48	0.95	0
pct_18_34	0.00	5.59	1.56	0.67	0.96	0
pct_35_64	1.01	18.36	6.35	2.30	0.96	0
pct_65_74	0.00	12.73	3.09	1.16	0.95	0
pct_75	0.00	11.13	3.87	1.19	0.97	0
male_pct	1.30	18.19	8.06	2.37	0.98	0
female_pct	1.91	19.94	7.90	2.26	0.98	0

Descriptive statistics

Calculate descriptive statistics for dependent COVID-19 rate and independent socio-demographic characteristics, reproducing the min, max, mean, and SD columns of original study table 1.

Planned deviation for reanalysis: We also calculate the Shapiro Wilk test for normality.

Compare reproduced descriptive statistics to original descriptive statistics. Difference is calculated as 'reproduction study - original study'. Identical results will result in zero.

The descriptive statistics are identical, except that the original study seems to have rounded the COVID-19 statistics to zero decimal places.

Analytical methods

Bivariate parametric correlation analysis

Calculate Pearson's R Correlation Coefficient of each independent variable and the COVID-19 incidence rate, reproducing the Pearson's R column of original study Table 1.

Compare the reproduced Pearson's r correlation coefficients to the original study's Pearson's r correlation coefficients. Stars indicates the significance level with two stars for `p < 0.01` and one star for `p < 0.05`. Correlation difference `rp_r_diff` is calculated between the reproduction study `rp_r` and original study `or_r` as `rp_r_diff = rp_r - or_r` Direction difference `rp_dir_diff` is calculated as `(rp_r > 0) - (or_r > 0)`, giving `0` if both coefficients have the same direction, `1` if the reproduction is positive and the original is negative, and `-1` if the reproduction is negative but the original is positive.

Reproduction correlation coefficients varied slightly from the original study coefficients by +/- 0.006. All but one Pearson's correlation coefficient was significant to the same level, and the exception was age 18 to 34. Counter-intuitively, the correlation coefficient was slightly closer to 0 but the p value was also found to be more significant, suggesting a difference in the estimation of t and/or p, or a typographical error. All of the coefficients had the same direction.

Unplanned Deviation for Reproduction: We should expect identical results for this correlation test, so we loaded the original author's data from `Aug1GEEdata.csv` to re-test the statistic, calculated as `unplanned_r`

	min	max	mean	SD
covid_rate	0	0.17	-0.1	-0.04
dis_pct	0	0.00	0.0	0.00
white_pct	0	0.00	0.0	0.00
black_pct	0	0.00	0.0	0.00
native_pct	0	0.00	0.0	0.00
asian_pct	0	0.00	0.0	0.00
other_pct	0	0.00	0.0	0.00
non_hisp_white_pct	0	0.00	0.0	0.00
hisp_pct	0	0.00	0.0	0.00
non_hisp_non_white_pct	0	0.00	0.0	0.00
bpov_pct	0	0.00	0.0	0.00
apov_pct	0	0.00	0.0	0.00
pct_5_17	0	0.00	0.0	0.00
pct_18_34	0	0.00	0.0	0.00
pct_35_64	0	0.00	0.0	0.00
pct_65_74	0	0.00	0.0	0.00
pct_75	0	0.00	0.0	0.00
male_pct	0	0.00	0.0	0.00
female_pct	0	0.00	0.0	0.00

variable	r	t	p
dis_pct	-0.060	3.350	0.000
white_pct	-0.332	19.612	0.000
black_pct	0.460	28.847	0.000
native_pct	0.019	1.072	0.142
asian_pct	0.094	5.272	0.000
other_pct	0.026	1.460	0.072
non_hisp_white_pct	-0.361	21.545	0.000
hisp_pct	0.119	6.686	0.000
non_hisp_non_white_pct	0.442	27.429	0.000
bpov_pct	0.106	5.914	0.000
apov_pct	-0.151	8.513	0.000
pct_5_17	0.084	4.688	0.000
pct_18_34	0.063	3.493	0.000
pct_35_64	-0.008	0.460	0.323
pct_65_74	-0.091	5.113	0.000
pct_75	-0.186	10.541	0.000
male_pct	-0.134	7.519	0.000
female_pct	0.023	1.305	0.096
pop	0.128	7.215	0.000
cases	0.209	11.891	0.000
x	0.099	5.540	0.000
y	-0.412	25.195	0.000

variable	or_r	or_stars	rp_r	rp_stars	rp_r_diff	rp_stars_diff	rp_dir_diff
dis_pct	-0.056	2	-0.060	2	-0.004	0	0
white_pct	-0.326	2	-0.332	2	-0.006	0	0
black_pct	0.456	2	0.460	2	0.004	0	0
native_pct	0.020	0	0.019	0	-0.001	0	0
asian_pct	0.097	2	0.094	2	-0.003	0	0
other_pct	0.028	0	0.026	0	-0.002	0	0
non_hisp_white_pct	-0.355	2	-0.361	2	-0.006	0	0
hisp_pct	0.119	2	0.119	2	0.000	0	0
non_hisp_non_white_pct	0.439	2	0.442	2	0.003	0	0
bpov_pct	0.108	2	0.106	2	-0.002	0	0
apov_pct	-0.146	2	-0.151	2	-0.005	0	0
pct_5_17	0.083	2	0.084	2	0.001	0	0
pct_18_34	0.066	1	0.063	2	-0.003	1	0
pct_35_64	-0.005	0	-0.008	0	-0.003	0	0
pct_65_74	-0.089	2	-0.091	2	-0.002	0	0
pct_75	-0.181	2	-0.186	2	-0.005	0	0
male_pct	-0.131	2	-0.134	2	-0.003	0	0
female_pct	0.028	0	0.023	0	-0.005	0	0

below.

The author's original data produced coefficients identical to the original publication! Is it possible that the data values are correct but have been reassigned / transposed to different counties?

Unplanned Deviation for Reproduction: Considering the precise bitwise reproduction of descriptive statistics and of correlation statistics from author-provided data, we decided to recalculate the COVID-19 incidence rate with author-provided case and population data for comparison to the author-provided incidence rate.

We found that 13 counties had incorrect COVID-19 incidence scores, and the scores seem to be transposed from other counties, such that the overall descriptive statistics were accurate but the correlation coefficients were inaccurate. This finding implies that subsequent analyses using the COVID-19 Incidence rate will be slightly different and more accurate in this reproduction study than in the original study.

Unplanned deviation for reproduction: Join the original author's Incidence data into our reproduction data frame so that we can later test for sensitivity to this error. Then report any counties for which the reproduced COVID incidence rate differs from the original author's COVID incidence rate.

The join worked, highlighting the same 13 counties with inconsistent incidence rates. This also confirms that our reproduced dependent variable is identical to the original dependent variable with the exception of these three counties.

Bivariate nonparametric correlation analysis

Unplanned Deviation for Reproduction: The dependent and independent variables in this study do not have normal distributions, as shown in the Shapiro-Wilk test results above. Therefore, we deviate from the original study to use the Spearman's Rho non-parametric correlation test.

Compare the Spearman's *rho* correlation coefficients to the reproduced Pearson's r correlation coefficients. Differences are calculated as *Spearman's Rho - Pearson's R*.

Three variables change significance levels, with *Native American* and *Other* races gaining significance and *age 18-34* losing significance. Two correlations change direction, with both *Native American* race (illustrated in scatterplot below) and *Female* households switching from positive correlations to negative correlations. Instabilities between the parametric and non-parametic correlations arise from variables with very skewed distributions and/or weak correlations at the county level. Some difference may also be attributable to the 13

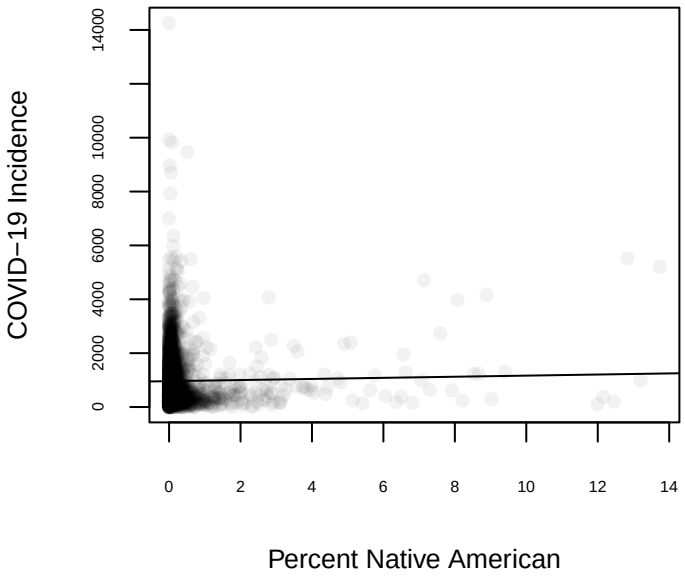
variable	unplanned_r	or_r	diff
dis_pct	-0.056	-0.056	0
white_pct	-0.326	-0.326	0
black_pct	0.456	0.456	0
native_pct	0.020	0.020	0
asian_pct	0.097	0.097	0
other_pct	0.028	0.028	0
non_hisp_white_pct	-0.355	-0.355	0
hisp_pct	0.119	0.119	0
non_hisp_non_white_pct	0.439	0.439	0
bpov_pct	0.108	0.108	0
apov_pct	-0.146	-0.146	0
pct_5_17	0.083	0.083	0
pct_18_34	0.066	0.066	0
pct_35_64	-0.005	-0.005	0
pct_65_74	-0.089	-0.089	0
pct_75	-0.181	-0.181	0
male_pct	-0.131	-0.131	0
female_pct	0.028	0.028	0

COUNTY_FIPS	ST_Name	Countyname	Total_POP	Cases	Incidence	recalc_Incidence	Incidence_diff
1115	Alabama	St. Clair	88690	1151	1349.52	1297.78	-51.74
1117	Alabama	Shelby	215707	2911	1297.78	1349.52	51.74
5123	Arkansas	St. Francis	25439	1112	704.16	4371.24	3667.08
5125	Arkansas	Saline	121421	855	397.33	704.16	306.83
5127	Arkansas	Scott	10319	41	314.15	397.33	83.18
5129	Arkansas	Searcy	7958	25	1322.08	314.15	-1007.93
5131	Arkansas	Sebastian	127753	1689	5420.39	1322.08	-4098.31
5133	Arkansas	Sevier	17139	929	570.08	5420.39	4850.31
5135	Arkansas	Sharp	17366	99	4371.24	570.08	-3801.16
8039	Colorado	Elbert	26282	85	626.60	323.42	-303.18
8041	Colorado	El Paso	713856	4473	323.42	626.60	303.18
8065	Colorado	Lake	7824	70	349.85	894.68	544.83
8067	Colorado	La Plata	56310	197	894.68	349.85	-544.83

county_st	covid_rate	or_incidence
St. Clair County, Alabama	1297.78	1349.52
Shelby County, Alabama	1349.52	1297.78
St. Francis County, Arkansas	4371.24	704.16
Saline County, Arkansas	704.16	397.33
Scott County, Arkansas	397.33	314.15
Searcy County, Arkansas	314.15	1322.08
Sebastian County, Arkansas	1322.08	5420.39
Sevier County, Arkansas	5420.39	570.08
Sharp County, Arkansas	570.08	4371.24
Elbert County, Colorado	323.42	626.60
El Paso County, Colorado	626.60	323.42
Lake County, Colorado	894.68	349.85
La Plata County, Colorado	349.85	894.68

Variable	Pearson's		Spearman's		Difference		
	R	Stars	Rho	Stars	Rho - R	Stars	Direction
dis_pct	-0.060	2	-0.113	2	-0.053	0	0
white_pct	-0.332	2	-0.421	2	-0.089	0	0
black_pct	0.460	2	0.575	2	0.115	0	0
native_pct	0.019	0	-0.084	2	-0.103	2	-1
asian_pct	0.094	2	0.194	2	0.100	0	0
other_pct	0.026	0	0.104	2	0.078	2	0
non_hisp_white_pct	-0.361	2	-0.454	2	-0.093	0	0
hisp_pct	0.119	2	0.231	2	0.112	0	0
non_hisp_non_white_pct	0.442	2	0.481	2	0.039	0	0
bpov_pct	0.106	2	0.062	2	-0.044	0	0
apov_pct	-0.151	2	-0.205	2	-0.054	0	0
pct_5_17	0.084	2	0.079	2	-0.005	0	0
pct_18_34	0.063	2	0.034	1	-0.029	-1	0
pct_35_64	-0.008	0	-0.020	0	-0.012	0	0
pct_65_74	-0.091	2	-0.151	2	-0.060	0	0
pct_75	-0.186	2	-0.285	2	-0.099	0	0
male_pct	-0.134	2	-0.201	2	-0.067	0	0
female_pct	0.023	0	-0.014	0	-0.037	0	-1

counties with data errors in the COVID-19 Incidence Rate. In such distributions, outlier observations have more weight in the parametric Person's R test than in the non-parametric Spearman's Rho test.



Kulldorff spatial scan statistic

We use a Kulldorff spatial scan statistic to detect spatial clusters of high COVID-19 incidence (workflow step 6). The statistic uses a Monte Carlo simulation to calculate statistical significance, and therefore may not

produce identical results each time.

The original study uses SaTScan software to implement the Kulldorff spatial scan statistic model. In SaTScan, models may be specified with many parameters having significant implications for results. The original manuscript only specifies that Poisson model should be used. We can also intuit that the model is discrete (locations are stationary and non-random), and spatial only (there is no temporal dimension). The author-provided SaTScan results `SatScan_results.txt` contains additional parameters which appear to adhere closely to the software's default settings. These include the maximum cluster size of "50 percent of population at risk", and the "GINI optimized cluster collection" and "no geographical overlap" options for detecting secondary clusters. The "P-value Cutoff" for significant clusters option did not appear in the v9.6 output, suggesting that the software only allowed the default "no" option for this at the time of the original study.

SaTScan software can also output two versions of geographic data:

- The `col` cluster polygon shapefile contains a circle for each cluster, where each polygon is a circle defined by the cluster center and radius. The attributes include a variable `REL_RISK` for cluster relative risk
- The `gis` location point shapefile contains one point for each county in a cluster. The attributes include variables `LOC_RR` for local relative risk and `CLU_RR` for cluster relative risk

The SaTScan software implementation of the Kulldorff spatial scan statistic calculates two relative risk scores for locations:

- cluster relative risk is the incidence rate of the population within the cluster divided by the incidence rate of the population outside of the cluster. This is calculated as `REL_RISK` in the `col` cluster polygon shapefile and as `CLU_RR` in the `gis` location point shapefile.
- local relative risk is the incidence rate of population within a location divided by the incidence rate of the population outside of the location. This is calculated as `LOC_RR` in the `gis` location shapefile, and is not calculated in the `col` cluster polygon shapefile.

For the purposes of interpreting the spatial scan statistic, a *location* is a *county centroid* while a *cluster* is a *collection of counties* with high incidence rates, defined in the shape of a circle with a *center* location (a county centroid) and a *radius*.

The original study is not clear about using the cluster geographic data *vs* the location geographic data or the cluster relative risk *vs* local relative risk. However, The author-provided `SatScan_results.txt` results file indicates a geographic cluster file but no location file, and the author-provided `Aug1GEEdata.csv` data table contains a `REL_RISK` field but no `CLU_RR` field or `LOC_RR` field. This suggests that in the original study, the `col` polygon cluster shapefile and *cluster* relative risk were used to represent COVID-19 risk and define GEE clusters.

The spatial scan statistic is based on case counts and total population, and is therefore unaffected by errors in the COVID Incidence rate.

Planned deviation for reproduction: We opted to use the SpatialEpi package in R, selecting open source software with R integration over SatSCan software, which is free but not open. The Kulldorff spatial scan statistic model in SpatialEpi also supports a discrete Poisson spatial model, and uses the GINI coefficient to select secondary clusters with no geographical overlap that maximize the difference between locations inside of clusters and locations outside of clusters. We expected that this set of software options could reproduce identical results compared to SaTScan.

First, calculate the Kulldorff spatial scan statistic using SpatialEpi. Optionally, skip this code block due to long run times of more than 10 minutes.

Load pre-calculated Kulldorff spatial scan results. Alternatively, skip or modify this code block to use your own version of the SpatialEpi Kulldorff results.

Report Kulldorff spatial scan results.

```
## [1] Most likely cluster:

## $location.IDs.included
##  [1] 1824 1835 1797 1818 1825 1749 1854 1742 1837 1747 1838  280 1760 1846 1756
##
## $population
## [1] 16949211
##
## $number.of.cases
## [1] 469091
##
## $expected.cases
## [1] 233805.6
##
## $SMR
## [1] 2.006329
##
## $log.likelihood.ratio
## [1] 97983.07
##
## $monte.carlo.rank
## [1] 1
##
## $p.value
## [1] 0.001

## [1] Number of Secondary clusters: 134
```

The `SpatialEpi` implementation of Kulldorff spatial scan statistics provides output in the form of hierarchical lists analogous to the text output of SaTScan, but does not output a simple data frame or tabular output analogous to the shapefiles from SaTScan. Therefore, additional steps are required to append the Kulldorff scan results to the `acs_covid` simple features data frame. This can be done by assigning unique cluster ID's to each county within a cluster. Clusters include the county at the center of a cluster and all of the other counties within the cluster radius. Therefore, we use the FIPS code of the county at the center of each cluster as the unique cluster ID.

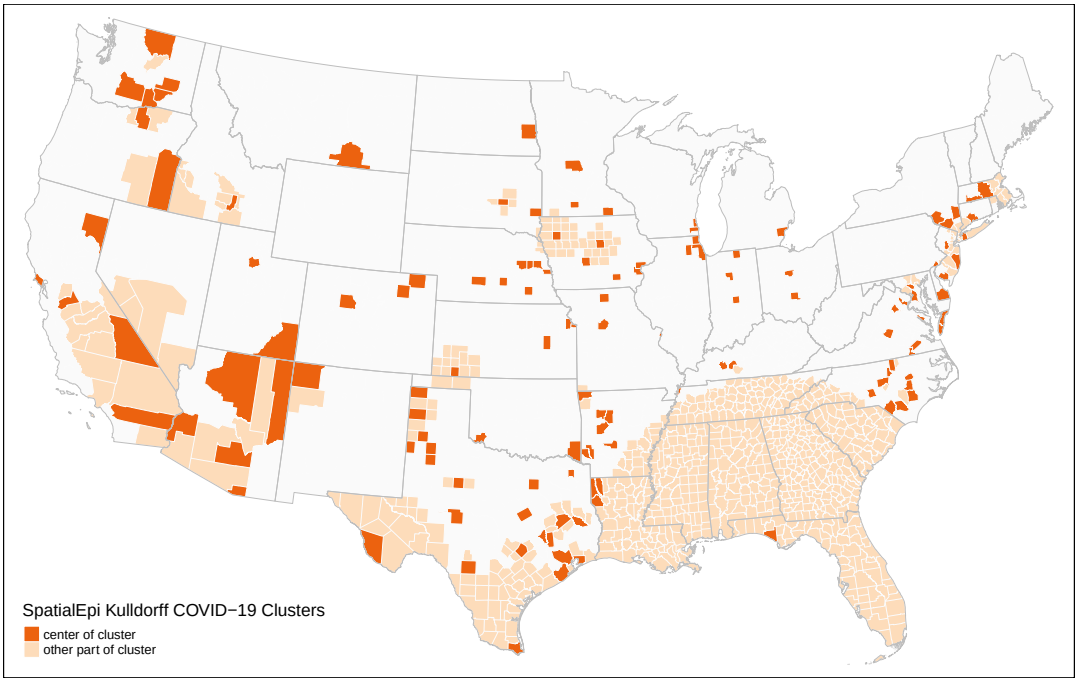
Map Kulldorff clusters

Unplanned deviation for reproduction: The original study does not include visualizations of the spatial structure and distribution of COVID-19 clusters.

First, we must join the Kulldorff spatial scan cluster IDs to the acs_covid simple features dataframe. Although this was planned in workflow step 9, the order of operations between steps 9 and steps 7 and 8 is not important.

Next, calculate a new field `isCluster` to identify counties in COVID-19 clusters. Additionally, distinguish between counties defining the center of a cluster from counties constituting other parts of a cluster by comparing the cluster ID (equivalent to the center county's fips code) to the county fips code.

Planned deviation for reproduction: Map the `SpatialEpi` cluster results.



Unplanned deviation for reproduction: The `SpatialEpi` implementation of Kulldorff spatial scan statistics does not calculate local relative risk or cluster relative risk. Therefore, the next step is to calculate local and cluster relative risk (workflow step 7).

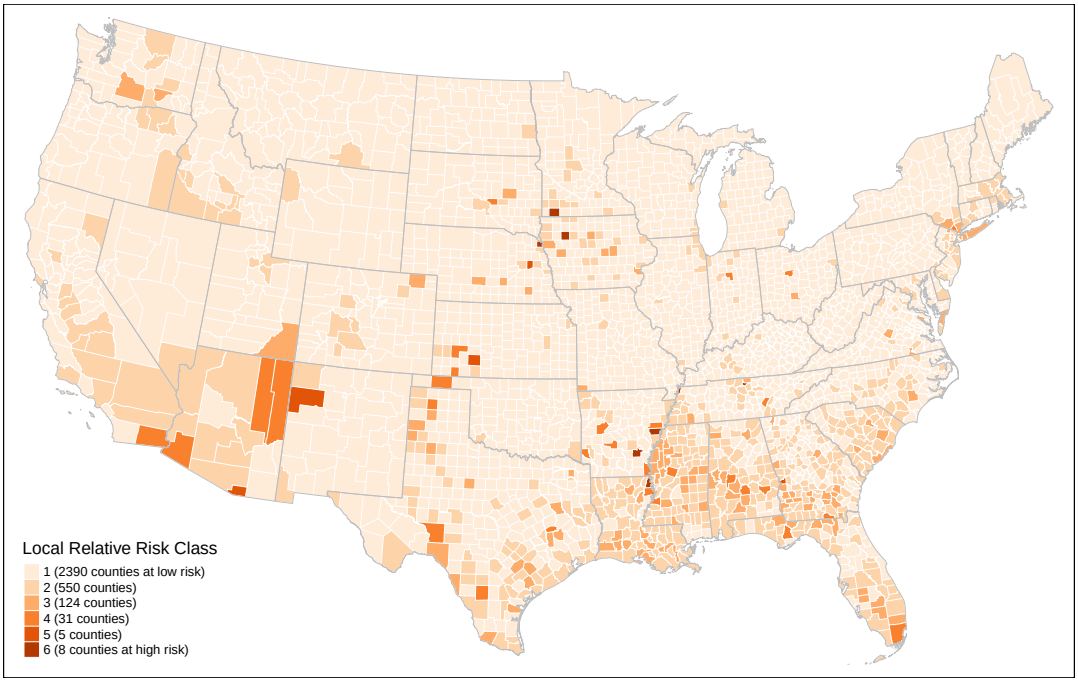
Classify relative risk on a scale from 1 to 6 (workflow step 8). Counties falling outside of any cluster are assigned a score of 1.

```
## # A tibble: 6 x 5
##   clusterID rr_cluster cluster_class rr_loc loc_class
##   <chr>          <dbl>         <dbl>  <dbl>     <int>
## 1 <NA>           NA                1 0.273         1
## 2 <NA>           NA                1 0.244         1
## 3 21031           1.34             2 1.65          2
## 4 <NA>           NA                1 0.0613        1
## 5 <NA>           NA                1 0.266         1
## 6 <NA>           NA                1 0.336         1
```

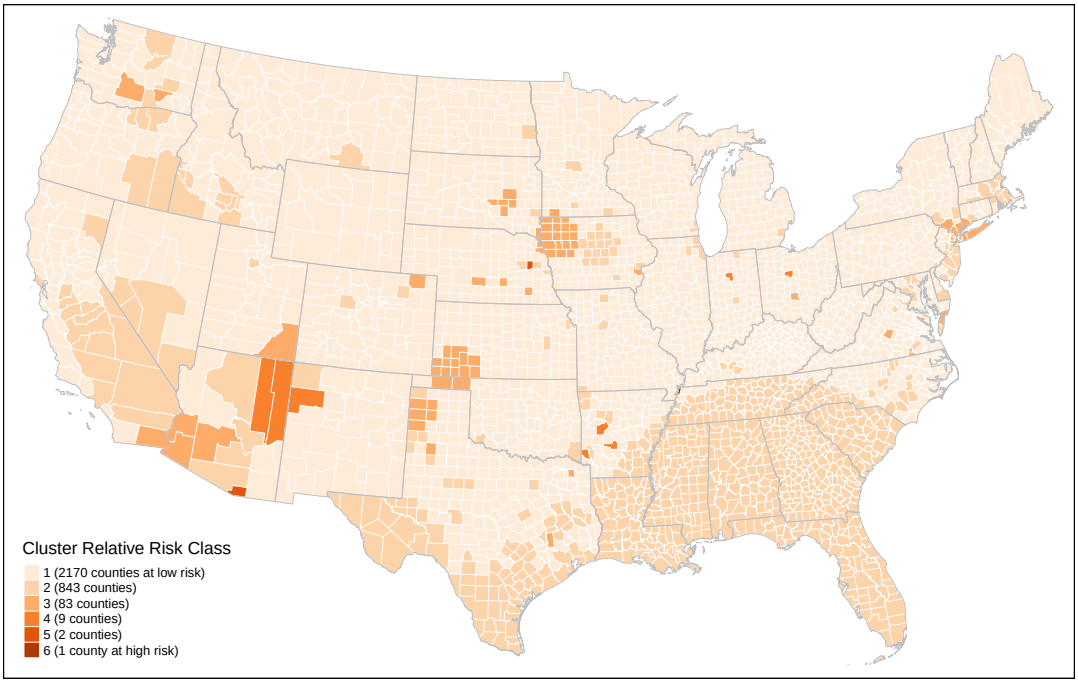
Map relative risk scores

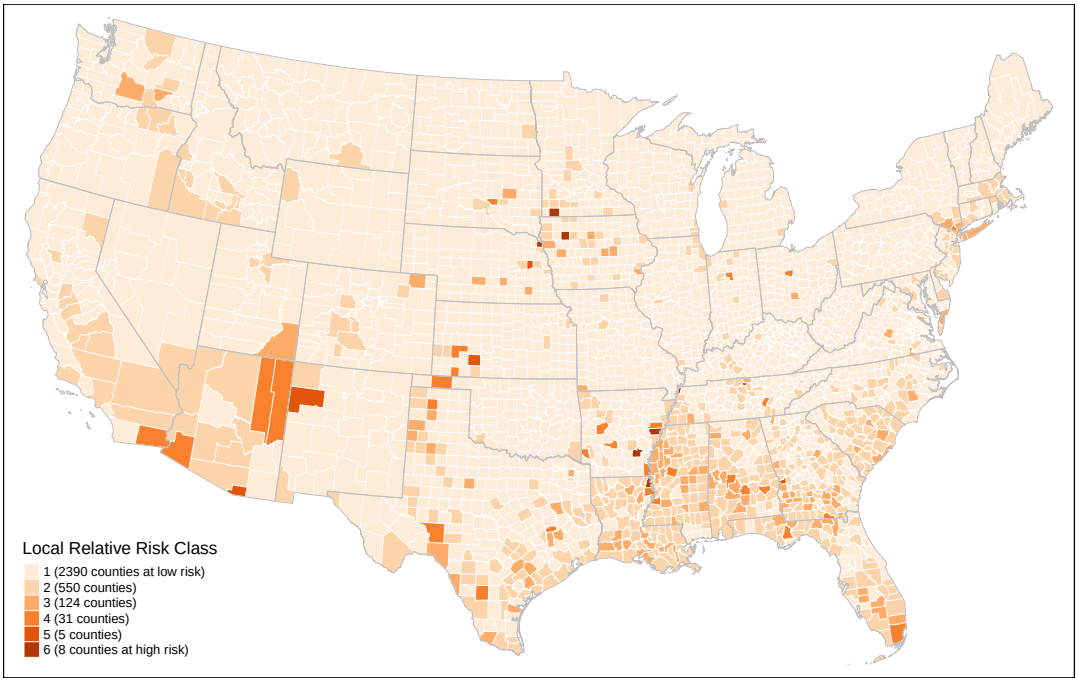
Unplanned deviation for reproduction: It would be helpful to visualize the spatial distributions of local relative risk classes and Kulldorff cluster relative risk classes in advance of using these classes to control for spatial heterogeneity in GEE models.

First, map the spatial distribution of local relative risk score classifications.



Next, map the cluster relative risk scores for comparison. Note that following the original study classification methodology, counties outside of clusters are assigned the lowest risk class of `1`.





Comparing the cluster and local relative risk classifications for regions like the Southeast, it is apparent that some areas of high risk are represented with large clusters that have an averaging effect on the cluster-based relative risk score. This effect is more pronounced for clusters with low compactness (e.g. the Southeast cluster stretched over the "black belt" region from Louisiana and Arkansas to Georgia) than clusters with higher compactness (e.g. New York City) because the circular shape of clusters includes more low-risk counties.

Compare clusters

The original study did not directly report any results from the Kulldorff spatial scan statistic. However, the Kulldorff cluster relative risk scores were combined with states to create clusters for GEE models, hereafter called "GEE clusters". The original study reported `102` unique GEE clusters having a range of `1` to `245` counties in each cluster.

In order to compare results, we first create cluster IDs as combinations of the relative risk class and the state ID. Then, we find the number of unique clusters and frequency counties per cluster in our reproduction study for comparison to the original study.

```
## 111 unique clusters based on spatialEpi CLUSTER relative risk
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    7.00   27.56   50.50  159.00
```

We failed to reproduce the same configuration of GEE clusters as the original study, finding 9 more clusters than the original study and a much smaller maximum cluster of 159 counties compared to 245 counties.

Reproduce Kulldorff spatial scan statistic in SaTScan

Unplanned deviation for reproduction: Upon failing to reproduce an identical number of GEE clusters using SpatialEpi in R, we reproduced the procedure in the free but not open SaTScan software, using the current software version 10.1. The input data files (`case`, `Coordinates.geo`, and `Population.pop`), and output data files (`sat_scan_rpr.txt`, `sat_scan_rpr.col.shp`, and `sat_scan_rpr.gis.shp`) are found in the `data/derived/public/satscan` directory. The `sat_scan_rpr.txt` file reports the model parameters used in addition to results.

Although it is not ideal to intercede with this unplanned deviation at this step, is the first step in the methodology following the Kulldorff spatial scan statistic with a result reported in the original publication.

First, load and verify whether our SaTScan reproduction data compares to the author-provided SaTScan data.

```
## 96 reproduced relative risk observations
## 96 author-provided relative risk observations
```

```
## 96 reproduced relative risk values match the original author's relative risk values
```

Our SaTScan results exactly reproduced the author-provided SaTScan results data.

Map SaTScan spatial clusters Join the SaTScan results to `acs_covid` for mapping and analysis.

```
## Joining 96 records with 96 unique LOC_ID county values
```

Unplanned deviation for reproduction: Visualize the spatial distribution of the author-provided Kulldorff COVID-19 Clusters.

In the map above, clusters containing only one county have no visible circle. Clusters containing two counties are encircled, but have no label. Clusters containing three or more counties are encircled and labelled with the number of counties.

Note that this version of data only includes the 96 counties defining cluster centers, visualized with fill colors above. The data excludes all of the non-center counties in clusters with more than one county. The extent of these larger clusters is visualized by unfilled circles defined by cluster radii.

Additionally, the SaTScan software confusingly merges two sets of clusters in the results when the user uses the (default) option for GINI-optimized clusters. One set of results is a hierarchical non-overlapping set of clusters. These clusters are noted with `GINI_CLUST = F` in the results. The second set of results is a set of hierarchical non-overlapping clusters designed to maximize the GINI coefficient of inequality between counties within clusters and counties outside of clusters. These clusters are noted with `GINI_CLUST = T` in the results.

Merged together as they are, the two sets of secondary clusters overlap one another geographically, causing ambiguity in terms of which cluster-based relative risk score should be used at each location.

Unplanned deviation for reproduction: Can we also use these reproduced SaTScan results to exactly reproduce the author-reported frequency of original GEE classes and maximum counties per class? If the results match, it will confirm that the problems identified above have propagated through the original study analysis.

```
## 102 unique clusters based on spatialEpi CLUSTER relative risk
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    3.00   29.99   58.75  245.00
```

Using SaTScan Kulldorff clusters, we have exactly reproduced the author-reported frequency of original GEE classes and maximum counties per class. We have confirmed that the original study used the *cluster relative risk* of the *center county* of each cluster, including both the *hierarchical* and *GINI-optimized* sets of clusters.

Compare SaTScan clusters to SpatialEpi clusters Unplanned Deviation for Reanalysis: At this point it is clear that the best decision will be to shift from a *reproduction* study to a *reanalysis* study, intentionally altering methodological decisions to achieve a more valid outcome. We prefer to include *all counties* contained in each cluster, and to use only *one set of non-overlapping clusters*, as produced by the `SpatialEpi` algorithm.

Given the shifting goal, how sensitive is this study to the choice of computational environment for the Kulldorff scan statistics? To answer this question, we must load the local SaTSCan results inclusive of all counties within clusters, filter the results to focus on the standard hierarchical set of clusters, and compare the spatial distributions of the SaTScan and SpatialEpi results.

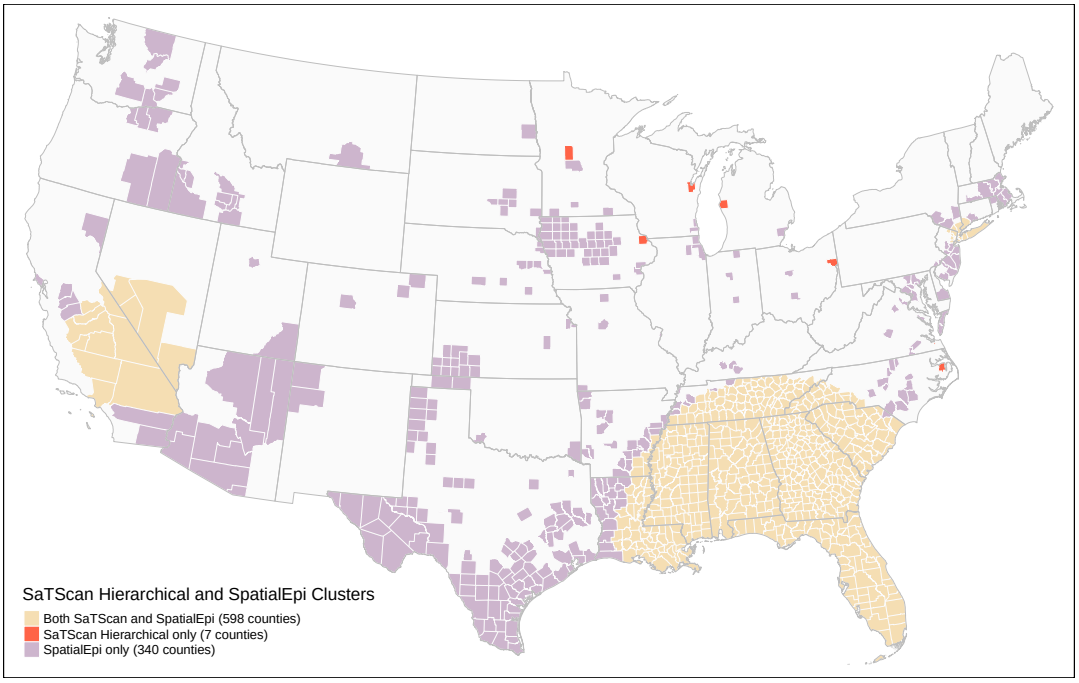
```
## SaTScan combined GIS output has 1306 records with 922 unique county values
```

```
##
##
## SaTScan Hierarchical clusters include 605 records with 605 unique county values
```

```
##
##
## SaTScan GINI-optimized clusters include 701 records with 701 unique county values
```

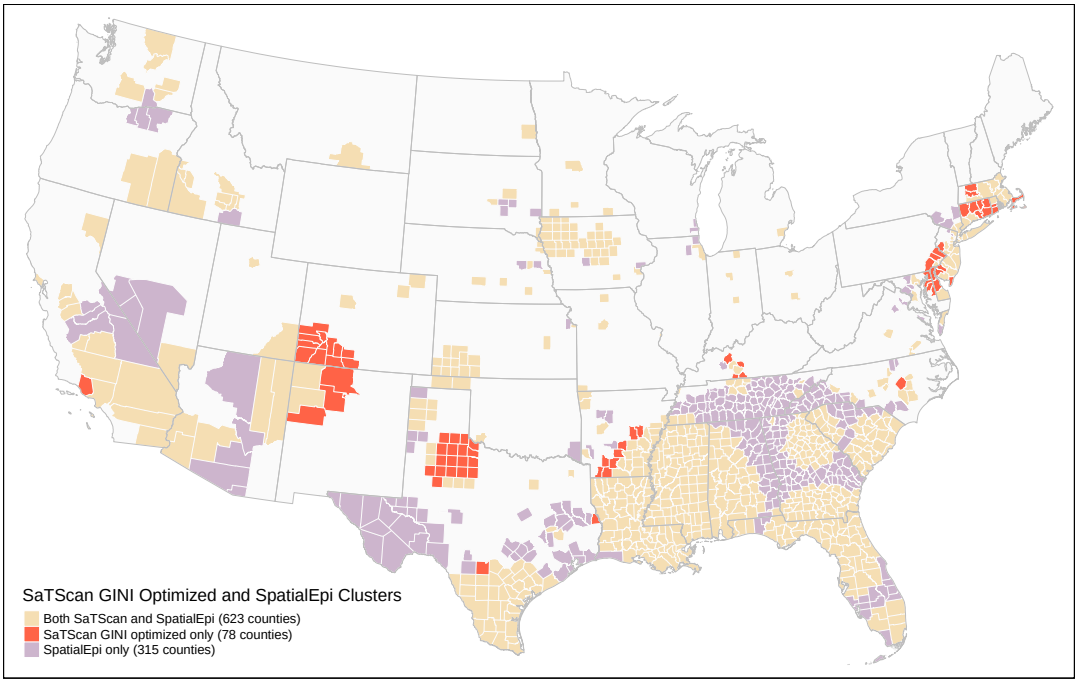
It was necessary to divide the Hierarchical clusters from the GINI clusters to avoid duplicates and geographic overlap.

Compare the SaTScan Hierarchical clusters to the SpatialEpi clusters.



The two methods only agree on the definition of the largest clusters in distant regions. Thereafter, SpatialEpi detects many secondary clusters in the vicinity of the largest ones, while SaTScan detects seven isolated and low-probability counties.

Compare the SaTScan GINI Optimized clusters to the SpatialEpi clusters.



There is more agreement overall between SpatialEpi and SaTScan GINI Optimized clusters. The two algorithms agree the most for smaller and less significant clusters above the 95% confidence threshold. Because the SaTScan clusters are more limited in size, SaTScan detects several smaller clusters with gaps in

Table 1: COVID-19 Risk Class by County

SpatialEpi	SatScan					
	1	2	3	4	5	6
1	2092	78	0	0	0	0
2	306	528	9	0	0	0
3	7	6	69	1	0	0
4	1	3	0	5	0	0
5	1	0	0	0	1	0
6	0	0	0	0	0	1

place of the largest SpatialEpi clusters.

Keeping in mind that the final analysis uses a classification of cluster relative risk for GEE models, are there important differences between the two results with regard to classification of risk? We can check by calculating cluster relative risk classes based on the SaTScan GINI clusters, and cross-tabulating with the SpatialEpi risk classes.

Indeed, SpatialEpi has identified more than 300 counties with above normal risk that were not identified by SaTScan. Meanwhile, SaTScan identified 78 counties with above normal risk that were not identified by SpatialEpi.

The maps and crosstabulation above indicate that there are important differences between the SaTScan and SpatialEpi computational environments for calculating secondary clusters.

We summarize our understanding of the computational differences for default settings below, based on close examination of our software outputs, technical documentation for SaTScan, and the documentation and code repository for SpatialEpi.

	SaTScan Hierarchical Clusters	SaTScan GINI Clusters	SpatialEpi Clusters
possible shapes	circle (default) or ellipse	circle (default) or ellipse	circle
possible cluster centers	locations with rates > normal	locations with rates > normal	all locations
maximum cluster size	50% of cases	varies, not exceeding 50% of cases	50% of population
maximum p of cluster	1.00	1.00	0.05
distance	spherical great circle	spherical great circle	spherical equidistant cylindrical projection

To further interrogate the differences in sets of secondary clusters, we must understand that theoretically each location (county), may be the center of many different circular clusters defined by different radii, starting with a radius of 0 and the one county at the center, and expanding until the maximum cluster size is reached.

SaTScan Hierarchical Clusters

- Select locations (counties) with above-normal COVID incidence rates
- For each location, find log-likelihood of all possible cluster sizes (from minimum of 2 cases to maximum of 50% of cases) with $p < 1$
- For each location, select the cluster with the maximum log-likelihood, resulting in one possible cluster for each location with above-normal COVID incidence rates
- Sort the remaining clusters by log-likelihood from greatest to least
- Select the most likely cluster

- Select secondary clusters by iterating over the remaining clusters, adding new clusters to the set of secondary clusters if they do not geographically overlap any of the clusters already identified as most likely or secondary (pg 68).

SaTScan GINI-Optimized Clusters

- Follow the same procedure as SaTScan Hierarchical clusters, but iterate the procedure with different maximum cluster sizes. By default the cluster sizes include 1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 45, and 50 percent of cases. With the default setting, the result is 17 different sets of clusters.
- For each set of clusters, calculate the GINI coefficient of the COVID-19 incidence inside the clusters vs outside the clusters.
- Select the set of clusters with the highest GINI coefficient (i.e. the most difference between COVID-19 incidence inside the clusters vs outside the clusters).

SpatialEpi Clusters

- For all locations, calculate the log likelihood of all the possible clusters below the maximum cluster size (50% of population)
- Select the clusters with $p < 0.05$
- Sort the remaining clusters by log-likelihood from greatest to least (line 186)
- Select the most likely cluster
- Select secondary clusters by iterating over the remaining clusters, adding new clusters to the set of secondary clusters if they do not geographically overlap any of the clusters already identified as most likely or secondary (line 199).

The differences between these three approaches have very significant impacts on the results (see the differences in results in the **two maps** above) and it is impossible to control for all of the differences with the available parameters. Most fundamentally, SaTScan develops sets of secondary clusters from a universe of just one most likely cluster per location with no default limitation its statistical significance, whereas SpatialEpi may consider multiple possible cluster sizes for each location with a default limitation of maximum 0.05 p for each cluster. These fundamental differences are evident in the spatial distribution of clusters. For example, New York City is the most likely cluster in all analyses. For counties near New York, the radius of the most likely cluster is large and geographically overlaps New York City. Therefore, if only the most significant cluster radius is considered as a possible secondary cluster for counties near New York City, all such clusters are disqualified by their geographical overlap. This is what happens in the SaTScan Hierarchical Clusters model, for which the next nearest clusters are in Ohio and Virginia. In the SaTScan GINI-optimized model, the maximum cluster size is apparently smaller, such that the most likely cluster in New York City is also smaller. This change allows for two other non-overlapping secondary clusters in Rhode Island and New Jersey. In contrast, the SpatialEpi algorithm still considers a variety of possible cluster sizes for each county, allowing for detection of smaller clusters adjacent to more significant ones.

Of course, the *relative risk* score of each cluster is contingent on the cluster size, so each difference in geographic configuration of clusters also impacts the cluster risk classification of individual locations. The most stable results are for the most likely clusters in distinct regions (New York City, Southeast U.S., Southern California & Nevada), while the most variability appears for secondary clusters close enough for their most likely radius to overlap the more likely clusters. The circular shape could be considered a major limitation of Kulldorff cluster detection, for which the SaTScan methodology enhances the limitation by constraining the possibility of nearby clusters while the SpatialEpi methodology can detect smaller adjacent secondary clusters.

The high significance threshold in the default SaTScan analysis allows the inclusion of many small clusters with low likelihoods, adding noise to the results. This could be controlled by overriding the maximum p value parameter. Combining all of the default parameters, SaTSCan includes small clusters of relatively low-risk counties in the Midwest, but excludes relatively high-risk counties adjacent to the major clusters of New York, the Southeast, and Southern California/Nevada. This problem does not exist in the SpatialEpi implementation, and the SpatialEpi parameter *alpha level* parameter cannot be practically increased to `1` to match the SatSCan default. This is because SpatialEpi does not filter counties by those with local relative

risk greater than 1—therefore an *alpha value* of `1` results in *all* counties being included as clusters.

In sum, there are three construct validity issues with the original study COVID-19 high-risk clusters as implemented in SaTScan.

1. Two sets of overlapping secondary clusters are included in the SaTScan output: hierarchical clusters and clusters optimized by GINI coefficient.
2. Only the 96 counties at the center of a cluster are considered in the risk classification.
3. The geographic patterns and cluster relative risk scores of secondary clusters are limited to circles or ellipses and are apparently sensitive to both the geographic shape and situation of high-risk clusters and to subjective decisions in parameters and algorithms.

Preprocess data for GEE modelling

Unplanned deviation for reanalysis: Based on the three observations above, we think that it would be more valid to choose one set of secondary clusters based on a single method rather than combining a set of hierarchical clusters with a set of GINI optimized clusters. We also think that it would be more valid to include risk levels for all counties within a cluster (i.e. all counties within any of the circles above), rather than only the county at the center of a cluster. Finally, we think it would be more valid to treat clusters as a single category rather than five tiers of above-normal risk.

To complete the reproduction/reanalysis study, we will therefore calculate and compare multiple versions of the GEE models:

1. Original study results
2. Original study data in geepack
3. SpatialEpi cluster classification in geepack
4. SpatialEpi binary clusters in geepack

Unique GEE cluster IDs

First, calculate GEE cluster IDs.

We have already calculated: - `rp_clusID` based on our SpatialEpi clusters - `ss_clusID` based on our SaTScan cluster centers, and shown to be identical to the original author's data - `gini_clusID` based on our SaTScan GINI-optimized clusters

Filter and standardize data

Second, filter the data for non-zero COVID-19 rates and z-score standardize the independent variables. This accomplishes step 10 of the workflow diagram.

Unplanned deviation for reproduction: We assumed that we should filtered for COVID rates > 0 first and then calculate z-scores, however after comparing data in the next code block, we realized that the original study had *first* calculated z-scores and *then* filtered for COVID rates > 0. Therefore, we placed the z-score standardization first and the filtering next in the following codeblock.

Compare independent variables for GEE models by subtracting the original values from the reproduced values, and finding the average and standard deviation of difference for each variable.

```
## Summary of difference between reproduction independent variables and original independent variables

## Mean:

##              z_white_pct              z_black_pct             z_native_pct
##                        0                        0                        0
##              z_asian_pct              z_other_pct     z_non_hisp_white_pct
##                        0                        0                        0
##               z_hisp_pct z_non_hisp_non_white_pct               z_bpov_pct
##                        0                        0                        0
```

```
##               z_apov_pct               z_pct_5_17             z_pct_18_34
##                        0                        0                       0
##             z_pct_35_64             z_pct_65_74                z_pct_75
##                        0                        0                       0
##              z_male_pct            z_female_pct
##                        0                        0
```

```
##
## Standard deviation:
```

```
##              z_white_pct              z_black_pct            z_native_pct
##                    0.000                    0.000                   0.000
##              z_asian_pct              z_other_pct    z_non_hisp_white_pct
##                    0.001                    0.001                   0.000
##               z_hisp_pct z_non_hisp_non_white_pct              z_bpov_pct
##                    0.000                    0.000                   0.000
##               z_apov_pct               z_pct_5_17             z_pct_18_34
##                    0.000                    0.001                   0.001
##              z_pct_35_64             z_pct_65_74                z_pct_75
##                    0.000                    0.000                   0.000
##              z_male_pct            z_female_pct
##                    0.000                    0.000
```

When we had first filtered for COVID rates > 0 first and then z-score standardized second, the means of differences ranged from -0.012 to 0.004, and standard deviations of differences ranged from 0.000 to 0.016.

After changing the order to first z-score standardize and then filter for COVID rates > 0, we observed no mean difference between our reproduced variables and the original variables, and we find no standard deviation > 0.001 for the difference between reproduction independent variables and original variables. There are no major differences between the independent variables.

Save final derived data Optionally, you may save the preprocessed data to `data/raw/public/gee_data.gpkg`

Optionally, you may load the preprocessed data from `data/raw/public/gee_data.gpkg`

GEE models

This accomplishes the step 11 of the workflow diagram

Generalized Estimating Equation parameters:

"The **'exchangeable' correlation matrix** was selected for the results reported here, since this specification yielded the best statistical fit based on the QIC (quasi- likelihood under the independence) model criterion." (Chakraborty 2021, Methods paragraph 5)

"The **gamma distribution** with **logarithmic link function** was chosen for all GEEs since this model specification provided the lowest QIC value." (Chakraborty 2021, Methods paragraph 5)

Useful Reference: https://data.library.virginia.edu/getting-started-with-generalized-estimating-equations/

GEE Function

Define a function for calculating and summarizing five GEE models

Original Clusters and Original COVID-19 Rate

Calculate GEE models with: - Clustering: SaTScan cluster centers & State ID - Dependent variable: original COVID-19 incidence (including errors).

Table 3: Original Cluster IDs and Original COVID-19 Incidence

	estimate	std.error	conf.low	conf.high	statistic	stars	p.value
race model intercept	6.757	0.020	6.718	6.796	115250.600	2	0.000
z_white_pct	-0.201	0.026	-0.251	-0.150	61.397	2	0.000
z_black_pct	0.281	0.020	0.242	0.321	191.476	2	0.000
z_native_pct	0.019	0.023	-0.025	0.064	0.703	0	0.402
z_asian_pct	0.064	0.024	0.016	0.112	6.944	2	0.008
z_other_pct	0.112	0.019	0.076	0.149	36.371	2	0.000
ethnicity model intercept	6.742	0.020	6.704	6.780	118274.177	2	0.000
z_non_hisp_white_pct	-0.226	0.026	-0.277	-0.175	76.367	2	0.000
z_hisp_pct	0.137	0.017	0.105	0.170	67.957	2	0.000
z_non_hisp_non_white_pct	0.278	0.019	0.241	0.315	217.468	2	0.000
poverty status model intercept	6.816	0.021	6.774	6.858	101377.567	2	0.000
z_bpov_pct	0.240	0.027	0.186	0.293	76.612	2	0.000
z_apov_pct	-0.303	0.030	-0.362	-0.245	102.166	2	0.000
age model intercept	6.820	0.021	6.779	6.862	102327.297	2	0.000
z_pct_5_17	0.069	0.028	0.015	0.123	6.185	1	0.013
z_pct_18_34	0.029	0.034	-0.038	0.097	0.723	0	0.395
z_pct_35_64	0.015	0.043	-0.069	0.099	0.127	0	0.722
z_pct_65_74	-0.044	0.042	-0.126	0.039	1.082	0	0.298
z_pct_75	-0.169	0.029	-0.227	-0.112	33.215	2	0.000
biological sex model intercept	6.814	0.021	6.772	6.855	105802.891	2	0.000
z_male_pct	-0.402	0.042	-0.484	-0.319	90.998	2	0.000
z_female_pct	0.298	0.044	0.212	0.385	45.558	2	0.000

Original Clusters and Fixed COVID-19 Rate

Calculate GEE models with: - Clustering: SaTScan cluster centers & State ID - Dependent variable: reproduced COVID-19 incidence (fixed errors).

GINI Clusters and Fixed COVID-19 Rate

Calculate GEE models with: - Clustering: Reproduced SaTScan GEE clusters & State ID - Dependent variable: reproduced COVID-19 incidence (fixed errors).

SpatialEpi Clusters and Fixed COVID-19 Rate

Calculate GEE models with: - Clustering: Reproduced SpatialEpi clusters & State ID - Dependent variable: reproduced COVID-19 incidence (fixed errors).

Compare GEE results

Load digitized version of Table 2 from the original publication.

To Do

Subtract the results of GEE models for with reproduced COVID-19 incidence (fixed errors) from original COVID-19 incidence (with errors).

First, here are two notes about interpreting the difference between models: - A positive number for the difference of coefficients indicates that the reproduction found a greater coefficient. - A positive number for the p-value indicates that the reproduction found a larger p-value, for a *less statistically signficant* result.

Table 4: Original Cluster IDs and Reproduced COVID-19 Incidence

	estimate	std.error	conf.low	conf.high	statistic	stars	p.value
race model intercept	6.755	0.020	6.716	6.794	114890.669	2	0.000
z_white_pct	-0.213	0.024	-0.261	-0.165	75.656	2	0.000
z_black_pct	0.284	0.020	0.244	0.325	192.591	2	0.000
z_native_pct	0.016	0.023	-0.029	0.062	0.500	0	0.480
z_asian_pct	0.055	0.028	0.001	0.110	3.920	1	0.048
z_other_pct	0.109	0.019	0.073	0.146	34.651	2	0.000
ethnicity model intercept	6.740	0.020	6.702	6.779	117978.135	2	0.000
z_non_hisp_white_pct	-0.238	0.024	-0.285	-0.191	99.695	2	0.000
z_hisp_pct	0.132	0.016	0.100	0.164	65.165	2	0.000
z_non_hisp_non_white_pct	0.281	0.019	0.243	0.318	211.893	2	0.000
poverty status model intercept	6.816	0.021	6.774	6.858	100871.356	2	0.000
z_bpov_pct	0.241	0.028	0.186	0.295	75.725	2	0.000
z_apov_pct	-0.311	0.030	-0.370	-0.252	106.800	2	0.000
age model intercept	6.820	0.021	6.779	6.862	102219.285	2	0.000
z_pct_5_17	0.074	0.028	0.019	0.130	6.889	2	0.009
z_pct_18_34	0.026	0.035	-0.042	0.095	0.573	0	0.449
z_pct_35_64	0.009	0.043	-0.075	0.092	0.042	0	0.838
z_pct_65_74	-0.041	0.042	-0.124	0.041	0.960	0	0.327
z_pct_75	-0.176	0.029	-0.233	-0.118	35.495	2	0.000
biological sex model intercept	6.815	0.021	6.774	6.856	105760.114	2	0.000
z_male_pct	-0.401	0.042	-0.484	-0.319	90.775	2	0.000
z_female_pct	0.292	0.044	0.206	0.377	44.805	2	0.000

Table 5: Reproduced SaTScan GINI Cluster IDs and Reproduced COVID-19 Incidence

	estimate	std.error	conf.low	conf.high	statistic	stars	p.value
race model intercept	6.769	0.020	6.730	6.809	111949.365	2	0.000
z_white_pct	-0.210	0.024	-0.258	-0.162	73.302	2	0.000
z_black_pct	0.278	0.018	0.242	0.314	231.024	2	0.000
z_native_pct	0.028	0.021	-0.013	0.070	1.798	0	0.180
z_asian_pct	0.054	0.021	0.014	0.094	6.858	2	0.009
z_other_pct	0.098	0.017	0.065	0.130	34.042	2	0.000
ethnicity model intercept	6.753	0.020	6.714	6.793	113392.736	2	0.000
z_non_hisp_white_pct	-0.236	0.024	-0.284	-0.189	95.116	2	0.000
z_hisp_pct	0.127	0.016	0.096	0.158	64.698	2	0.000
z_non_hisp_non_white_pct	0.272	0.018	0.237	0.307	230.689	2	0.000
poverty status model intercept	6.833	0.022	6.789	6.876	95331.421	2	0.000
z_bpov_pct	0.243	0.025	0.195	0.292	98.521	2	0.000
z_apov_pct	-0.303	0.030	-0.361	-0.244	102.163	2	0.000
age model intercept	6.840	0.022	6.798	6.882	100886.810	2	0.000
z_pct_5_17	0.071	0.025	0.022	0.120	8.025	2	0.005
z_pct_18_34	0.023	0.030	-0.036	0.083	0.579	0	0.447
z_pct_35_64	0.026	0.036	-0.044	0.096	0.529	0	0.467
z_pct_65_74	-0.040	0.037	-0.113	0.032	1.199	0	0.273
z_pct_75	-0.179	0.029	-0.235	-0.122	38.781	2	0.000
biological sex model intercept	6.833	0.021	6.791	6.874	103348.018	2	0.000
z_male_pct	-0.399	0.042	-0.481	-0.317	90.805	2	0.000
z_female_pct	0.297	0.043	0.214	0.381	48.721	2	0.000

Table 6: Reproduced SpatialEpi Cluster IDs and Reproduced COVID-19 Incidence

	estimate	std.error	conf.low	conf.high	statistic	stars	p.value
race model intercept	6.755	0.020	6.716	6.795	112262.225	2	0.000
z_white_pct	-0.204	0.024	-0.250	-0.157	74.622	2	0.000
z_black_pct	0.265	0.021	0.225	0.305	166.023	2	0.000
z_native_pct	0.021	0.023	-0.024	0.065	0.814	0	0.367
z_asian_pct	0.051	0.020	0.011	0.090	6.397	1	0.011
z_other_pct	0.086	0.018	0.050	0.122	22.211	2	0.000
ethnicity model intercept	6.743	0.020	6.704	6.781	115038.889	2	0.000
z_non_hisp_white_pct	-0.230	0.024	-0.276	-0.184	95.542	2	0.000
z_hisp_pct	0.115	0.016	0.083	0.146	51.343	2	0.000
z_non_hisp_non_white_pct	0.267	0.019	0.229	0.305	191.127	2	0.000
poverty status model intercept	6.817	0.021	6.775	6.859	101003.814	2	0.000
z_bpov_pct	0.219	0.029	0.162	0.277	56.181	2	0.000
z_apov_pct	-0.285	0.031	-0.346	-0.224	83.714	2	0.000
age model intercept	6.823	0.022	6.781	6.866	100526.111	2	0.000
z_pct_5_17	0.063	0.023	0.018	0.108	7.655	2	0.006
z_pct_18_34	0.032	0.035	-0.037	0.101	0.840	0	0.359
z_pct_35_64	-0.004	0.041	-0.085	0.076	0.012	0	0.914
z_pct_65_74	-0.025	0.041	-0.105	0.055	0.375	0	0.540
z_pct_75	-0.160	0.027	-0.214	-0.107	34.839	2	0.000
biological sex model intercept	6.816	0.021	6.774	6.858	102169.185	2	0.000
z_male_pct	-0.368	0.039	-0.445	-0.292	89.145	2	0.000
z_female_pct	0.266	0.041	0.185	0.347	41.502	2	0.000

Table 7: Original Publication Table 2

	beta	std_error	lower_95_ci	upper_95_ci	wald_chi_square	p_stars
Race model intercept	7.106	0.083	6.997	7.322	7465.214	2
White	-0.203	0.020	-0.242	-0.164	102.958	2
Black	0.111	0.016	0.079	0.143	46.214	2
Native American	0.051	0.009	0.033	0.069	31.438	2
Asian	0.080	0.018	0.046	0.115	21.060	2
Other race	0.077	0.017	0.044	0.115	21.030	2
Ethnicity model intercept	7.186	0.083	7.023	7.348	7525.648	2
Non-Hispanic White	-0.237	0.022	-0.280	-0.194	116.954	2
Hispanic	0.119	0.031	0.058	0.180	14.708	2
Non-Hispanic non-White	0.118	0.016	0.086	0.149	53.248	2
Poverty model intercept	7.183	0.072	7.043	7.324	10066.930	2
Below poverty level	0.148	0.022	0.105	0.190	46.913	2
Above poverity level	-0.267	0.023	-0.312	-0.222	134.297	2
Age model intercept	7.242	0.076	7.093	7.391	9093.179	2
Age 5-17	0.047	0.016	0.016	0.078	8.732	2
Age 18-34	0.038	0.022	-0.005	0.081	3.008	0
Age 35-64	-0.026	0.023	-0.071	0.019	1.300	0
Age 65-74	-0.089	0.021	-0.131	-0.047	17.597	2
Age 75 or more	-0.108	0.020	-0.148	-0.069	29.479	2
Biological sex model intercept	7.223	0.072	7.081	7.365	9963.672	2
Male	-0.298	0.028	-0.353	-0.243	113.460	2
Female	0.153	0.029	0.097	0.209	28.577	2

Table 8: Reproduced model minus Original model

	estimate	std.error	conf.low	conf.high	statistic	stars
race model intercept	-0.349	-0.063	-0.279	-0.526	107785.386	0
z_white_pct	0.002	0.006	-0.009	0.014	-41.561	0
z_black_pct	0.170	0.004	0.163	0.178	145.262	0
z_native_pct	-0.032	0.014	-0.058	-0.005	-30.735	-2
z_asian_pct	-0.016	0.006	-0.030	-0.003	-14.116	0
z_other_pct	0.035	0.002	0.032	0.034	15.341	0
ethnicity model intercept	-0.444	-0.063	-0.319	-0.568	110748.529	0
z_non_hisp_white_pct	0.011	0.004	0.003	0.019	-40.587	0
z_hisp_pct	0.018	-0.014	0.047	-0.010	53.249	0
z_non_hisp_non_white_pct	0.160	0.003	0.155	0.166	164.220	0
poverty status model intercept	-0.367	-0.051	-0.269	-0.466	91310.637	0
z_bpov_pct	0.092	0.005	0.081	0.103	29.699	0
z_apov_pct	-0.036	0.007	-0.050	-0.023	-32.131	0
age model intercept	-0.422	-0.055	-0.314	-0.529	93234.118	0
z_pct_5_17	0.022	0.012	-0.001	0.045	-2.547	-1
z_pct_18_34	-0.009	0.012	-0.033	0.016	-2.285	0
z_pct_35_64	0.041	0.020	0.002	0.080	-1.173	0
z_pct_65_74	0.045	0.021	0.005	0.086	-16.515	-2
z_pct_75	-0.061	0.009	-0.079	-0.043	3.736	0
biological sex model intercept	-0.409	-0.051	-0.309	-0.510	95839.219	0
z_male_pct	-0.104	0.014	-0.131	-0.076	-22.462	0
z_female_pct	0.145	0.015	0.115	0.176	16.981	0

Table 9: Reproduced model minus Original model

	estimate	std.error	conf.low	conf.high	statistic	stars	p.value
race model intercept	-0.002	0.000	-0.002	-0.002	-359.931	0	0.000
z_white_pct	-0.012	-0.002	-0.010	-0.015	14.259	0	0.000
z_black_pct	0.003	0.000	0.002	0.004	1.115	0	0.000
z_native_pct	-0.003	0.000	-0.004	-0.002	-0.203	0	0.078
z_asian_pct	-0.009	0.004	-0.015	-0.002	-3.024	-1	0.040
z_other_pct	-0.003	0.000	-0.003	-0.003	-1.720	0	0.000
ethnicity model intercept	-0.002	0.000	-0.002	-0.001	-296.042	0	0.000
z_non_hisp_white_pct	-0.012	-0.002	-0.008	-0.016	23.328	0	0.000
z_hisp_pct	-0.005	-0.001	-0.005	-0.006	-2.792	0	0.000
z_non_hisp_non_white_pct	0.003	0.000	0.002	0.003	-5.575	0	0.000
poverty status model intercept	0.000	0.000	0.000	0.000	-506.211	0	0.000
z_bpov_pct	0.001	0.001	0.000	0.002	-0.887	0	0.000
z_apov_pct	-0.008	0.000	-0.008	-0.007	4.634	0	0.000
age model intercept	0.000	0.000	0.000	0.000	-108.012	0	0.000
z_pct_5_17	0.005	0.000	0.004	0.007	0.704	1	-0.004
z_pct_18_34	-0.003	0.001	-0.004	-0.002	-0.150	0	0.054
z_pct_35_64	-0.006	0.000	-0.006	-0.007	-0.085	0	0.116
z_pct_65_74	0.003	0.000	0.002	0.002	-0.122	0	0.029
z_pct_75	-0.007	0.000	-0.006	-0.006	2.280	0	0.000
biological sex model intercept	0.001	0.000	0.002	0.001	-42.777	0	0.000
z_male_pct	0.001	0.000	0.000	0.000	-0.223	0	0.000
z_female_pct	-0.006	0.000	-0.006	-0.008	-0.753	0	0.000

As a result of fixing the COVID-19 incidence rate errors, the significance level of one variable (Asian people with disabilities) decreased. The statistical significance of the Age 35-64 group increased by 0.08, but the variable remained above the 0.05 significance level.

Overall, the impact of errors in COVID-19 incidence rates for 13 counties appears to have been minimial in the context of an overall sample size of 3108 counties.