Preanalysis Plan for Rpr-Chakraborty-2021: Reproduction of Social Inequities in the distribution of COVID-19: An intra-categorical analysis of people with disabilities in the U.S.

Joseph Holler, Department of Geography, Middlebury College, Middlebury VT 05753
Derrick Burt, Department of Geography, Middlebury College, Middlebury VT 05753
Drew An-Pham, Department of Geography, Middlebury College, Middlebury VT 05753
Peter Kedron, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe AZ 85281
Junyi Zhou, Department of Geography, Middlebury College, Middlebury VT 05753

Version 1.1 | Created Jul 7, 2021 | Last Updated June 2, 2022

Reproduction of: 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

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 study aims to examine public concern that persons with disabilities (PwDs) face disproportionate challenges due to COVID-19. To investigate this, Chakraborty examines the statistical relationship between confirmed county-level COVID-19 case rates and county-level socio-demographic and disability variables. Specifically, Chakraborty tests county-level bivariate correlations between COVID-19 incidence against the percentage of disability and socio-demographic category, with a separate hypothesis and model for each subcategory within disability, race, ethnicity, age, and biological sex. To control for differences between states and geographic clusters of COVID-19 outbreaks, Chakraborty uses five generalized estimating equation (GEE) models to predict the relationship and significance between COVID-19 incidence and disability subgroups within each socio-demographic category while considering inter-county spatial clusters. Chakraborty (2021) finds significant positive relationships between COVID-19 rates and socially vulnerable demographic categories of race, ethnicity, poverty, 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 reproduction 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. To the greatest extent possible, the reproduction will be implemented with R markdown using packages geepack for the generalized estimating equation and SpatialEpi for the spatial scan statistics.

Keywords

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

Study design

The reproduction study will try 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 figure 1, table 1, and table 2 of Chakraborty (2021). 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

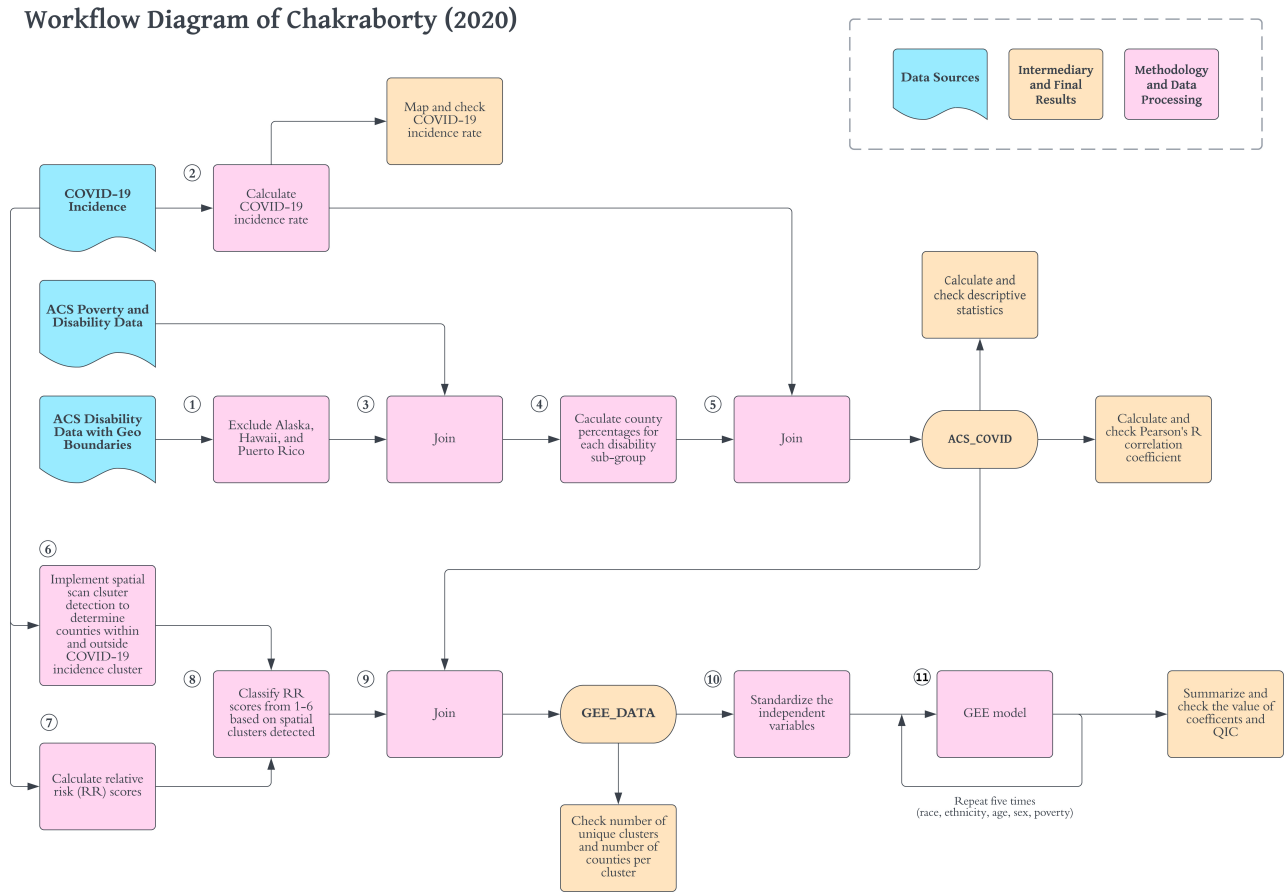
The original study is **observational**, with the **exploratory** objective of determining "whether COVID-19 incidence is significantly greater in counties containing higher percentages of socio-demographically disadvantaged [people with disabilities], based on their race, ethnicity, poverty status, 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.

The study was originally conducted using SaTScan software (unspecified version) to implement the spatial scan statistic. Other software are not specified in the publication; however data files and communication with the author show that spatial analysis and mapping was conducted in ArcGIS and statistics were calculated in SPSS.

Our understanding of the original study design and our plan for the reproduction analysis are visualized in the workflow diagram.



Sampling plan

Existing data and data exploration

This registration is based upon a thorough reading of the original research article, searching and calculating summary statistics for American Community Survey data, accessing the Johns Hopkins Coronavirus Resource Center, and acquiring some additional information and data from the original author, Jay Chakraborty. Specifically, Chakraborty informed us of the American Community Survey data table names used in the study (S1810 for demographic categories and disability status and C18130 for poverty status and disability status), provided Johns Hopkins county-level Coronavirus data downloaded on August 1, 2020, outputs from SaTScan spatial clustering analysis, and inputs for the GEE models. The data provided by the author is not available in an online repository, but we will include the data in our research compendium with permission of the author.

In our reproduction attempt, we will use publicly available American Community Survey data downloaded directly from the Census API using the tidycensus package for R. We will use Johns Hopkins Coronavirus data provided by the author because it is not possible to download that dynamic dataset in an archived form as it existed on August 1, 2020. Johns Hopkins still provides aggregated COVID-19 incidence rate data, but does not publicly provide archived data identical to those used in the original study. This pre-analysis plan is based on information from the original paper, correspondence with the original author (as described above), viewing metadata and data sources provided by the author and the U.S. Census, and calculating summary statistics.

Disclaimer: For demonstration purposes, we are registering this plan *after* running the full analysis in R-studio, based upon our documented plan and knowledge of the study prior to completing the analysis.

Data collection and spatial sampling

The study exclusively uses secondary data sources. The study does not sample from the secondary data sources.

The published results are based of COVID-19 cases reported at the county-level and this is not a sampled dataset. The disability data from the ACS are collected at the county level. Details on the data collection can be found at https://www.census.gov/topics/health/disability/guidance/data-collection-acs.html and details on sampling methods can be found at https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html.

Data description

Although the data specifications are described in detail in the original study, none of the data from the original study is provided in an online repository.

We received the COVID-19 case data from 8/1/2020 from the author, as there is no readily apparent way to access archived data from the Johns Hopkins University Center for Systems Science Engineering database. The COVID-19 case data expresses cumulative count of reported COVID-19 from 1/22/2020 to 8/1/2020. The data can be found at the John Hopkins CCSE COVID-19 Data Repository (https://github.com/CSSEGISandData/COVID-19). However, archived data only provides summaries at the national scale.

The 2018 ACS 5 year estimates for disabilities can be accessed from the U.S. Census website or through the Census API.

Variables

All variables in this study were derived from secondary data. There are no experimentally manipulated variables in this experiment. Eighteen independent variables, a percentage of total disabled persons per county and seventeen 'disaggregated' categories that break down socio-demographic characteristics of the disabled population. COVID-19 incidence rate can be seen as the dependent variables.

The socio-demographic variables are broken down into the following categories. Their table code from the ACS data has been included in this documentation

COVID-19 incidence rate

COVID-19 Incidence is calculated as the number of known cases per 100,000 people, based upon the Johns Hopkins University COVID-19 Resource Center database.

Persons with disabilities

The American Community Survey (ACS) variables used in the study are outlined below.

Table 1: Disability Subgroup Variables

Variable Name in Study	ACS Variable name
percent of total civilian non-institutionalized population with a disability	S1810_C03_001E
Race	
percent w disability: White alone	S1810_C03_004E

Variable Name in Study	ACS Variable name
percent w disability: Black alone	S1810_C03_005E
percent w disability: Native American	S1810_C03_006E
percent w disability: Asian alone	S1810_C03_007E
percent w disability: Other race	S1810_C03_009E
Ethnicity	
percent w disability: Non-Hispanic White	S1810_C03_0011E
percent w disability: Hispanic	S1810_C03_012E
percent w disability: Non-Hispanic non-White	(S1810_C02_001E - S1810_C02_011E - S1810_C02_012E) / (S1810_C01_001E - S1810_C01_011E - S1810_C01_012E) * 100
percent w disability: Other race	S1810_C03_009E
Poverty	
percent w disability: Below poverty level	(C18130_004E + C18130_011E + C18130_018E) / C18130_001E * 100
percent w disability: Above poverty level	(C18130_005E + C18130_012E + C18130_019E) / C18130_001E * 100
Age	
percent w disability: 5-17	S1810_C03_014E
percent w disability: 18-34	S1810_C03_015E
percent w disability: 35-64	S1810_C03_016E
percent w disability: 65-74	S1810_C03_017E
percent w disability: 75+	S1810_C03_018E
Biological sex	
percent w disability: male	S1810_C03_001E
percent w disability: female	S1810_C03_003E

Attribute variable transformations

The COVID-19 incidence rate is normalized at the county-level per 100,000 people. Most of the disability and sociodemographic variables are provided in the format that they are used, as a percentage of "people with disabilities in each subgroup by the total civilian non-institutionalized population relevant to the variable category" (Chakraborty 2011). Non-Hispanic non-White, Below poverty level and Above poverty level are calculated as shown in Table 1 above.

Before conducting the GEE, all independent variables are normalized into z-scores.

For the GEE, two different clustering scores are assigned to each county. The first clustering ID is just a categorical variable determined by the counties' state. The second clustering ID is a relative risk score calculated by identifying spatial clusters from a spatial scan statistic based on the Poisson Model. We will calculate the clusters using the SpatialEpi package in R. We then calculate

the relative risk score for each county using the formula: `(rate of cases within the cluster) / (rate of cases outside the cluster)`. The relative risk score is then classified into six categories based on the estimated relative risk values (<1.0, 1.00-1.99, 2.00-2.99, 3.00-3.99, 4.00-4.99, and 5.0 or more). The first clustering ID (State) and second clustering score (Classified Relative Risk) are combined to form IDs for each unique combination of state and relative risk class. The clustering ID's will then be joined with the American Community Survey data on disability subgroups to be used as input to the GEE models.

Geographic transformations

Although there are no explicit geographic transformations in this experiment, the variable transformations that occur during the SaTScan procedure are geographic in nature: they assign values based on spatial clustering of COVID-19 risk, which are subsequently used to define clusters in the GEE models.

Having looked at the SaTScan outputs from the original study, our best guess is that the author might have calculated centroids for each county before running the GEE, using a geographic coordinate system.

Analyses

Geographical characteristics

The **coordinate reference system** is not specified in the methodology. Census data is provided in the NAD1983 Geographic Coordinate System. We assume that the analysis was also conducted the NAD1983 Geographic Coordinate System because the SaTScan can perform a spherical distance calculation using latitude and longitude.

The **spatial extent** of the study were the contiguous 49 United States (including the District of Columbia).

The **spatial scale** and **unit of analysis** of the study is are U.S. counties.

Edge effects will not be accounted for in the analysis.

This analysis does create **spatial subgroups** based on **spatial clustering**. The purpose of this grouping is to control for **spatial heterogeneity** between regions (defined as states) and spatial correlation within regions. There are criteria for two different types of spatial clustering; we address these in the attribute variable transformation section.

This analysis does not measure or account for any **first order spatial effects**, **second order spatial effects**, or **spatial anisotropies**.

Temporal characteristics

The **temporal extent** of the study is based on the COVID-19 incidence rate, which covers cases from 1/22/2020-8/1/2020. The study also uses 5 year estimates for county disability and sociodemographic characteristics collected from 2014-2018. This range is not explicitly stated in the original study.

The **temporal support** for the COVID-19 incidence rate was case data collected from 1/22/2020-8/1/2020. The **temporal support** for the disability sociodemographic data was data collected from 2014-2018. **Temporal effects** are not measured or accounted for.

Data exclusion

There is no documentation of any **data exclusion** based on attribute criteria in the original study.

The study does not analyze the presence of **outliers**. The study does not **weight samples**.

Analytical specification

The county-level Pearson's rho correlation coefficient is used to test association between intra-categorical rates of disability and COVID-19 incidence rates. As this is a parametric test, normality should be tested. A separate hypothesis is formulated for each sociodemographic disability characteristic.

The generalized estimating equation (GEE) models are used to test association between intra-categorical rates of disability and COVID-19 incidence rates while accounting for spatial clustering. As specified by the author, "GEEs extend the generalized linear model to accommodate clustered data, in addition to relaxing several assumptions of traditional regression (i.e., normality)". Additionally, the author notes that "clusters of observations must be defined based on the assumption that observations within a cluster are correlated while observations from different clusters are independent." Following Chakraborty, all five GEE models will be specified with exchangeable correlation matrices, gamma distributions, and logarithmic link function. These specifications were chosen after testing each alternative and choosing the models with the best quasilikelihood under the independence model criterion (QIC).

Inference criteria and robustness

Bivariate inference will be assessed with correlation coefficients and p-values.

Multivariate inference will not be made because GEE models provide estimated coefficients for the independent variables and these are best interpreted as exploratory estimates. Model fit will be assessed with QIC. Statistical significance of independent variable coefficients will be tested with Wald Chi Square and assessed for the 0.01 and 0.05 confidence levels.

Exploratory analyses and contingency planning

There are no **exploratory** analyses in this analysis. There is no need for a **contingency plan** in this study.

Reproduction study design

Planned differences from the original study

We plan to implement the analysis to the greatest extent possible in R / RStudio, using the geepack package for the generalized estimating equation and SpatialEpi package for the spatial scan statistics, whereas the original study was conducted using ArcGIS (Desktop v 10.7), SPSS, and SaTScan (v9.6).

We will plan to check the normality of our distribution of our independent variables before correlations. If they are not normal, we may choose to calculate the bivariate correlation using a Spearman's Rho.

Evaluating the reproduction results

Before comparing results from our reproduction of the statistical models, we plan to compare summary statistics for all of our independent variables to those of the original study to confirm we are using the same inputs. We will compare the summary statistics (in table 1) and geographic distribution (in figure 1) of the dependent variable, COVID-19 incidence.

Considering that we will use a different computational environment from the original authors, we will compare the bivariate correlation coefficients and significance levels expecting extremely similar coefficients and p-values.

Considering that both the computational environment and some analytical decisions will vary in our reproduction of the clusters for GEE modeling, we will compare the coefficients and significance levels with expectation that the direction and significance level will be the identical, but magnitudes and Chi Square values may vary.

In order to test the results for both our bivariate correlation and GEE, we plan to construct tables (or matrices) that show the difference between our correlation coefficients and the original study's correlation coefficients. If there are any non-zeroes, we will investigate further.

We will consider the reproduction an exact reproduction only if we can create identical coefficients for the Pearson's Rho bivariate tests of table 1 and the GEE models of table 2. We will consider the reproduction to be approximate if we find coefficients with the same direction and significance levels as the original study. We will consider the reproduction to have at least partially failed if we find coefficients with different directions or significance levels.

References

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