Replication of a Research Claim from Fielding-Miller et al. (2020),
from medRxiv

Replication Team: Kent Jason Cheng and Radoslaw Panczak

Research Scientist: Nick Fox

Action Editor: Gustav Nilsonne

Independent Reviewers
(add name below when you initiate review, comment “DONE” on your name when you finish):

Reviewer #1: [Samuel Smith]

Reviewer #2: Michael Mullarkey

Reviewer #3: [NAME]

Review Period: October 28 - November 2

View-only links to: Original Paper, Replication Data, Replication Analysis

Privacy Statement: Other teams are making predictions about the outcomes of many different studies, not knowing which studies have been selected for replication. As a consequence, the success of this project requires full confidentiality of this peer review process. This includes privacy about which studies have been selected for replication and all aspects of the discussion about these replication designs.

Instructions for Data Analysts

The preregistration for this replication study was started by a separate team of researchers who were responsible for identifying data sources and constructing them into a replication dataset(s) for your use in the analysis. They have completed sections 1-13 of the preregistration below, and included additional materials in the OSF project that document how the dataset was constructed.

In cases where all of the underlying data sources were able to be freely shared and posted, the constructed dataset(s) have been posted to the OSF as well, which you are free to use in designing the analysis plan (see below for details). In cases where some or all of the data sources could *not* be freely shared or posted, the replication dataset(s) are not provided on the OSF. Rather, you will need to follow the instructions and code to first reconstruct the datasets, and then proceed with your work. In such cases, the team responsible for creating the dataset(s) has provided summary statistics in the OSF that correspond to the constructed datasets, so you can verify that the datasets you create match what they intended.

You'll be responsible for filling out sections 16-25 of the preregistration below. Before you do so, **please review the original study, sections 1-15 of the preregistration, and the materials provided on the OSF**, so that you are familiar with all of the decisions that have been made to date. In many cases, the 'data preparer' will have left you instructions and suggestions on how the provided data can be used in the analysis, as well as idiosyncrasies and discrepancies in the data that you should be aware of. The data preparers have tried to be thorough in including all variables that you might need, but please keep in mind the following:

- Some of the variables included in the constructed dataset(s) may not be needed in the final analysis, so please do not feel the need to necessarily use all of the provided variables.
- Some of the variables needed might have mistakenly been excluded from the constructed datasets. If you find that this is the case, please let Andrew or Anna know, and they will work with you to supplement the datasets as needed.

For these secondary data replications, we would like the analysis plan to be completed before the preregistration goes through review, so that after review, the only remaining steps are registration and running the analysis code on the full datasets. To facilitate that, we are asking that you include in section 19 a link to the code you will use that takes the constructed dataset(s) provided to you and produces the focal analysis (including all of the cleaning, merging, and transforming required). When developing your analysis plan and code, please randomly sample 5% of the data for use in your work and demonstrate that the focal analysis produces sensible results using just that random sample (see section 19 for details). **Do not use the rest of the data until after your study is registered and it is time to run the final analysis**. In section 19, you will find a statement that we are asking you to bold that confirms you've only used 5% of the data when developing and testing your code. If this approach will not work for any reason, please let Andrew or Anna know and disclose deviations from this plan somewhere in the preregistration.

- In cases where we are providing you a complete dataset, you can just sample out 5% of the observations and hold the rest out until you are ready to perform the final analysis.
- In cases where we are providing you multiple datasets that need to be combined prior to analysis, please sample out 5% of the observations in whatever way is most sensible.
    - For example, in cases where each dataset contains complete observations on its own (a typical 'row bind' situation), it makes the most sense to sample out 5% of each dataset separately and then combine them together to develop and test your code.

    - In cases where datasets need to be merged in order to create complete observations (a typical 'column bind' situation), it makes the most sense to merge the separate datasets into a full dataset first, and then sample out the 5% before proceeding with the rest of the analysis code.
- We leave the decision on how to sample out the random subset of data to you, so long as (a) you are not performing any analyses on the complete dataset until after your study is registered and (b) whatever decision you make is documented in the preregistration.

Finally, in cases where the replication data combines observations from the original study with observations that were not used in the original study (what we are calling 'hybrid replications'), please perform up to three analyses (details immediately below). This will likely require you to subset your data, based on the description of the original analysis provided in the study.

- When the 'new' data alone can clear the minimum power threshold, please perform three analyses: one analysis that only uses data that was not used in the original analysis (the focal analysis); one analysis that combines all available data; and a third analysis that only uses data that was used in the original analysis. Please make sure all three analyses are documented (with code) in section 19 below.
- When the 'new' data alone *cannot* clear the minimum power threshold, please perform one analysis that combines all available data, and a second that only uses the old data. Please make sure both analyses are documented (with code) in section 19 below.

Please contact Andrew or Anna if you have any questions. After you've completed the remaining sections of the preregistration and uploaded all the necessary materials to the OSF, please contact the SCORE coordinators regarding next steps.

Preregistration of Fielding-Miller_covid_R3pV

Existing Data Replication

Study Information

1. Title (provided by SCORE)

RR TEAM INSTRUCTIONS: *This has been determined by SCORE.*

Replication of a research claim from Fielding-Miller et al. (2020).

2. Authors and affiliations

RR TEAM INSTRUCTIONS: *Fill in the names and affiliations of your team below.*

Kent Jason Cheng, Data finder[1]
Radoslaw Panczak Data analyst[2]

1 Department of Social Science, Maxwell School of Citizenship and Public Affairs, Syracuse University, Syracuse, New York, USA
2 Institute of Social and Preventive Medicine, University of Bern, Mittelstrasse 43, CH-3012 Bern, Switzerland

3. Description of study (provided by SCORE)

RR TEAM INSTRUCTIONS: *This description has been provided by SCORE. Please review and make a SCORE project coordinator aware of any edits, additions, and corrections you would suggest to the paragraph. You are free to add additional descriptions of your project in a separate paragraph.*

There is a negative association between insured status and mortality. This reflects the following statement from the paper's abstract: "Percentage of uninsured individuals was associated with lower reported COVID-19 mortality (b = -0.36, p = 0.001)." The claim is tested with a spatial autoregressive model to assess the association between number of deaths and percentage of uninsured individuals, adjusting for potential confounders, and fitted the model with a spatial lag of the dependent variable based on a contiguity matrix. The finding is that the percentage of uninsured individuals was associated with lower reported COVID19 mortality (b = -0.36, p = 0.001).

4. Hypotheses (provided by SCORE with possible Data Analyst additions)

RR TEAM INSTRUCTIONS: *The focal test for SCORE is indicated as H*. If you will test additional hypotheses (or use alternate analyses) that help you to evaluate the claim your replication/reproduction is testing, number them H1, H2, H3 etc. (You can place H* in the list wherever makes sense). Please make sure that any additional hypotheses are logical*

deductions/operationalizations of the selected SCORE claim or are necessary to properly interpret the focal H hypothesis. Research that is outside this scope should be described in a separate preregistration.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Are the listed hypotheses specific, concise, clearly testable, and specified at the level of operationalized variables?*
- *Are hypotheses identified as directional or non-directional, and, if applicable, have the direction of hypotheses been stated? (Example: "Customers' mean choice satisfaction will be higher in the CvSS architecture condition than in the standard attribute-by-attribute architecture condition.")*
- *Does the list of hypotheses/tests indicate whether additional hypotheses are taken from the original study or modified/added by the team?*

H*: At the county level, a higher percentage of uninsured individuals will be associated with lower reported COVID-19 mortality.

Design Plan

5. Study type

NOTE: *The study type selected should be based on the data collected for the replication, and not necessarily the data used in the original study.*

- Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.
- **Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.**
- Meta-Analysis - A systematic review of published studies.
- Other

6. Blinding

RR TEAM INSTRUCTIONS: *Select any/all of the below that apply for your study by bolding them. You will give a longer description in the next question.*

- **No blinding is involved in this study.**
- For studies that involve human subjects, they will not know the treatment group to which they have been assigned.
- Personnel who interact directly with the study subjects (either human or non-human subjects) will not be aware of the assigned treatments. (Commonly known as “double blind”)
- Personnel who analyze the data collected from the study are not aware of the treatment applied to any given group.

[QUESTION 6 - BOLD YOUR RESPONSE ABOVE]

7. Blinding

RR TEAM INSTRUCTIONS: *Since all existing data replications are based on data that has already been collected, in most cases it will not be necessary to comment on participant blinding. In the rare instance when an existing experiment is being re-analyzed for an existing data replication and blinding is a relevant consideration, please provide below any details regarding blinding that are important for a reviewer to be aware of.*

Blinding is not applicable in the data.

8. Study Design

RR TEAM INSTRUCTIONS: *Please describe how data was collected in the original study and how it compares to the data that was selected for the replication attempt. Explain why the data selected for the replication study is suitable for a replication and if any substantial deviations exist between the two.*

If the data used in the replication combines observations from the original study with new observations (e.g. if the data selected for the replication attempt comes from the same longitudinal survey as the original study), describe how 'original' and 'new' observations relate to each other and an estimate for what proportion of the final dataset's observations will be comprised of original vs. new observations.

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify the unit of analysis?*
- *Does the preregistration provide sufficient detail about how the data selected for the replication attempt deviates from or is congruent with the data employed in the original study?*
- *Does the preregistration describe whether and how 'original' and 'new observations' are combined together for the replication dataset?*

The study aims to assess the social determinants of COVID-19 deaths in the US on a county-level basis. The focal claim of interest is the negative association between insured status and COVID-19 mortality in non-urban counties. The authors used spatial regression models to test their hypotheses. As far as the focal claim of interest is concerned, original data includes non-urban counties that had at least 1 COVID-19 deaths as of April 26. The SCORE pre-identified replication dataset allows the analysis to be extended up to the present; in this replication dataset's case, up to July 16, 2020.

9. Randomization (free response)

RR TEAM INSTRUCTIONS: *If the variables used for this replication attempt were randomized, state how they were randomized, and at what level.*

The data is a compilation of all reported COVID-19 patients in the US. Therefore, no randomization is involved.

Sampling Plan

This section describes how the data sources for the replication were selected, how they were prepared into a replication dataset, and the number of observations that will be analyzed from these data. Please keep in mind that the data described in this section are the actual data used for analysis, so if you are using a subset of a larger dataset, please describe the subset that will actually be used in your study.

10. Existing data (multiple choice question, provided by SCORE)

1.1.1. Registration prior to creation of data
1.1.2. Registration prior to any human observation of the data
1.1.3. Registration prior to accessing the data
1.1.4. Registration prior to analysis of the data
1.1.5. Registration following analysis of the data

11. Explanation of existing data

NOTE: *For replications that rely on existing data sources, this question refers to the data that will be used for the replication analysis (i.e. the final replication dataset), and not (a) the data from the original study or (b) the data sources accessed to construct the replication dataset. Since no new data will be created for 'existing data replications,' 1.1.1 should never be selected. Since all analyses will occur after registration, 1.1.5 should also never be selected.*

The dataset referenced above has been accessed and cleaned prior to registration. Variables were selected based on their expected relevance to the replication analysis. None of the variables were selected because of their likelihood (or not) of leading to a confirmatory result.

12. Data collection procedures

RR TEAM INSTRUCTIONS: *Please describe the process for constructing the replication dataset in as much detail as you can. The sections below should be used to provide the following information:*

- *Which variables are needed from the original study to perform a good-faith, high-quality replication.*
- *Which data sources were used, why they were selected, any deviations between the original study design and the replication study design that these selections present, and the procedures used to access the data.*
- *Which of the variables from the original study are available in the replication data sources, including relevant details about each measure.*
- *The procedure for creating the replication dataset, in both narrative and script form.*
- *A data dictionary that documents each variable included in the replication dataset.*

In the sections below, please provide links to the original materials whenever possible -- including descriptions of the original datasets and corresponding codebooks. If materials can be shared on the OSF, please do so, and provide view-only links to those materials.

Specific points to keep in mind for reviewers:

- *Does the preregistration describe which data sources were selected for the replication study and why each is suitable?*
- *Does the preregistration make clear how the data sources were used to construct the replication dataset?*

(a) Data Needed

RR TEAM INSTRUCTIONS: *List below the datasets and variables the original author used to analyze the focal claim. Include details regarding the sample size, waves or years used, and other details pertinent to finding an existing dataset for replication. Please include page numbers when excerpting from the original article. If possible, categorize the list of variables as one of the following: dependent variable, focal independent variable, control variable, or sample parameters/clustering variable. Finally, include the sample size of the original study's focal analysis, if it is available.*

Dependent Variable(s)

COVID-19 mortality

- "We built a series of spatial autoregressive models to assess county-level associations between COVID-19 mortality…" (p 2)
- "We next built three separate spatial autoregressive models to assess the association between number of deaths and our hypothesized social determinants…" (p 3)
- Note: To the best of my understanding of the original analysis based from quotes above, the dependent variable is the number of deaths. This is unusual since most analysis of this kind I have seen used rate per 100,000 population as dependent variable instead. Or alternatively, count models like Poisson or negative binomial regression have been used if dependent variable is kept as a count.

Focal Independent Variable(s)

% residents uninsured

- "percentage of uninsured individuals under the age of 65" (p 2)
- "Percent uninsured was based on the US Census Small Area Health Insurance Estimates (SAHIE) program's 2018 estimates." (p 2)

Control Variable(s)

Percentage of Non-English speaking households

- "defined as households in which no one 14 years or older reports speaking English at least "very well" (p 2)
- "The proportion of households with limited English speaking ability was drawn from the American Community Survey's (ACS) 2014 5-year estimate."

Percentage of individuals engaged in hired farm work
- "percentage of individuals engaged in hired farm work in the county as of 2018" (p 2)
- "percentage of farmworkers was taken from the US Bureau of Economic Analysis" (p 2)

Percentage of individuals living at or below the poverty line
- "percentages of individuals living below poverty… were from 2017 ACS data" (p 2)

Percentage of residents over the age of 65
- "percentages of individuals… over the age of 65 were from 2017 ACS data" (p 2)

Population density
- "Density was measured as number of individuals per square mile, based on US census data."

Stage of the local epidemic
- "number of days since a county reported its first case of COVID-19, and the number of days between the 100th case in a state." (p 2-3)

Shelter-in-place control
- "Arkansas, Iowa, Nebraska, North Dakota, and Wyoming were assigned a '0', denoting that they had not yet implemented an SIP order at the time of these analyses." (p 3)
- Note: This may no longer be applicable to the replication dataset since all states have implemented some sort of stay at home order one way or another. See link.

Sample Parameters

Urbanicity
- "A non-urban model with population density less than or equal to 1000 individuals per square mile." (p 3)
- Note: the focal hypothesis to be tested is concerned with non-urban counties only.

Date
- "Spatial regression models, predictors of number of deaths across urban, non-urban, and all US counties reporting at least 1 COVID-19 case as of April 26, 2020" (p 6)

(b) Data Access

RR TEAM INSTRUCTIONS: *Describe below the data sources that will provide the replication variables. Include information such as the name of the data source (e.g., Indonesian Family Life*

Survey), the description and link of the data source, and the waves needed to create a final replication dataset.

Also describe the process for accessing the data sources that will be used to create the final replication dataset; specify how long long it took for the registration to be approved and what information was required (e.g., writeup of the purpose of the project, email address from an IPCSR institution, etc.); and verify that the data can be opened as expected. If applicable, provide a link to the page where you registered to access the data.

Describe in detail any restrictions on data access and data-sharing, as well as any additional terms of data use that will be relevant for the replication study and final report (e.g. citations that will need to be made). If you were able to access the data because of special permissions that you have, but that you expect other researchers might not have, please document those as well.

All of the data sources are accessible without the need for registration. The COVID-19 deaths data come from New York Times (right click "U.S. County-Level Data (Raw CSV)" to download .txt file). The proportion of households with limited English speaking ability, percentages of individuals living below poverty and over the age of 65, the percentage of uninsured, population, sourced through various waves of the American Community Survey (ACS) can be searched and downloaded through the US Census Bureau's website. Filter the search results to the desired wave of the ACS. In some instances, data is more complete for the 5-year estimates (e.g. for the uninsured and poverty variables).

The authors claimed that ACS was used to obtain population density but I could not find the data on the US Census Bureau's website. Thus, I decided to Social Explorer which my university subscribes to. To get the data, go to this link, click the dropdown entitled "American Community Surveys (5-year estimates)" which will show several options. Click "Begin Report" beside "American Community Surveys 2014--2018 (5-year estimates)." Then filter to "All counties," then click "Proceed to Tables." Select "Social Explorer Tables: ACS 2018 (5-year estimates)" from the dropdown, pick "A00002: Population Density (per Sq. Mile)," then click "show results." Click on the "Data download" tab and check the desired format. I downloaded the STATA-ready delimited .txt files. The set of files included .do, .dct, and .txt. One must update the links to the file locations of the .dct and .txt in the .do file before running the .do file which will provide the .dta file.

I could not find the percent of farmworkers from the pre-identified links by SCORE so instead, I searched for the variable through the US Department of Agriculture National Agricultural Statistics Service (NASS) website. To find the parameter in NASS, specify "Census" under Program box, "Economics" under Sector box, "Labor" under Commodity box, "Labor, Hired - Number of Workers" under Data Item box, "Total" under Domain box, "County" under geographic level box, highlight all states under State box, "2017" under Time box, and then click "get data." After the data preview appears, click "Spreadsheet" to get the .csv format of the data.

Note that the authors claimed they use 2018 percent of farmworkers but I could only find the 2017 version of the data through NASS. This should not be a major issue because the variable is unlikely to drastically change from 2017 to 2018. Moreover, the original analyses did settle for data from earlier years based on availability - e.g. the authors used 2014 ACS 5-year estimates for percentages of individuals living below poverty and 2017 ACS for percent of those over the age of 65.

There seems to be no recommended way to cite the New York Times data and the ACS data. Use this to cite NASS:

USDA National Agricultural Statistics Service, 2017 Census of Agriculture. Complete data available at www.nass.usda.gov/AgCensus.

Use this format to cite Social Explorer:

Social Explorer Tables: ACS 2018 (5-Year Estimates)(SE), ACS 2018 (5-Year Estimates), Social Explorer; U.S. Census Bureau

(c) Variable Availability

RR TEAM INSTRUCTIONS: *For each variable required for the replication analysis (listed above), describe the variables from the replication data that can be used to measure it (including which data files or sources each measure is found in),* ***any notes a data analyst should consider when using the measure in a replication analysis****, and any important differences between the original variable and the proposed replication variable.*

If there are multiple variables in the replication data that correspond to a required variable (e.g. two different measures of education in the replication data), include all of those options below. If a variable from the original study ***cannot*** *be measured using the replication data, please make that clear as well.* ***Finally, include a description of the identifiers used to merge multiple datasets, if applicable.***

Dependent Variable(s)

Variable name: deaths

- Cumulative counts of COVID-19 deaths per county as of the cutoff date
- File name: us-counties.txt
- Folder: Raw files
- Note: The original data's cutoff date is April 26 whereas the replication data's cutoff is July 16. In addition, use this to compute for the stage of the local epidemic - "number of days since a county reported its first case of COVID-19, and the number of days between the 100th case in a state case and the declaration of a state-wide shelter in place (SIP)." (p 2-3)

Focal Independent Variable(s)

The analyst can use these variables to construct the focal independent variable - the percent uninsured for those aged 65 and below.

Variable name: S2701_C05_011E

- Percent Uninsured under Age 19 from 2018 5-year ACS
- File name: ACSST5Y2018.S2701_data_with_overlays_2020-07-16T134058.csv
- Folder: Raw files/ ACSST5Y2018.S2701_2020-07-24T152258
- Note: The authors did not specify which ACS this came from. ACS 2018 only had 840 counties available whereas 2018 5-year estimates had more than 3000 counties' worth of data. Thus, ACS 2018 5-year estimates were chosen over the 2018 one. This must be added with S2701_C05_012E to get percent uninsured for those below age 65.

Variable name: S2701_C05_012E

- Percent Uninsured from age 19 to 64 from 2018 5-year ACS
- File name: ACSST5Y2018.S2701_data_with_overlays_2020-07-16T134058.csv
- Folder: Raw files/ ACSST5Y2018.S2701_2020-07-24T152258
- Note: The authors did not specify which ACS this came from. ACS 2018 only had 840 counties available whereas 2018 5-year estimates had more than 3000 counties' worth of data. Thus, ACS 2018 5-year estimates were chosen over the 2018 one. This must be added with S2701_C05_011E to get percent uninsured for those below age 65.

Control Variable(s)

Variable name: S1602_C01_001E

- Percentage of Non-English speaking households from 2014 ACS 5-year estimates
- File name: ACSST5Y2014.S1602_data_with_overlays_2020-07-16T234716.csv
- Folder: Raw files/ ACSST5Y2014.S1602_2020-07-24T132411.zip

Variable name: LABOR

- Individuals engaged in hired farm work from NASS
- File name: 4B63B4A6-D90C-3A4A-BB0B-F2B683CB4DB4.csv; Note that the file name was automatically generated during the downloading process. It may generate a different file name.
- Note: must be divided by county population to get the percentage. The original analysis used 2018 values but as mentioned, I could only find the 2017 values for this variable.

Variable name: S1701_C03_001E

- Percent below poverty line for population for whom poverty status is determined from 2018 ACS 5-year estimates
- File name: ACSST5Y2018.S1701_data_with_overlays_2020-07-19T095640.csv

- Folder: Raw files/ ACSST5Y2018.S1701_2020-07-24T153138.zip

Variable name: DP05_0001E

- Total population from 2017 ACS 5-year estimates
- File name: ACSDP5Y2017.DP05_data_with_overlays_2020-07-24T143015.csv
- Folder: Raw files/ ACSDP5Y2017.DP05_2020-07-24T143112.zip
- Note: The authors specified that this came from 2017 ACS this came from but it only had 840 counties available whereas 2017 5-year estimates had more than 3000 counties' worth of data. Thus, ACS 2017 5-year estimates were chosen over the 2017 one. Use this to obtain percent farm workers.

Variable name: DP05_0024PE

- Percent population aged 65 and over from 2017 ACS 5-year estimates
- File name: ACSDP5Y2017.DP05_data_with_overlays_2020-07-24T143015.csv
- Folder: Raw files/ ACSDP5Y2017.DP05_2020-07-24T143112.zip
- Note: The authors specified that this came from 2017 ACS this came from but it only had 840 counties available whereas 2017 5-year estimates had more than 3000 counties' worth of data. Thus, ACS 2017 5-year estimates were chosen over the 2017 one.

Sample parameters

Variable name: A00002_002

- Population density (per sq. mile)
- File name: R12591036_SL050.dta
- Folder: Raw files/ From Social Explorer
- Note: The authors used this to determine urbanicity. Specifically, authors defined non-urban counties as those with <=1,000 population per square mile.

Variable name: date

- This is the time variable for the data in MM/DD/YYYY format.
- File name: us-counties.txt
- Folder: Raw files
- Note: Use this to indicate the dependent variable which is the cumulative deaths as of July 16, 2020. Also, use this to compute for the stage of the local epidemic - "number of days since a county reported its first case of COVID-19, and the number of days between the 100th case in a state." (p 2-3)

The following are needed to merge the New York Times, NASS, ACS, and Social Explorer data:

Variable name: GEO_ID

- Description: According to the CENSUS Bureau, “GEOIDs are numeric codes that uniquely identify all administrative/legal and statistical geographic areas for which the Census Bureau tabulates data.” This is a combination of State and County FIPS codes.
- Notes: This is the identifier in all ACS-obtained data. Split this string variable to obtain FIPS code.

Variable name: fips

- Description: According to the CENSUS Bureau, “American National Standards Institute codes (ANSI codes) are standardized numeric or alphabetic codes issued by the American National Standards Institute (ANSI) to ensure uniform identification of geographic entities through all federal government agencies. The American National Standards Institute (ANSI) has taken over the management of geographic codes from the National Institute of Standards and Technology (NIST). Under NIST, the codes adhered to the Federal Information Processing Standards (FIPS). ANSI now issues two types of codes (available from the link above). They continue to issue the commonly used FIPS codes, although the acronym has now changed to Federal Information Processing Series, because it is no longer considered the standard. The tables included on this webpage (link above) provide the FIPS codes, which are the codes most commonly used by the Census Bureau.”
- Notes: This is the identifier in the New York Times data.

Variable name: countyansi, stateansi

- Description: ANSI is a different name for FIPS.
- Notes: These are the identifiers in the NASS data. Combine stateansi and countyansi to get FIPS specific to each county.

(d) Data Creation

RR TEAM INSTRUCTIONS: *Create a dataset using the data sources and variables listed above. Provide a detailed narrative describing how the various datasets were cleaned and merged into a final replication dataset. Provide a view-only link to a clearly commented script on the OSF that produces the replication data as described in the narrative. Our preference is that this be either an R script or a script from another language that similarly allows for open and reproducible analyses. Please let the SCORE team know if this is not possible.*

- *If the data can be freely shared and posted to OSF, please post it in your OSF project and provide a link to the completed dataset below.*
- *If any part of the dataset cannot be shared between researchers or posted to the OSF, please leave the final dataset off the OSF. Instead, include either below or in your script (commented out at the bottom) two pieces of information that will help an independent team verify they have created the dataset according to your instructions:*
    - *The dimensions of the final dataset(s) you’ve created (# of rows, # of columns)*
    - *A summary of 8-10 variables in the replication dataset. For numeric variables, the summary should include the mean, standard deviation, and count of NAs. For*

categorical variables, the summary should include each level present in the data and its count, as well as a count of NAs. If multiple datasets are submitted as part of your work, at least one variable should be included from each dataset.

The data from the replication sources should be preserved in as 'raw' a form as possible, in order to give the data analyst the most latitude to clean the variables as they see fit. Variables from the original source should be preserved in their original form (e.g. do not recode values of 99 to NA). New variables should only be created when they're needed to complete the merge or combine the datasets; in those cases, please preserve a version of the original, unaltered variable in the new dataset.

When combining multiple datasets by binding rows, please be sure that the data type and measurement units are equivalent across each dataset. If there is a discrepancy in how a variable is measured across datasets, rename the variable in each dataset to indicate the original dataset, and then carefully document the resulting measures below and in the data dictionary. See here for an example of how this should work.

Please also use this section to describe:

- *Any deviations between the original study design and the replication design that would result from using this replication dataset.*
- *Any notes about using these variables that you would like to pass along to the data analyst.*

The analyst must be aware that county identifiers are different across datasets. I had to generate the fips code or the county identifiers from each dataset's available identifier first before merging the dataset. The script I prepared called [1] Merging datasets_v2.do format and the all processed datasets in .dta can be found here. Note that the $ in the script must be replaced with the file destination. The main coding strategy I employed involved merging everything with the New York Times so that the data would follow the longitudinal format which would be closer to being ready for analyses than if I left the data in wide format. Specifically, the long format is needed to help the analyst derive the following covariates: number of days since a county reported its first case of COVID-19, and the number of days between the 100th case in a state.

The merged dataset is called merged_covid_usa_v2.dta in the same folder. Note that to give the analyst the freedom to reshape the data as s/he sees fit, I maintained all the available data from the New York Times database. The analyst must remove urban counties prior to analyses since the focal hypothesis of interest only concerns non-urban counties. Moreover, the dependent variable is cumulative deaths as of the cutoff date (original data's cutoff date is April 26 while the replication's is July 16). The analyst must be aware that the analyses must be filtered to these dates so long as the data is kept in long format.

(e) Data Dictionary

RR TEAM INSTRUCTIONS: *Create a data dictionary following this template. Provide below a view-only link to the completed data dictionary included in the OSF project. If the Data Analyst will need to create new variables using the variables in the final replication dataset (e.g. recoding the provided education variable to be in a better format for analysis), please document below your recommendation on how the analyst should do so. Please also document any additional notes regarding the variables in the dataset that do not fit within the provided data dictionary template or the other sections above.*

Data dictionary called data_dictionary_covid_usa_v2.xlsx can be found here.

13. Sample size

RR TEAM INSTRUCTIONS: *Please report below the analytic sample size(s) in the replication dataset, with reference to however many units or levels are in the data. Please report as much information here as will be helpful for the review committee to be aware of, including differences in sample size resulting from various analytic decisions (e.g. listwise deletion vs multiple imputation).* ***Finally, when the replication combines observations from the original study with new observations, please estimate what proportion of the analytic sample's observations will be comprised of original vs. new observations.***

The original study included 2,629 non-urban counties. After conducting listwise deletion, the replication dataset is estimated to have 2,862 non-urban counties. This increases the original data's number of observations by 6%.

Required sample size [to be filled out by the SCORE team]: The primary unit of analysis is non-urban US counties. An estimate of the minimum viable sample size for the data analytic replication is: 933. For comparison, the stage1 required sample size would be: 4,536 and the stage2 sample size would be: 10,203.

14. Sample size rationale

For data analytic replications in SCORE, three sample sizes are calculated:

- *A minimum threshold sample size, defined as the sample size required for 50% power of 100% of the original effect*
- *A stage 1 sample size, defined as the sample size needed to have 90% power to detect 75% of the original effect*

- *A stage 2 sample size, defined as the sample size needed to have 90% power to detect 50% of the original effect*

Details of the power analysis for the replication can be found here: https://osf.io/kx6ar/?view_only=33de775a25654219969defa5f60fd6ea

15. Stopping rule (provided by SCORE)

For this replication, SCORE recommends three analyses be performed:

- **one analysis that only uses the dates that have occurred since the original analysis**
- **one analysis that combines all available dates**
- **a third analysis that only uses dates that were used in the original analysis**

Variables

RR TEAM INSTRUCTIONS: *The preregistration form divides variables across three questions: manipulated variables, measured variables, and indices (i.e. analytic variables derived from raw variables). For existing data replications, only fill out the "Measured variables' and 'Indices' sections. Please do not fill out anything in the 'Manipulated variables' section.*

The raw data of any transformed variable (e.g. reaction time → log reaction time) or any created index should be defined in the 'Measured variables' section. Details regarding the variable transformation should be specified in the 'Transformations' section. Details regarding the creation of an index should be specified in the 'Indices' section.

Across these questions, you should define all variables that will later be used during your analysis (including data preparation/processing). You can describe all variables in the preregistration and/or summarize and link to a data dictionary (codebook) in your repository to answer these questions.

If you will share data from your replication, this is also the place to state whether any variables will be removed prior to sharing the dataset (e.g. to reduce risk of participant identification or comply with copyright restrictions on scale items.)

16. Manipulated variables

RR TEAM INSTRUCTIONS: *Manipulated variables in this preregistration refer specifically to variables that have been randomly assigned in an experiment. The use of data from an experiment should be rare in existing data replications. If your existing data replication relies on experimental data, please document each manipulated variable as a measured variable, and*

use the codebook to indicate what each level of the variable corresponds to (e.g. participants assigned to the treatment condition = 1; participants assigned to the control condition = 0). The default language in bold below has been copied into all existing data replication preregistrations.

N/A -- not documented for existing data replications.

17. Measured variables

RR TEAM INSTRUCTIONS: *Please use this section to document each variable that was used in the original study's analysis and the role it served (e.g. dependent variable, control variable, sample parameter, etc). For each variable, provide the description of the variable offered in the paper and/or codebook of the original study, the variable in the replication dataset that it corresponds to, and explain any deviations between the two. In cases where an equivalent replication variable was not found, explain how, if at all, you expect it will affect the replication attempt. In cases where you are adding a variable that was not present in the original study, please explicitly state that you are doing so, and explain how, if at all, you expect it will affect the replication attempt.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration surface all of the variables needed to replicate the focal analysis?*
- *Are deviations between the original variables and replication variables documented when needed?*

VARIABLE NAME

- [Use in the analysis]
- [Description from the original study]
- [Variables used in the replication (if it needs to be constructed from multiple measures, include all of them here)]
- [Deviations between the original study and the replication study]

deaths

- [Use in the analysis]: outcome
- [Description from the original study]: *COVID-19 mortality on the county level at the date of the preprint / replication*

date_proper

- [Use in the analysis]: selection of time frame for analyses; calculation of time_case1 & time_case100 variables
- [Description from the original study]: *NA*

fips

- [Use in the analysis]: unique identifier of geographical area used to merge tabular data with spatial data; selection of states included in the analysis; generation of sample
- [Description from the original study]: *NA*

uninsured (S2701_C05_011E, S2701_C05_012E)

- [Use in the analysis]: focal exposure in regression
- [Description from the original study]: *percentage of uninsured individuals under the age of 65*
- [Variables used in the replication (if it needs to be constructed from multiple measures, include all of them here)]: constructed from sum of S2701_C05_011E and S2701_C05_012E variables

nonenglish (S1602_C01_001E)

- [Use in the analysis]: non-focal exposure in regression
- [Description from the original study]: p*ercentage of Non-English speaking households (defined as households in which no one 14 years or older reports speaking English at least "very well")*

farmwork (LABOR, DP05_0001E)

- [Use in the analysis]: non-focal exposure in regression
- [Description from the original study]: *percentage of individuals engaged in hired farm work in the county as of 2018*
- [Variables used in the replication (if it needs to be constructed from multiple measures, include all of them here)] derived from LABOR variable divided by DP05_0001E

poverty (S1701_C03_001E)

- [Use in the analysis]: non-focal exposure in regression
- [Description from the original study]: *percentage of individuals living at or below the poverty line*

older (DP05_0024PE)

- [Use in the analysis]: non-focal exposure in regression
- [Description from the original study]: *percentage of residents age 65 or older*

pop_dens (A00002_002)

- [Use in the analysis]: non-focal exposure in regression & selection of nonurban counties
- [Description from the original study]: *county density, measured as number of residents per square mile*

nonurban

- [Use in the analysis]: stratification of the analyses

- [Description from the original study]: urban - counties with population greater than 1000 individuals per square mile, non-urban - population density less than or equal to 1000 individuals per square mile

time_case1

- [Use in the analysis]: non-focal exposure in regression
- [Description from the original study]: *number of days since a county reported its first case of COVID-19*
- [Variables used in the replication (if it needs to be constructed from multiple measures, include all of them here)]: derived using date of first case in the county

time_case100

- [Use in the analysis]: non-focal exposure in regression
- [Description from the original study]: *the number of days between the 100th case in a state and the declaration of a state-wide shelter in place (SIP) order*
- [Variables used in the replication (if it needs to be constructed from multiple measures, include all of them here)] derived from the date where cumulative number of cases went over 100 and a date of SIP order in state (sip_effect); five states had it hard coded to zero as per preprint specifications

18. Indices

RR TEAM INSTRUCTIONS: *If any of the measured variables described in Section 17 will be combined into a composite measure (including simply a mean), describe in detail what measures you will use and how they will be combined. Please be sure this preregistration includes a link to a clearly commented script that constructs the index according to the narrative.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify each of the composite measures (e.g. mean scores, factor scores) that are needed for the focal analysis, and which of the measured variables in Section 17 are used in each one (e.g. the happiness, joy, and satisfaction items will be used to create the 'positive feelings' measure)?*
- *Does the preregistration link to a clearly commented script that constructs the indices according to the narrative description?*

NA

Analysis Plan

19. Statistical models

RR TEAM INSTRUCTIONS: *This section should describe in detail the analysis that will be performed to replicate the focal result. This analysis must align as closely as possible with the*

original study's analysis, even if you have identified limitations in the original study. The level of detail should allow anyone to reproduce your analyses from your description below. Examples of what should be specified: the model; each variable; adjustments made to the standard errors and to case weighting; additional analyses that are required to set up the focal analysis; and the software used.

Beyond the replication of the focal analysis from the original study, it is at your discretion to test the claim using other analytic approaches as a check of the robustness of the claim. The original test should be listed first and be clearly distinguished from any other tests. If you are testing additional confirmatory hypotheses, describe them in the same order as you numbered them in the "Hypotheses" section above and make clear reference to the specific hypothesis being tested for each.

Please provide a link to a clearly commented script that performs the analysis described in the narrative provided below. Our preference is that this be either an R script or a script from another language that similarly allows for open and reproducible analyses. Please let the SCORE team know if this is not possible. Please also test that the code runs without error on a random subset of 5% of the replication dataset, and provide verification that the code has produced a sensible result below (a screenshot of the results is preferable). Finally, please confirm that you have only developed and tested your analysis plan and code using 5% of the data.

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify which statistical model will be used to provide the 'focal evidence' for the SCORE test (e.g. a regression coefficient in a larger multiple regression model), and does it correspond closely to the model and evidence from the original study?*
- *Does the preregistration describe each variable that will be included in the focal analysis, and what role each variable has (e.g. dependent variable, independent variable)?*
- *Does the preregistration include a detailed specification of the focal analysis, including interactions, lagged terms, controls, etc., in both narrative form and in a clearly commented script?*
- *Does the preregistration verify that the code runs without error on a random subset of the replication dataset?*

This statement confirms that only 5% of the data have been randomly sampled in developing the analysis plan and code contained in this preregistration. Extra care has been placed to select spatially contiguous sample of observations in order to obtain the correct neighbourhood matrix used in the models needed to replicate the study.

In order to replicate the analysis we need to fit a spatially autoregressive model with the cumulative counts of deaths in a county at a given date as an outcome. The model has to be fitted to a subsample of non-urban counties selected on the basis of population density. In order

to fit spatial models, the geographical dataset of county boundaries has to be used in order to construct spatial contiguity matrix, which is then used in the model fitting. Spatial contiguity matrix can be specified using “queen” or “rook” contiguity. Preprint does not specify that. Since first option is the far more prevalent in case of analyses of administrative boundaries this method will be chosen for the analyses and sensitivity analysis using the second option will be conducted to assess the impact on the results.

Model is adjusted for the covariates listed in section **17**.

Preprint does not specify software that was used nor any further details regarding modelling beyond that.

If Stata is used, which is reasonable to guess judging from the structure of the analyses and wording used by the authors, then analyst is required to choose between one of the two methods of estimation when building spatial models with `spregress` command, with no default option provided.

- *ml* - use maximum likelihood estimator
- *gs2sls* - use generalized spatial two-stage least-squares estimator.

The Stata’s manual informs that the result should be consistent across two methods when data are homoskedastic, and that deviation might occur when *ml* method when working with heteroskedastic data. Given the constraints of working with ~5% of data where we cannot determine the nature of the relationship between the variables we will resort to use *gs2sls* method to guard the analysis of full dataset against the potential presence of heteroskedasticity

~~Without any further information the most reasonable strategy is to use both and compare results.~~

Publicly available code that aims to prepare the data and perform analyses using two time frames of the data is available in github repository of the data analyst. In a series of commented literate programing markdown documents a series of steps is proposed to replicate the analyses in the preprint. More specifically,

- file `01_spatial-sample.Rmd` is R code used to create spatially contiguous sample of 5% counties and discussing details of geography used
- file `02_data-preparation-extended.stmd` is a Stata code to prepare the data to required format, deal with missing and select sample data.
- file `03_analysys-5perc-sample-extended.stmd` is a Stata code required to run analyses on sample of 5% data using the most recent data available
- files 04 & 05 replicate two steps above but constraining dataset to the dates specified in the original study

All files above were run without issues and output of analyses is saved in a series of html files with names matching the scripts that can be found in the repository. This output was created from stmd files using `markstat` command. If an analyst prefers to use standard Stata do files those are preserved in the repository as well.

20. Transformations

RR TEAM INSTRUCTIONS: *This section should describe how any of the measured variables or composite measures mentioned above will be transformed prior to the analyses listed in Section 19. These are adjustments made to variables **after** measurement or measure creation, and might include centering, logging, lagging, rescaling etc. Please provide enough detail such that anyone else could reproduce the transformations based on the description below. Please be sure this preregistration includes a link to a clearly commented script that performs the transformations described in the narrative provided below.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify which of the measured variables or composite measures will need to be transformed prior to the focal analysis?*
- *For each variable needing transformation, does the preregistration adequately describe the transformations, including any centering, logging, lagging, recoding, or implementation of a coding scheme for categorical variables?*
- *Does the preregistration link to a clearly commented script that performs each transformation?*

NA

21. Inference criteria

RR TEAM INSTRUCTIONS: *This section describes the precise criteria that will be used to assess whether the hypotheses listed above were confirmed by the analyses in Section 19. The default language below only applies to the test of the SCORE claim, **H***. It is at your discretion to describe the inferential criteria you will use for any additional analyses. They need not rely on p-values and/or the same alpha level we have specified for **H***.*

If the additional analyses will use multiple comparisons, the inference criteria is a question with few "wrong" answers. In other words, transparency is more important than any specific method of controlling the false discovery rate or false error rate. One may state an intention to report all tests conducted or one may conduct a specific correction procedure; either strategy is acceptable.

Criteria for a successful replication attempt for the SCORE project is a statistically significant effect (alpha = .05, two tailed) in the same pattern as the original study on the focal hypothesis test (**H***). For this study, this criteria is met by obtaining a statistically significant (p- value

specified in the preprint is 0.001~~ - it is assumed it means p <= 0.001~~) regression coefficient from the adjusted model run on the subsample of non-urban counties.

22. Data exclusion

RR TEAM INSTRUCTIONS: *The section below should describe the rules you will follow to exclude collected cases from the analyses described in Section 19. Note that this refers to exclusions **after** the creation of the replication dataset; exclusion criteria that prevent a case from entering the replication dataset in the first place should be detailed in the 'Data Collection Procedure' section above. Please be as detailed as possible in describing the rules you will follow (e.g. What is the specific definition of outliers you will use? Exactly how many attention checks does a participant need to fail before their removal from the analytic sample?).*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration comment on whether any cases included in the replication dataset will be excluded prior to data analysis?*
- *If yes, does the preregistration provided detailed instructions on how the exclusions will be performed (e.g. Is the definition of outlier provided? Is the number of attention checks failed before a participant is excluded specified?)*

County is the basic unit of analyses. Some of the units were excluded from the analyses. More specifically:

1. We will exclude counties that were labelled as "Unknown".
2. Data analyst identified information on explanatory variables for counties not present in the NYT dataset (that includes NY counties - see info below). Since these counties do not have information about the outcome they will be excluded from the analysis.
3. We will exclude counties that are missing fips codes: Joplin and Kansas City. Explanatory variables were not provided for those counties. Most likely it was due to the missing codes. It was possible that data were excluded from the original analyses. It was also possible that counties were merged to other counties, however without any specifications available from the preprint the most convservative approach would be to exclude them
4. 5 counties of New York City are aggregated in a dataset of cases. Data finder provided characteristics of individual counties however it is not possible to determine how the data was handled exactly. It was possible that all cases were assigned five time to each county, one county was kept with aggregated explanatory variables or some other technique was used. Since focal analysis focuses on non-urban counties it might be reasonable to leave this issue unresolved and exclude these counties.
5. The analysis focuses on "50 states" - we will assume this excludes the counties from Northern Mariana Islands" and Puerto Rico and remove them from the analyses.

23. Missing data

RR TEAM INSTRUCTIONS: *The section below should describe how missing or incomplete data will be handled. Please be as detailed as possible in describing the exact procedures you will follow (e.g. last value carried forward; mean imputation) and any software required (e.g. We will use Amelia II in R to perform the imputation).*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration comment on how missing or incomplete data will be addressed (e.g. casewise removal, missing data imputation)?*
- *If applicable, does the preregistration specify how many missing variables will lead to a case's removal (e.g. If a subject does not complete any of the three indices of tastiness, that subject will not be included in the analysis.)?*
- *If applicable, does the preregistration describe how missing data imputation will be performed, including relevant software?*

1. Analyses are adjusted by a variable specifying the number of days between the 100th case and the declaration of the state-wide shelter in place (SIP) order. There is a problem with generating this variable for counties that did not reach 100 cases by the time analyses were conducted. Preprint is silent on what was done in such cases. The only two reasonable strategies would be to run analyses without counties with missing information (which would result in drastic sample size reduction) or impute it to 0. The latter has been applied in this case.
2. There are some counties in the dataset missing information on few of the covariates. Particularly affected was the LABOUR variable. Again the preprint is silent on what was done in such a case. Data finder failed to provide information on the origin of the missing (counties not present in original data vs counties present in original data but with missing information). Datasets as provided so not allow any reasonable imputation method to be performed. It is not clear if missing could be replaced by zero in terms of percentages. Default Stata behaviour would be to drop such cases during the analysis and run models on complete cases. In case of spatial models additional difficulty arises since after such deletion the tabular data is shrank it does not match the spatial matrix any longer. All counties with information missing on any covariates used in the regression will be then dropped and documented in the last stage.

24. Exploratory analysis (Optional)

RR TEAM INSTRUCTIONS: *If you plan to explore your data set to look for unexpected differences or relationships, you may describe those tests here. An exploratory test is any test where a prediction is not made up front, or there are multiple possible tests that you are going to use. A statistically significant finding in an exploratory test is a great way to form a new confirmatory hypothesis, which could be registered at a later time. If any exploratory analyses*

involve additions to the data collection procedure beyond what was performed in the original study (e.g. additional items on the survey; running another condition in the experiment), please describe them below.

NA

25. Other

RR TEAM INSTRUCTIONS: *This section serves two purposes. First, please use this section to discuss any features of your replication plan that are not discussed elsewhere. Literature cited, disclosures of any related work such as replications or work that uses the same data, plans to make your data and materials public, or other context that will be helpful for future readers would be appropriate here. Second, please also re-surface any major deviations from earlier in the preregistration that you expect a reasonable reviewer could flag for concern. Give a summary of these deviations, focusing on larger changes and any possible challenges for comparing the results of the original and replication study.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration reference other sections of the preregistration where substantial deviations from the original study have been described (including deviations due to differences in location or time compared to the original study)?*
- *Does the preregistration comment on plans to make the data and materials from the replication study public?*

Code of the analysis is public.

Final review checklist

REVIEWER INSTRUCTIONS: *For the following questions, reviewers please indicate whether you can 'sign off' on the following items by adding a comment. You can update this response as the lab moves through revisions during the review period!*

- Included in this pre-registration are specific materials needed to create a replication dataset:
    - Is the final replication dataset that the research team constructed suitable for performing a high-quality, good-faith replication of the focal claim selected from the original study?
    - Is the procedure for constructing the final replication dataset sufficiently documented that an independent researcher could construct the same dataset following the procedures and code they lay out?
- Included with this pre-registration is a narrative description of how the replication dataset will be used to perform the focal replication analysis, as well as the specific analytic scripts/code/syntax that will be used:
    - Is the analysis plan (including code) that's documented in the preregistration consistent with a high-quality, good-faith replication of the focal claim selected from the original study?
    - Has the data analyst demonstrated that the analysis code works as expected on a random 5% of the final replication dataset?
- I have reviewed all sections of this pre-registration, and I believe it represents a good-faith replication attempt of the original focal claim.