[RP] Report

HE&G ReScience

Reproduction of

Analysis of COVID-19 Cases' Spatial Dependence in US Counties Reveals Health Inequalities

by Saffary, T., Adegboye, O.A. Gayawan, E., Elfaki, F., Kuddus, M.A., Saffary, R.
in: Frontiers in Public Health, 8, pp. 1-10

Created
18 February 2021

Revised
13 June 2021

Reproduction Authors:

Peter Kedron[1,2] / Joseph Holler[3] / Sarah Bardin[†,1,2] / Summer Cliff / Kimberly Fuller / Joshua Gilman / Bryant Grady / Megan Seeley / Addison Van Zanbergen / Wenxin Wang / Xin Wang

[1]Arizona State University, School of Geographical Sciences and Urban Planning, Tempe, AZ, USA
[2]Arizona State University, Spatial Analysis Research Center (SPARC), Tempe, AZ, USA
[3]Middlebury College, Department of Geography, Middlebury, VT, USA
[†]Graduate Student Author

Reproduction Materials Available at:

Report – https://github.com/HEGSRR/RP-Saffary-2020/tree/main/docs/report
Data – https://github.com/oyeadegboye/USA_COVID- 19
Code – https://github.com/HEGSRR/RP-Saffary-2020/tree/main/procedure/code

Correspondence should be addressed to Peter.Kedron@asu.edu. The authors declare that no competing interests exist

Research Hypotheses to Reproduce

H1: (a) COVID-19 cases and deaths were non-randomly distributed globally across the counties of the contiguous United States, and (b) non-randomly distributed around individual counties within the contiguous United States.

Original test: The authors found support for the non-random distribution of cases (Moran's $I = 0.228$, $p < 0.0001$) and deaths (Moran's $I = 0.477$, $p < 0.0001$), and identified clusters around several large cities (LISA, Fig 2c).

H2: Non-zero global and local correlations exist between COVID-19 cases and deaths and county-level health factors (a) count of ICU beds, (b) count of primary care physicians, (c) adult obesity, (d) diabetes prevalence, (e) flu vaccination coverage, and (f) the proportion of the population without health insurance.

Original Test: The authors found support for the existence of non-zero spatial correlations between (a) cases-ICU (bivariate Moran's I=0.08, p<0.0001; Fig 2b), deaths-ICU (bivariate Moran's I=0.15, p<0.0001; Fig 2c). All other global and local bivariate Moran's I were reported as non-significant at the p=0.05 level. Non-random local patterns were reported (Fig 2, Fig3)

H3: Non-zero global and local correlations exist between COVID-19 cases and deaths and county-level demographics (a) percent black, (b) percent Hispanic, (c) percent white.

Original Test: The authors found support for the existence of non-zero spatial correlations between (a) cases-black (bivariate Moran's $I = 0.174$, p=0.0001, Fig 4f), deaths-black (bivariate Moran's $I = 0.264$, p=0.0001, Fig 4f). No significant global correlations were found for the Hispanic or white populations. Non-random local patterns were reported (Fig 4).

Supplemental Tests: The authors retested the above hypotheses making Bonferroni and False Discovery Rate adjustments for multiple testing. Under the Bonferroni adjustment the majority of statistically significant clustering and association disappeared. Under the False Discovery Rate, some clusters were removed.

Key Findings

- We were able to partially reproduce the original analyses. Reproductions were hindered by omission of a key variable (PCP) from the publicly available data and a lack of code availability.
- The inferences made in the paper may not be credible because the authors i) fail to account for multiple testing in their main results, ii) fail to adjust for population densities and age-structure, iii) misinterpret the Bivariate Moran's I
- The supplemental materials that accompany the original paper should be reported as the main results presented in the paper. When adjustments for multiple testing are made, the results of the analysis are radically different. The adjustments that were completed may have been misapplied.

Original Study Information

Description:

Saffary et al. (2020) examined whether socio-demographics and healthcare resources are correlated in space with COVID-19 cases and deaths across 3,108 counties located within the contiguous United States. The original analyses are retrospective and use observational data collected from federal and private sources. The data used in the original analysis is available. The code used for the original analysis was not made available by the authors.

Analytical Plan:

Sampling Plan and Data Description: Saffary et al. collected data from online, publicly available datasets. The datasets included socioeconomic, healthcare, and health data for counties within the contiguous United States (3,108). COVID-19 cases and death data up to May 22, 2020 were collected from USA facts and presented as per 100,000 population. The number of intensive care unit (ICU) beds per county was collected from Kaiser Health News. All other data were gathered from the World Health Organization (WHO), U.S. Department of Agriculture, U.S. Census Bureau, governmental, non-governmental, and educational institutions. No further manipulation of the data was described.

Variables: The authors examined spatial patterns and relationships among:

1. COVID-19 cases and deaths per 100,000 population prior to May 22, 2020
2. The count of ICU beds in each county
3. The number of primary care physicians (PCP) per 10,000 people
4. Adult obesity, defined as the percentage of adults with Body Mass Index (BMI) greater than 30
5. The percentage of adults over 20 with diabetes
6. The percentage of people under 65 without insurance
7. The percentage of annual Medicare enrollees with an annual flu vaccination
8. The percentage of the population that is non-Hispanic Black, Hispanic, and Non-Hispanic white

Analytical Specification: All original analyses were performed in R statistical software (version 3.6.2). No coordinate system or projections were specified, but the data was aggregated at the county level. Edge effects were not addressed. Global and local univariate and bivariate Moran's I with a first order Queen's contiguity spatial weight matrix defined the spatial relationships between counties and was used to evaluate all hypotheses.

Inference Criteria, Results, and Robustness: The original authors conducted all hypothesis tests using the p=0.05 significance threshold. Local spatial statistics were also visually evaluated using national scale maps. Included only in the supplemental material, the authors applied Bonferroni and False Discovery Rate adjustments to account for multiple testing during local spatial statistical testing. However, these results were not discussed in detail in the original analyses, so the inference criteria used in their evaluation is unclear.

Reproduction Procedure

Protocol:

We followed the data preparation and analytical procedures of the original study, making as few modifications as possible. The data used in the original analyses was made available by Saffary et al. and we use this original data in our reproduction. We retrieved the data used in the reproduction from Github (https://github.com/oyeadegboye/USA_COVID- 19) on March 21, 2021. The code used for the original

analysis was not made available by the authors. Details of the authors' analytical workflow were gathered from their published article and accompanying supplemental material.

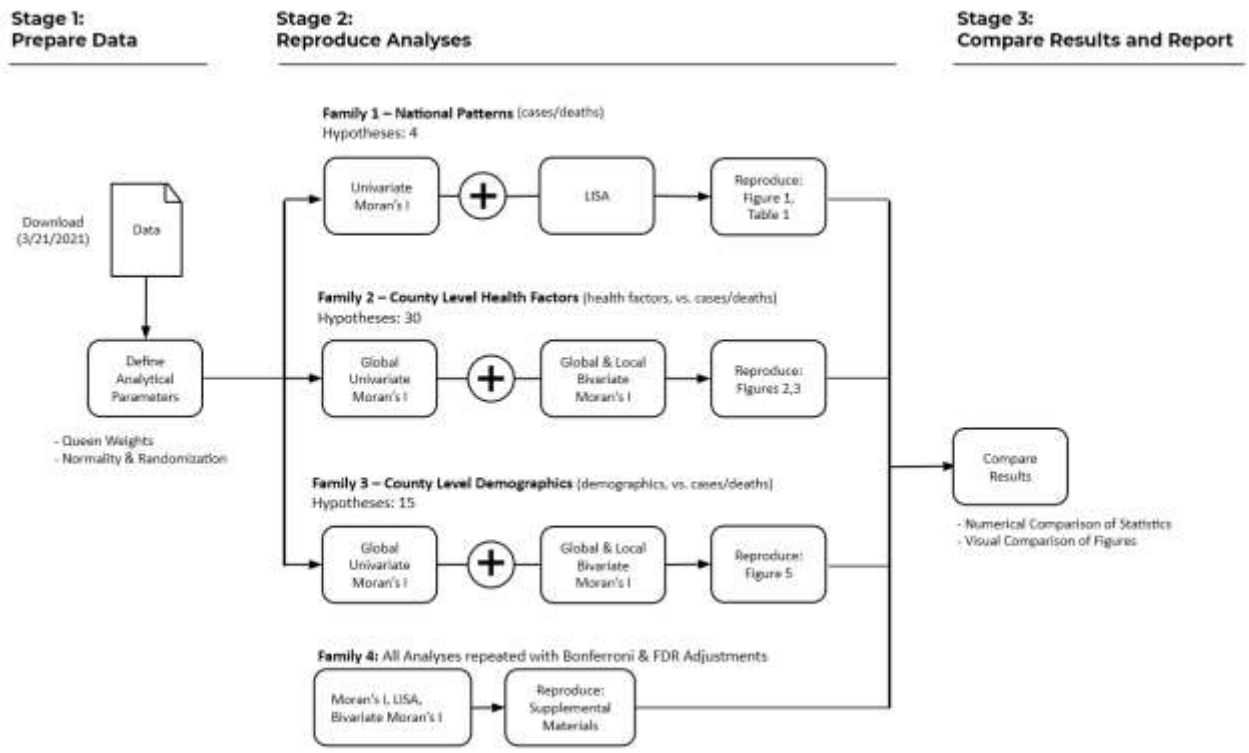


Fig 1. Reproduction workflow

Following the original authors, we tested a total of 49 hypotheses[1], which we organized into three families (Fig. 1). First, we used the univariate global and local Moran's I statistics to test the hypothesis that COVID-19 cases and deaths were non-randomly distributed globally across the counties of the contiguous United States. Second, we evaluated a family of hypotheses assessing whether non-zero global and local correlations exist between COVID-19 cases and deaths and county-level health factors. Third, we evaluated a family of hypotheses examining the possible existence of non-zero global and local correlations between COVID-19 cases and deaths and county-level demographics.

To correct for multiple testing, we included the Bonferroni bounds and false discovery rate (FDR) tests, which the original authors included in the supplemental materials. Given the large number of hypotheses being tested by the original authors in the univariate and bivariate LISA analyses, it is important to adjust for multiple hypothesis testing. For example, applying the standard $p = 0.05$ significance threshold even if no local clustering existed in all of the 3,108 U.S. counties, we would expect to observe approximately 156 false positive signals of local clustering (3,108 * 0.05 = 156) if we fail to adjust for multiple testing. Therefore, while the original authors viewed these findings as supplemental, we believe it is not only worthwhile to reproduce them, but that they should in fact be the focus of inference.

Matching the original authors, all of our tests are conducted within the R statistical environment using R version 4.0.1 and Rstudio version 1.3.959. The original authors did not specify the R package they used

[1] The local univariate and bivariate statistical tests in fact assessed 3,108 county-level hypotheses for each analysis. For simplicity we bundle these hypotheses together for each variable reproduction and discussion.

in their analysis or the statistical framework they used to make their inferences. We used the *localmoran* function from the **spdep** package to evaluate hypotheses associated with national patterns (Fig. 1; Family 1). Next, we used code developed by Pereira (2020) to calculate Bivariate Moran's I and evaluate hypotheses associated with county health factors and demographics (Fig. 1; Families 2, 3).[1] We evaluated all hypotheses under the randomization assumption. We created our figure reproductions using the **ggplot2** package.

Planned Differences from the Original Study: Our reproduction procedure matches that of the original authors, except for a small number of deviations. We depart from the original study in our reproduction through the use of R version 4.0.1 rather than version 3.6.2.

Lacking information from the original authors about the evaluation criteria used during statistical testing, we evaluated statistics under the randomization assumption. In the original study, both the Bonferroni bounds and FDR were used to correct for multiple testing, but the test results were only included in the supplementary material and the parameters used were not discussed in detail. For both tests, we investigated multiple methods and determined which best matches the results of the original study. For the Bonferroni, we both manually corrected the results by dividing the alpha value (α=0.05) by the number of comparisons (3,105) and used the *p.adjust* function from the R **stats** package (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/p.adjust). While performing the FDR correction, we again used the *p.adjust* function as well as the *fdrtools* function from the **fdrtools** package (https://cran.r-project.org/web/packages/fdrtool/fdrtool.pdf).

Assessment Criteria: We assessed the success of the reproduction based on whether we achieved bitwise reproduction - observing/estimating the exact same statistical values. For some of the tests used in the original article (e.g., those that assess significance through simulation) bitwise reproduction may not be possible, as the authors did not include information on the seed values of their randomizer. As a result, in cases where bitwise reproduction was not achieved, we compared the direction, magnitude, and significance of our results with those of the original authors.

Reproduction Results

Reproduction Results for the Descriptive Statistical Analysis (H1):

We combined the original data provided by Saffary et al. with county-level boundary information sourced from the U.S. Census Bureau 2018 Population Estimates API and County Health Rankings from 2020.[2,3] This file was subsequently reduced from 3,142 counties to 3,108 counties by restricting to counties within the contiguous United States. This 3,108 county dataset was the working dataset for all hypothesis testing.

To ensure this dataset matched the original dataset, we reproduced the descriptive statistics originally presented in Table 2. The means and medians were identical for nearly all variables except for those calculated for the count of primary care physicians (PCP), which were only slightly lower, and the cases and deaths. We were able to link the discrepancy in the mean and median cases and deaths to the original authors misreporting these values. The authors reported these values as rates per 100,000 population when they in fact reported raw counts.

Table 1. Descriptive summary of dependent and independent variables in original analysis and reproduction analysis

	Original		Reproduction	
	Mean	**Median (IQR)**	**Mean**	**Median (IQR)**
Cases per 100,000 population	503.1	32.0 (7.0-147.8)	268	110.1 (45.1-277.3)
Number of Cases[a]	n.a.	n.a.	503.1	32.0 (7.0-147.8)
Deaths per 100,000 population	30.5	1.00 (0-5)	11.5	2.0 (0-10)
Number of Deaths[a]	n.a.	n.a.	30.5	1.00 (0-5)
Health Factors				
Number of ICU beds	23.6	0 (0-12)	23.6	0 (0-12)
PCPs per 10,000 population	54.5	48.0 (32.0-71.0)	n.a.	n.a.
PCPs per 10,000 population (filtered)[b]	n.a.	n.a.	54.15	48.13 (31.88-70.23)
PCPs per 10,000 population (zero imputed)[b]	n.a.	n.a.	51.96	46.49 (28.77-69.46)
PCPs per 10,000 population (mean imputed)[b]	n.a.	n.a.	54.51	50.06 (32.58-69.46)
Adult Obesity (BMI > 30)	32.9	33 (29-37)	32.9	33 (29-37)
Diabetes	12.2	12 (9-15)	12.2	12 (9-15)
Uninsured	11.5	11 (7-14)	11.5	11 (7-14)
Flu vaccinations	41.7	43 (36-49)	41.5	43 (36-49)
Race Factors				
Black American	9	2.2 (0.7-10.2)	9	2.2 (0.7-10.2)
Hispanic	9.6	4.4 (2.4-10.0)	9.7	4.4 (2.4-10.0)
Non-Hispanic White	76	83.4 (64.3-92.3)	76	83.4 (64.3-92.3)

Sources: Saffary et al., 2020; County Health Rankings, 2020
Notes: Estimates are percentages out of 100, unless otherwise noted. All estimates were calculated using the full sample of 3,142 counties in the U.S.
[a]Case and death counts include all positive cases and deaths of COVID-19 reported between January 21, 2020 and May 22, 2020.
[b]To account for missing values on the PCP variable, three imputation approaches were explored: 1) filtering out missing values, 2) imputing missing values with a value of 0, and 3) imputing missing values with the mean value of PCP among counties with non-missing data.
Abbreviations: n.a. = not applicable; ICU = intensive care unit; PCP = primary care physician; BMI = Body Mass Index

Our inability to match PCP is tied to the fact that the data file published by the original authors does not include the PCP variable. We were able to use the provided reference to collect this data from the County Health Rankings 2020. However, those rankings contain counties with missing values. Saffary et al. provided no explanation of how they addressed these missing data. We investigated three alternative procedures to address this issue - filtering, zero imputation, and mean imputation. Our findings suggest Saffary et al. simply omitted missing values. Omitting those counties reduced our county sample from the 3,108 reported by the authors to the 2,962 presented above. We used this county dataset with 2,962 for our analysis involving the PCP measure. For all other analyses, we used the original dataset containing 3,108 counties.

As expected we were able to bitwise reproduce the bivariate global Moran's I values of the original authors for nearly all variables. However, we did find some discrepancies. Saffary et al calculated I = 0.001 for PCP-deaths and I = -0.01 for PCP-cases in their original analysis. We calculated I = 0.0593 for PCP-deaths and I = 0.0346 for PCP-cases in our reproduction. The bivariate Moran's I associating non-Hispanic white values with deaths and cases are identical, but switched. The value for deaths in the original was -0.0137 with cases at -0.203, while the reproduction had cases of -0.137 and deaths at -0.203. The difference appears to be the result of a transcription error in the original publication.

Table 2. Results from univariate and bivariate global Moran's I in original analysis and reproduction analysis

	Original		Reproduction	
	Deaths	**Cases**	**Deaths**	**Cases**
Cases per 100,000 population	n.a.	0.228****	n.a.	0.2286****
Deaths per 100,000 population	0.476****	n.a.	0.4766****	n.a.
Health Factors				
Number of ICU beds	0.15****	0.08****	0.1500***	0.0765***
PCPs per 10,000 population	0.001	-0.01	n.a.	n.a.
PCPs per 10,000 population (filtered)[a]	n.a.	n.a.	0.0593****	0.0346***
PCPs per 10,000 population (zero imputed)[a]	n.a.	n.a.	0.0583****	0.0283***
PCPs per 10,000 population (mean imputed)[a]	n.a.	n.a.	0.0560****	0.0300*
Adult Obesity (BMI > 30)	0.005	0.01	0.0049	0.0091
Diabetes	0.03	0.01	0.0262	0.0122
Uninsured	0.005	0.012	0.0055	0.0128
Flu vaccinations	-0.004	-0.006	-0.0035	-0.0057
Race Factors				
Black American	0.264****	0.174****	0.2643****	0.1745****
Hispanic	-0.002	0.008	-0.0018	0.0079
Non-Hispanic White	-0.137****	-0.203****	-0.2033****	-0.1371****

Sources: Saffary et al., 2020; County Health Rankings, 2020
Notes: Estimates are percentages out of 100, unless otherwise noted. All estimates were calculated using the sample of 3,108 counties in the contiguous U.S.
[a]To account for missing values on the PCP variable, three imputation approaches were explored: 1) filtering out missing values, 2) imputing missing values with a value of 0, and 3) imputing missing values with the mean value of PCP among counties with non-missing data.
* $p < 0.05$, *** $p < 0.001$, **** $p < 0.0001$
Abbreviations: n.a. = not applicable; ICU = intensive care unit; PCP = primary care physician; BMI = Body Mass Index

Our reproduction identified similar non-random distributions of cases and deaths across the contiguous United States (Figure 2). Figures 2A and 2B reveal identical 'High-High' areas, however the reproduction (Figure 2B) also contains 'High-Low' clusters that are not present in the original. Similarly, Figures 2C and 2D show similar 'High-High' clusters, but the reproduction (Figure 2D) identifies more clusters than the original analysis. There are also no 'Low-Low' clusters present in the reproduction. Overall, we were able to partially reproduce the analyses conducted to assess the first family of hypotheses.

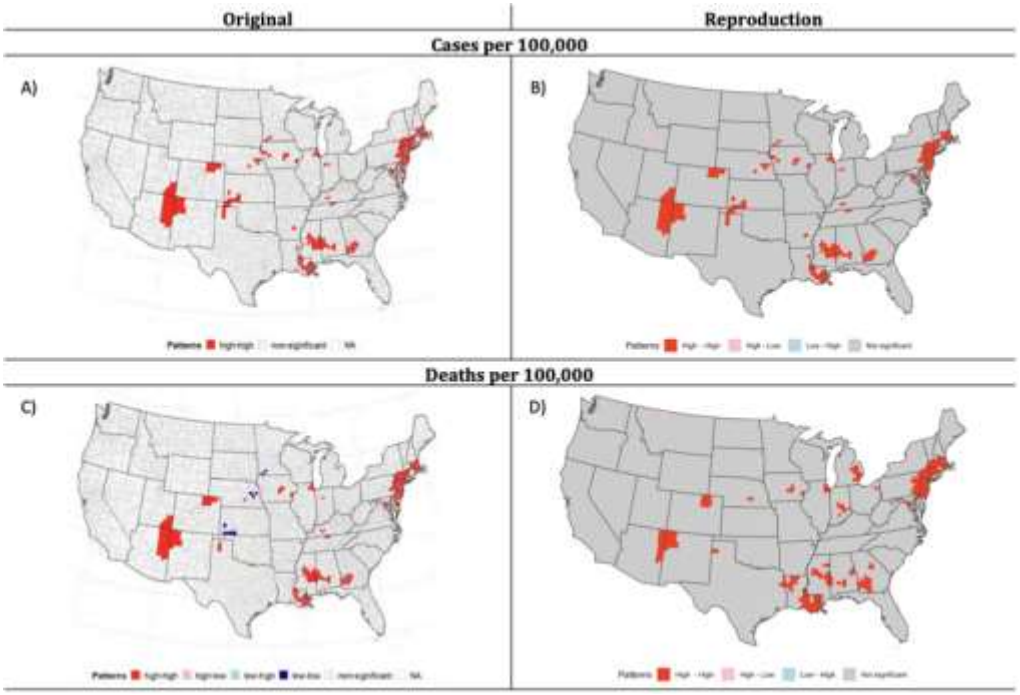
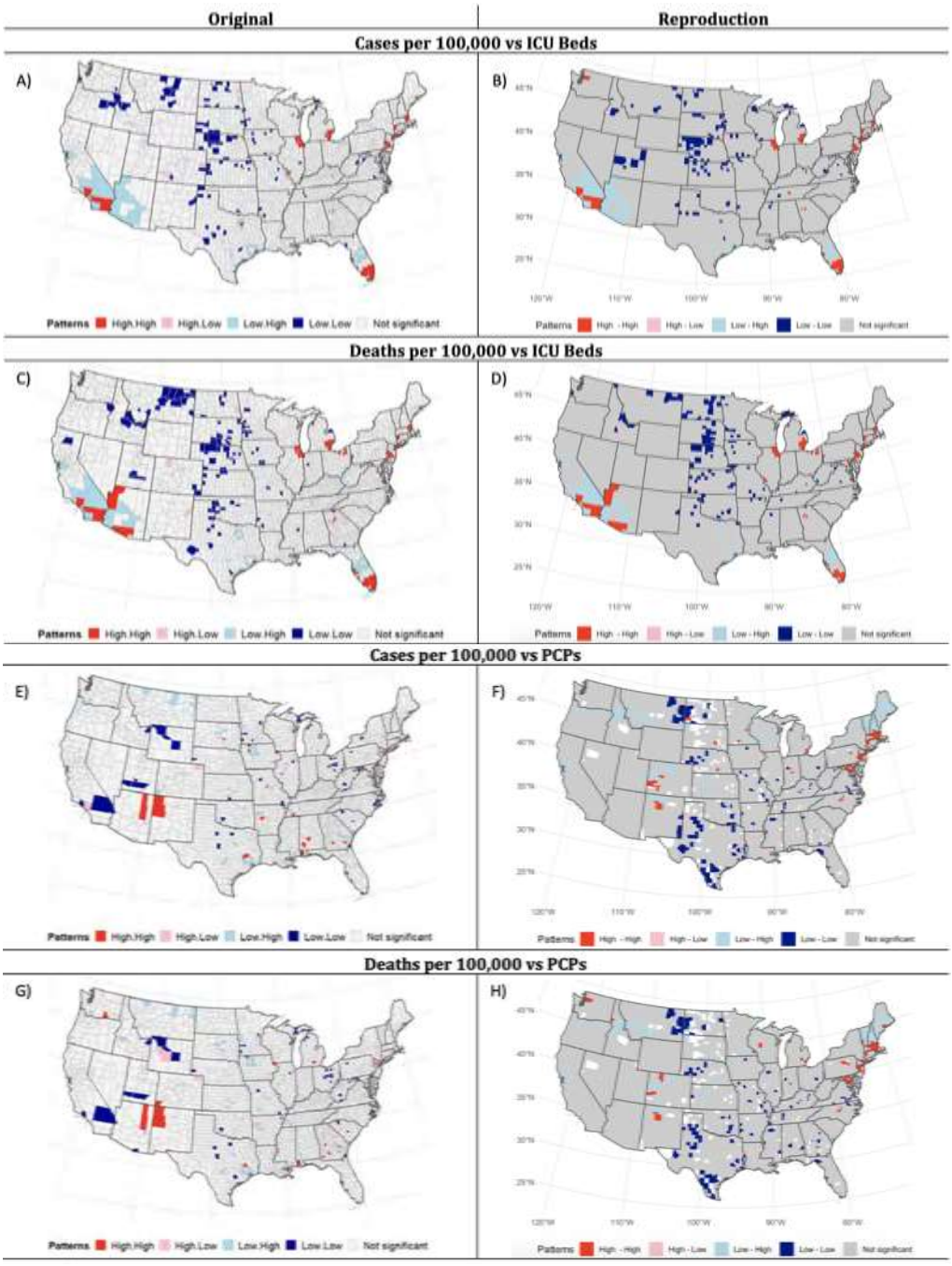


Figure 2. Results from univariate local Moran's I in original analysis and reproduction analysis
Note: Maps show statistically significant clusters, where statistical significance is based on $p < 0.05$.

Reproduction Results for the County-Level Health Factors (H2):

We were able to partially reproduce the relationship between county-level health factors and COVID-19 prevalence. In particular, the reproduction of the analyses of ICU beds and of diabetes produced results that were very similar to the original results (Figure 3, Panels A-D and I-L). However, we were unable to reproduce the associations between the rate of PCPs and COVID-19 reported in the original analysis (Figure 3, Panels E-H).



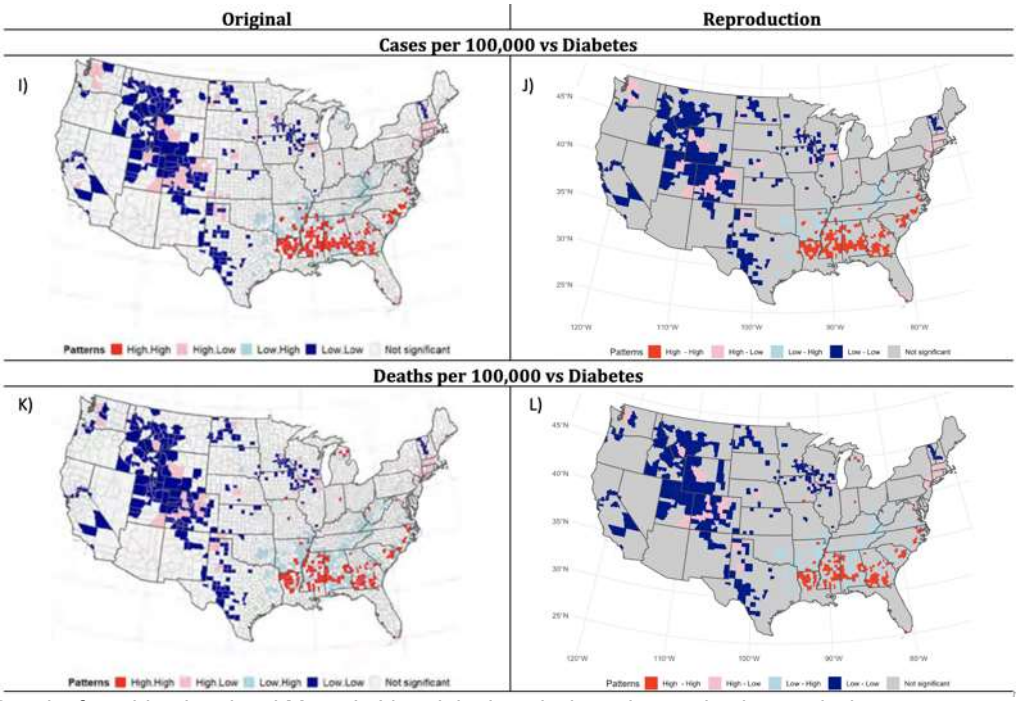
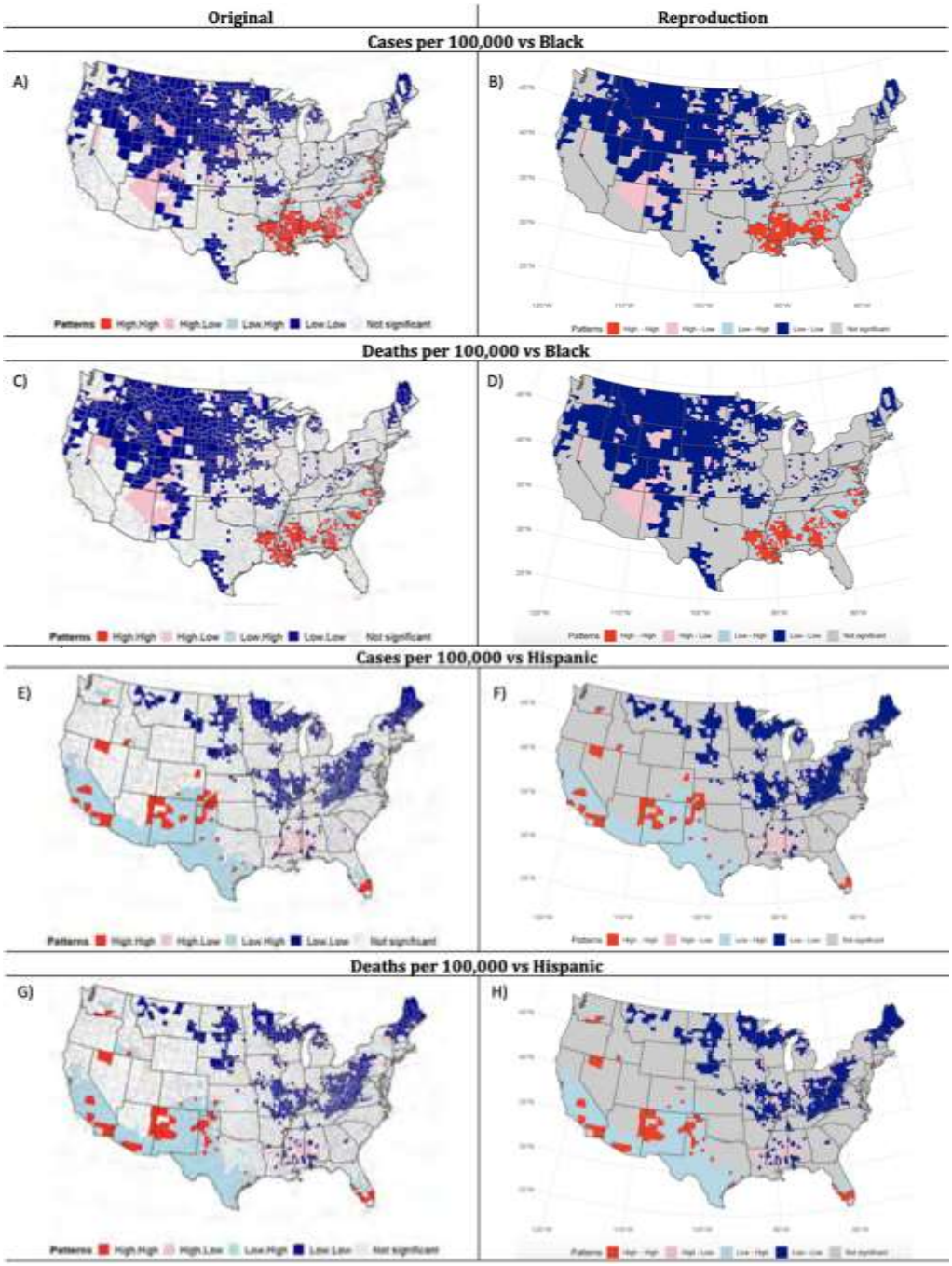


Figure 3. Results from bivariate local Moran's I in original analysis and reproduction analysis
Notes: Maps show statistically significant clusters, where statistical significance is based on $p < 0.05$. Counties omitted from the analysis appear in white in Figs. 3F and 3H.
Abbreviations: ICU = intensive care unit; PCP = primary care physician

Based on Saffary et al.'s results, we anticipated finding 'High-High' clusters in parts of Arizona and New Mexico, with a handful of 'Low-Low" clusters scattered across the western U.S (Figures 3E and 3G). However, our reproductions show 'High-High' clusters along the East Coast, and 'Low-Low' clusters concentrated in the central U.S. These discrepancies are not surprising given we were unable to reproduce the descriptive statistics for the PCP variable.

Reproduction Result for the Demographic Analysis (H3):

We were able to reproduce the local bivariate Moran's I analyses that explored the associations between COVID-19 and demographic characteristics (Fig. 4). Although not perfectly identical, there is strong visual similarity between the original maps and those of the reproduction for all racial and ethnic groups. In particular, large 'High-High' clusters of COVID and percent Black are observed throughout the southeast, while 'Low-Low' clusters are found across the upper Midwest. A series of 'Low-High' clusters along with a handful of 'High-High' clusters of CVOID and percent Hispanic are present in the southwest, with 'Low-Low' clusters across the northern U.S. Finally, 'Low-Low' clusters of Covid and percent non-Hispanic white were observed across the southeast, with 'Low-High' clusters throughout the Midwest and Northeast.


Original
Reproduction
Cases per 100,000 vs Black
A)
B)
Deaths per 100,000 vs Black
C)
D)
Cases per 100,000 vs Hispanic
E)
F)
Deaths per 100,000 vs Hispanic
G)
H)
Patterns
High-High
High-Low
Low-High
Low-Low
Not significant

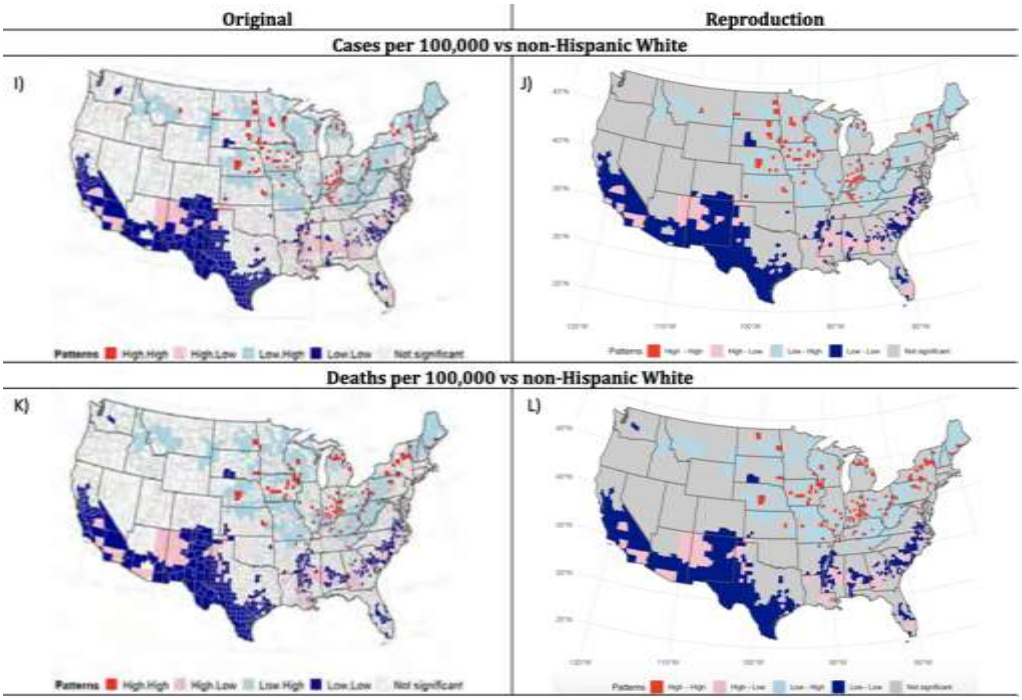


Figure 4. Results from demographic bivariate local Moran's I in original analysis and reproduction analysis
Notes: Maps show statistically significant clusters, where statistical significance is based on $p < 0.05$.

Reproduction Results for Tests Conducted with Corrections for Multiple Testing (H1-H3):

In addition to the analyses presented above, the original authors conducted all of their hypothesis tests making corrections for multiple testing. The authors used two alternative techniques in each case - (1) a Bonferroni adjustment and (2) a false-discovery rate (FDR) adjustment.

Because the FDR adjustment is a less conservative adjustment than the Bonferroni adjustment, both the original findings and the reproductions suggest that more counties have statistically significant estimates when using an FDR as opposed to a Bonferroni adjustment. In the original analysis, Saffary et al. found substantial clustering in the bivariate LISA analyses when examining the relationship between COVID and race, and, to a lesser extent, in the relationship between COVID and diabetes.

Reproduction of Adjusted H1: After applying a Bonferroni adjustment, the original analysis indicated that only a small cluster of counties in the U.S. had statistically significant low case rates after, and that no clusters emerged with respect to death rates. We were unable to reproduce either of these results (Figs. 5 and 6). In particular, we found that several of the high case and death rate clusters found in the unadjusted analysis persisted after applying a Bonferroni adjustment, and there was no evidence of clusters containing low case rates. Similarly, although fewer counties were found to be statistically significant in the adjusted bivariate LISA analyses, in all cases, we still find substantial clustering in these relationships, contrary to the original authors' findings.

Similar discrepancies emerge when comparing the supplemental results adjusted using an FDR adjustment to the unadjusted results for the univariate LISA analysis in the original paper. The authors find clusters of low case rates after applying an FDR adjustment that were not present in the unadjusted analyses. We were unable to reproduce these adjusted results.

It is worth noting that the maps of adjusted case rates presented in the supplement differ substantially from the map of the unadjusted case rates presented in the main paper. Because the Bonferroni adjustment only affects the threshold used for determining statistical significance, we would expect any county in the adjusted map to also be statistically significant in the unadjusted map, however this is not the case in the original paper. These stark differences between the adjusted and unadjusted maps presented by Saffary et al., suggest that the original authors may have implemented other changes in their supplemental analysis, beyond simply correcting for multiple testing, however, because no other analytic changes were discussed, we were unable to reproduce their results.

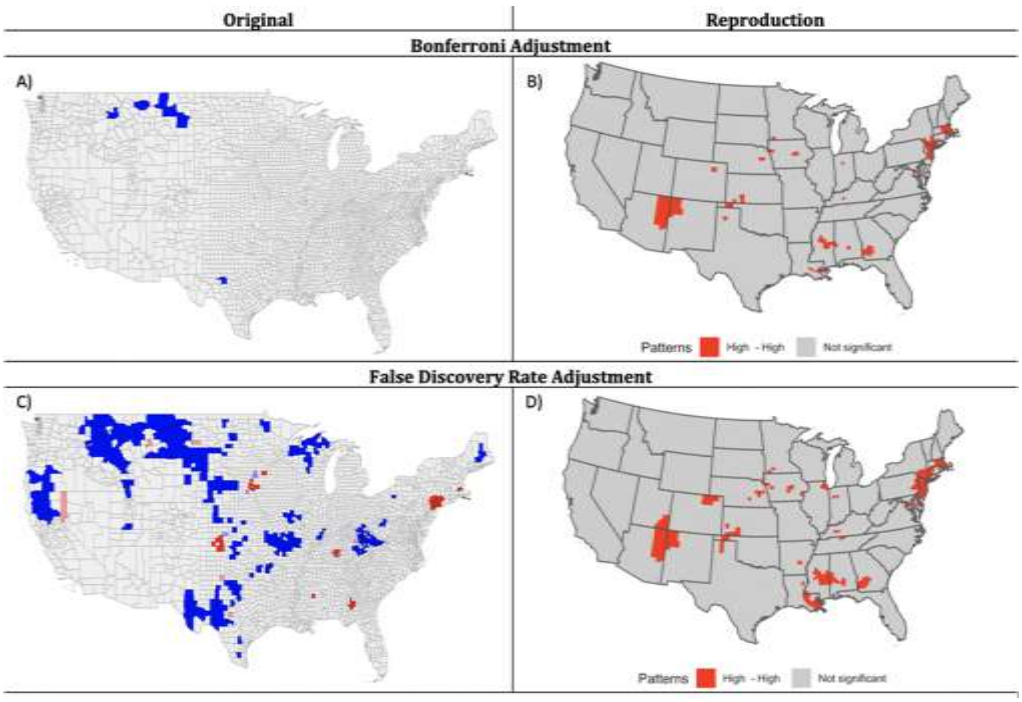


Figure 5. Results from adjusted univariate local Moran's I for case rate in original analysis and reproduction analysis

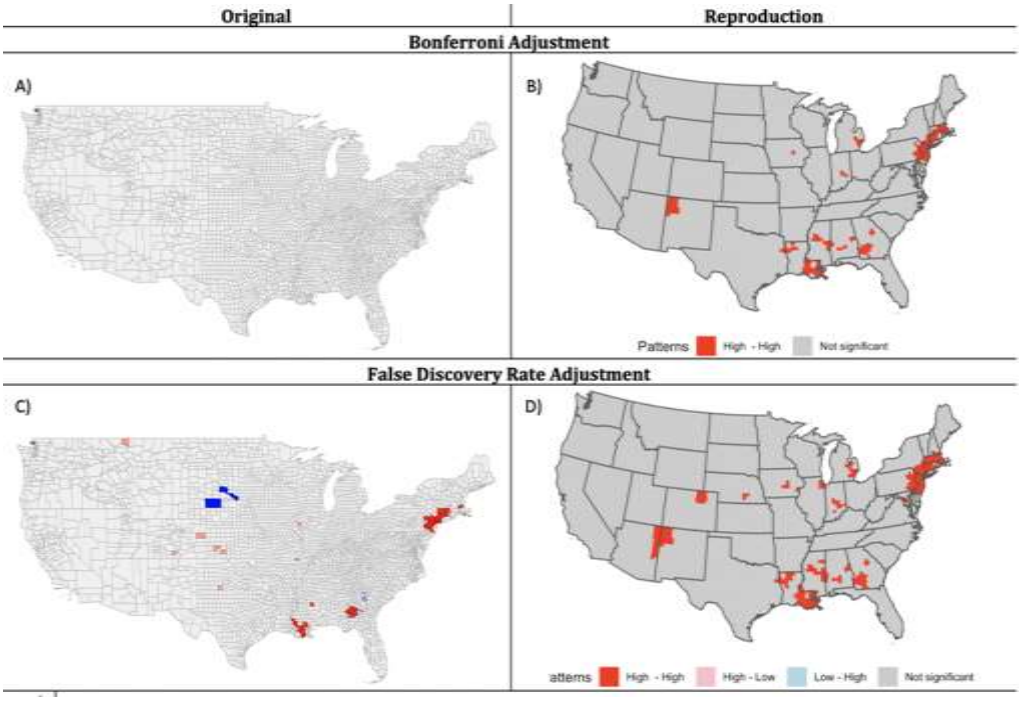


Figure 6. Results from adjusted univariate local Moran's I for death rate in original analysis and reproduction analysis

Reproduction of Adjusted H2: After applying a Bonferroni adjustment, Saffary et al. found that none of the local bivariate relationships between county-level health factors and COVID-19 rates were statistically significant. Contrary to these findings, our reproductions indicate that several clusters along the east and west coasts are present after applying an FDR adjustment for analyses examining the relationship between COVID and the number of ICU beds and primary care physicians (Figs.7-10). Our reproduction of the adjusted diabetes maps using FDR are similar to those produced by Saffary et al., although in our analysis, we observe a larger cluster of counties with low rates of COVID cases and deaths by diabetes in the western U.S. than were found by the original authors (Figs. 11-12).

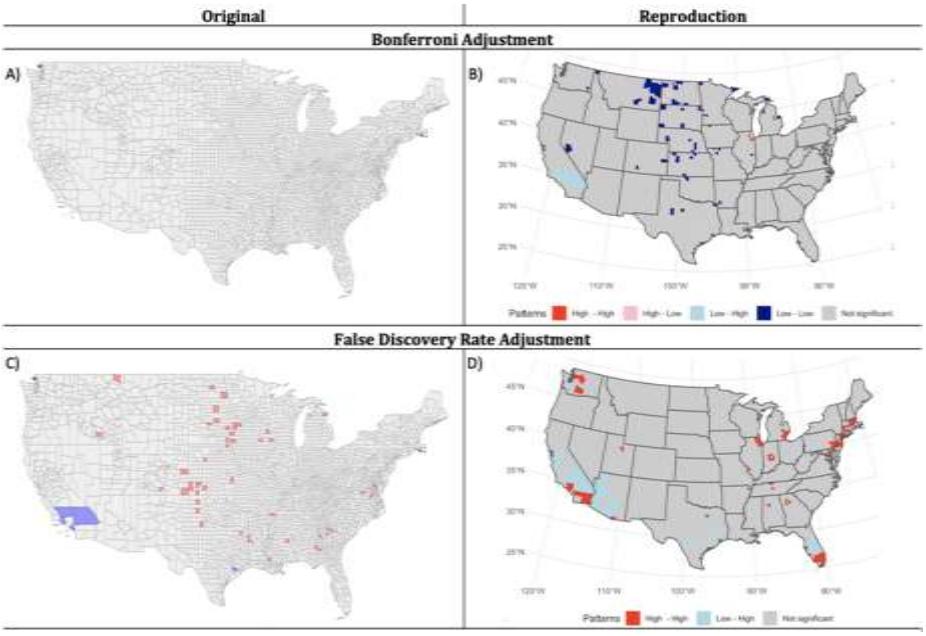


Figure 7. Results from adjusted bivariate local Moran's I for case rate vs ICU in original analysis and reproduction analysis

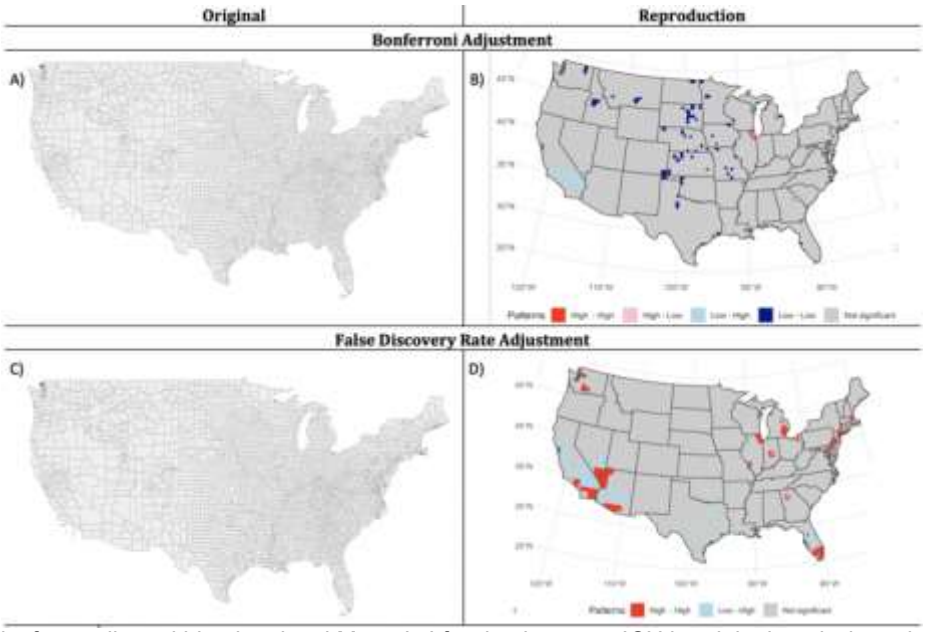


Figure 8. Results from adjusted bivariate local Moran's I for death rate vs ICU in original analysis and reproduction analysis

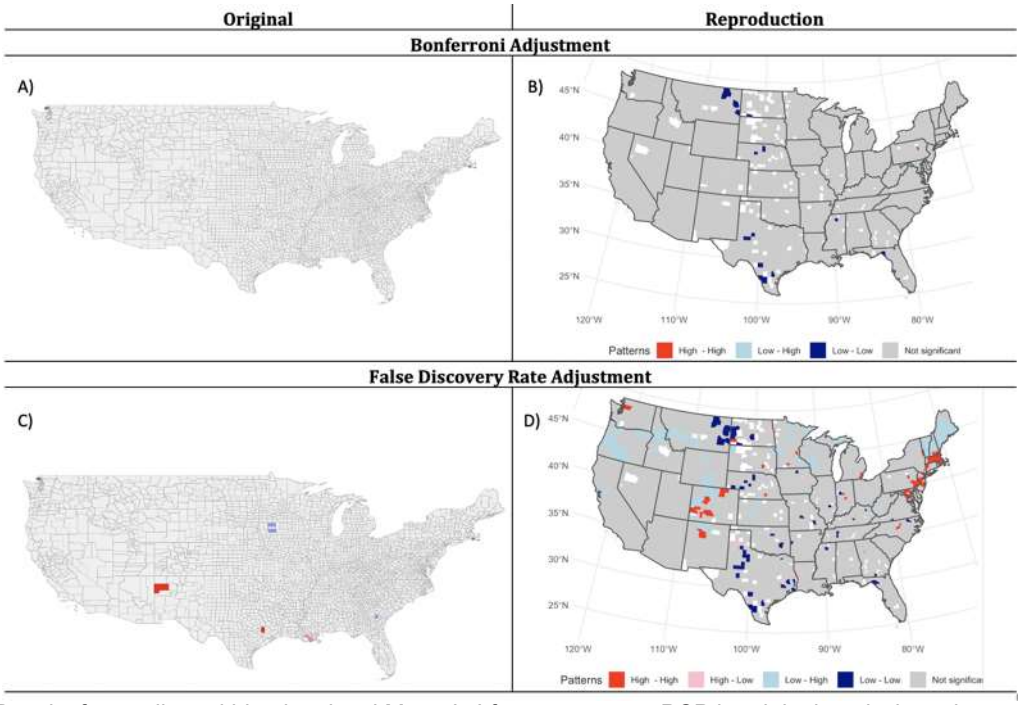


Figure 9. Results from adjusted bivariate local Moran's I for case rate vs PCP in original analysis and reproduction analysis

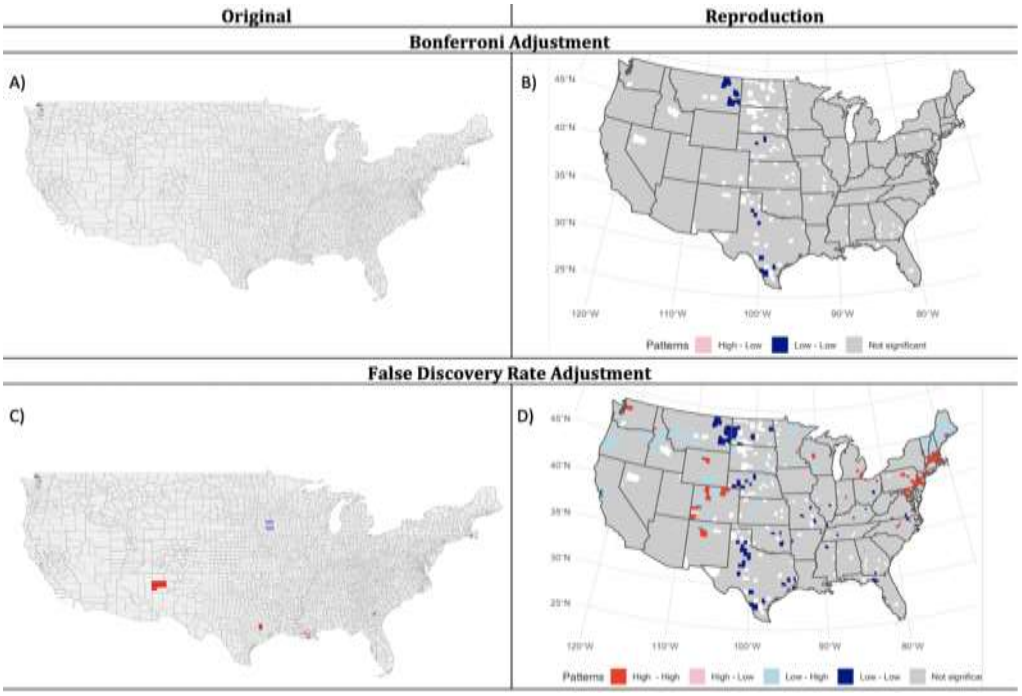


Figure 10. Results from adjusted bivariate local Moran's I for death rate vs PCP in original analysis and reproduction analysis

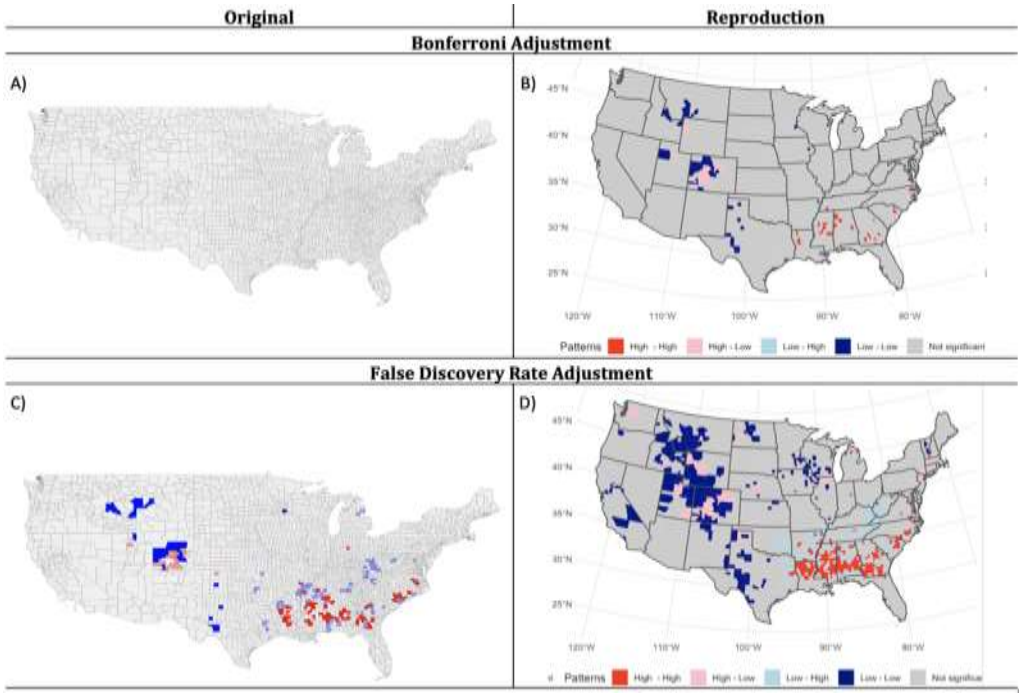


Figure 11. Results from adjusted bivariate local Moran's I for case rate vs diabetes in original analysis and reproduction analysis

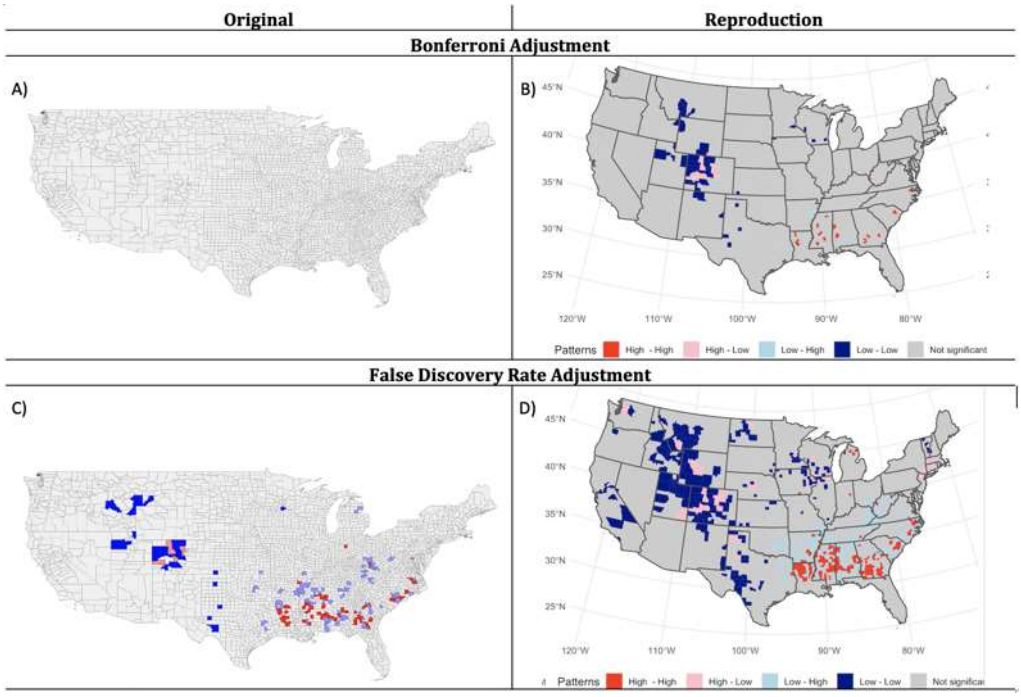


Figure 12. Results from adjusted bivariate local Moran's I for death rate vs diabetes in original analysis and reproduction analysis

Reproduction of Adjusted H3: Our reproductions of FDR adjusted bivariate LISA analyses more closely resembled those in the supplement than was the case with the Bonferroni adjusted findings. Although not identical, there appears to be strong similarities between our reproductions and the FDR adjusted results presented by Saffary with respect to the following independent variables: percent white, percent Hispanic, and percent Black (Figs. 13-18).

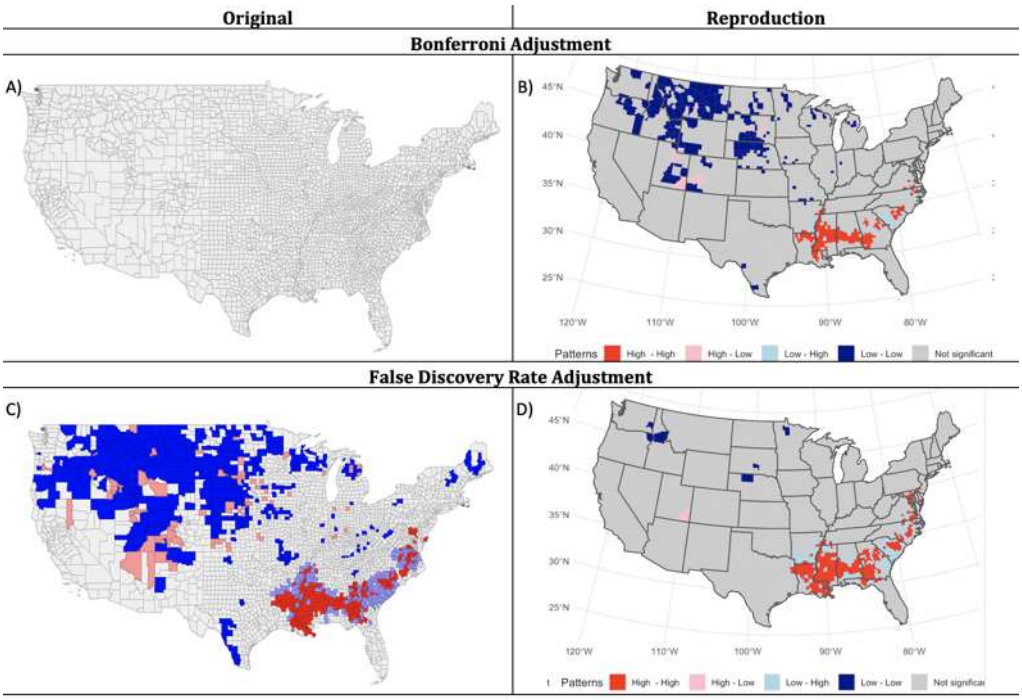


Figure 13. Results from adjusted bivariate local Moran's I for case rate vs black in original analysis and reproduction analysis

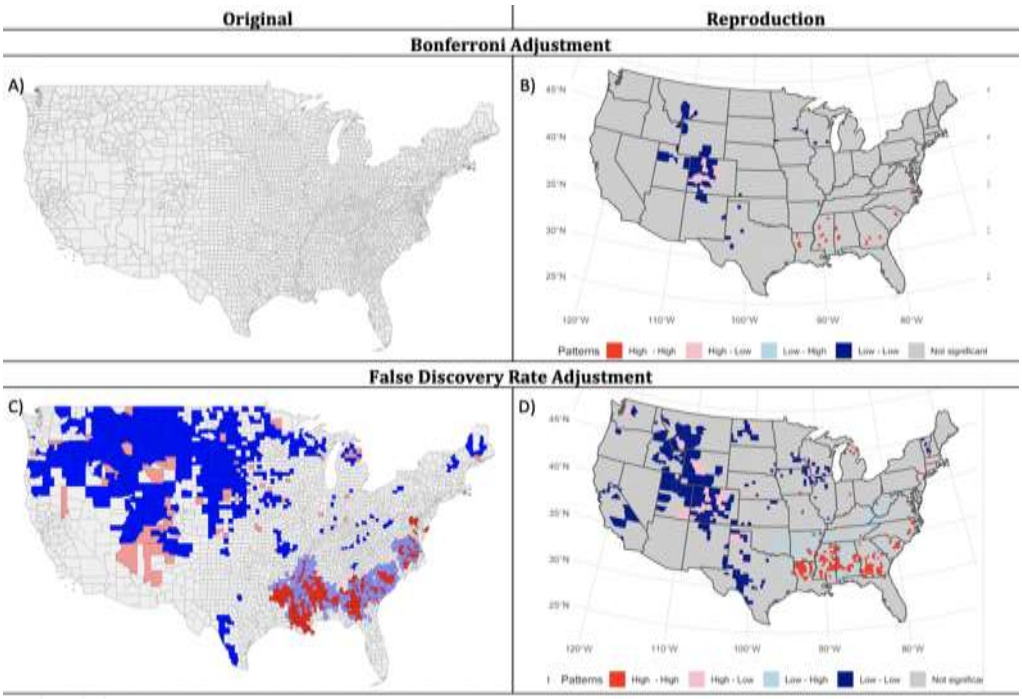


Figure 14. Results from adjusted bivariate local Moran's I for death rate vs black in original analysis and reproduction analysis

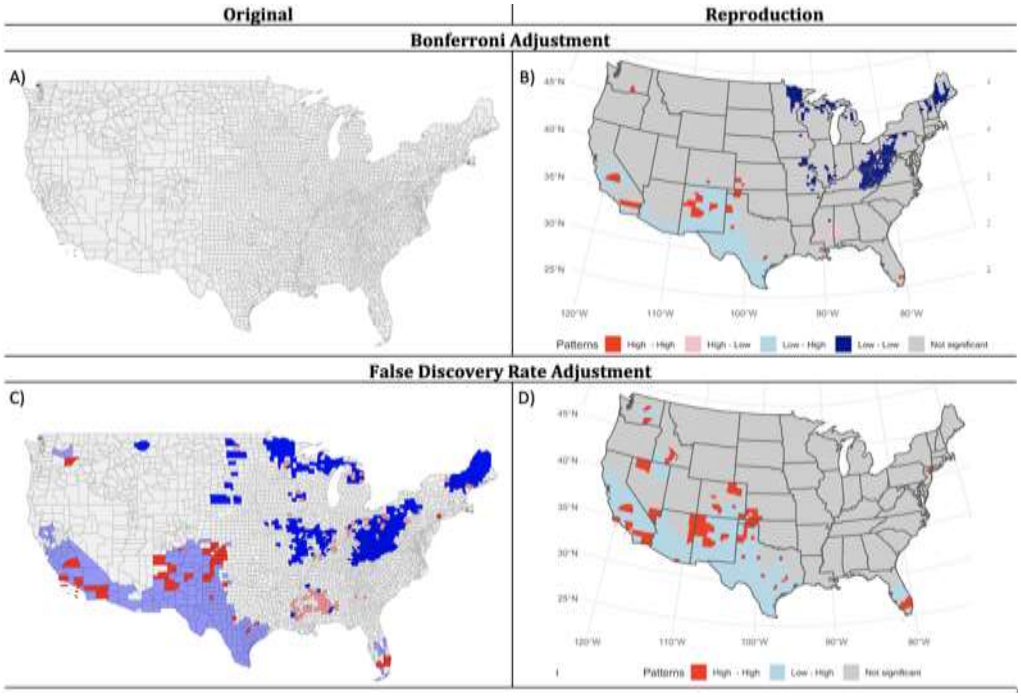


Figure 15. Results from adjusted bivariate local Moran's I for case rate vs Hispanic in original analysis and reproduction analysis

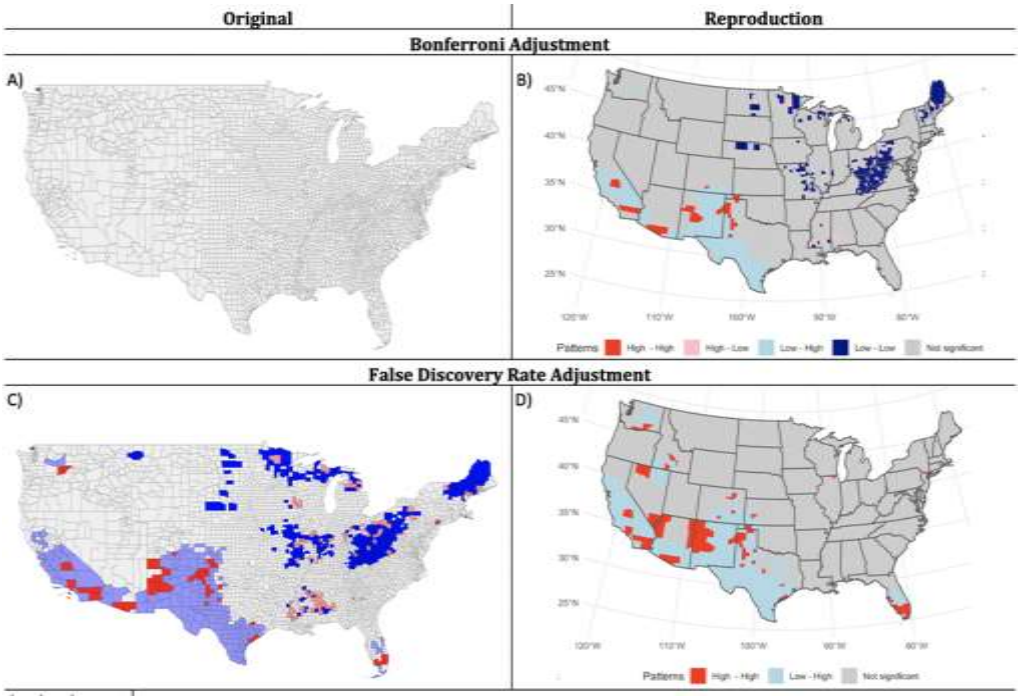


Figure 16. Results from adjusted bivariate local Moran's I for death rate vs Hispanic in original analysis and reproduction analysis

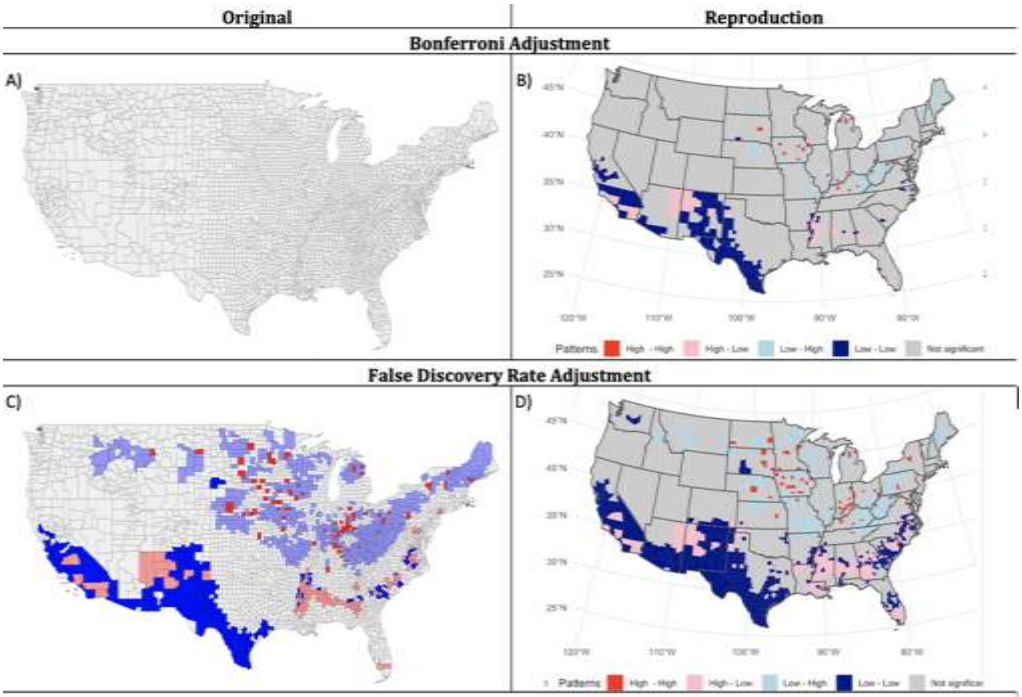


Figure 17. Results from adjusted bivariate local Moran's I for case rate vs non-Hispanic white in original analysis and reproduction analysis

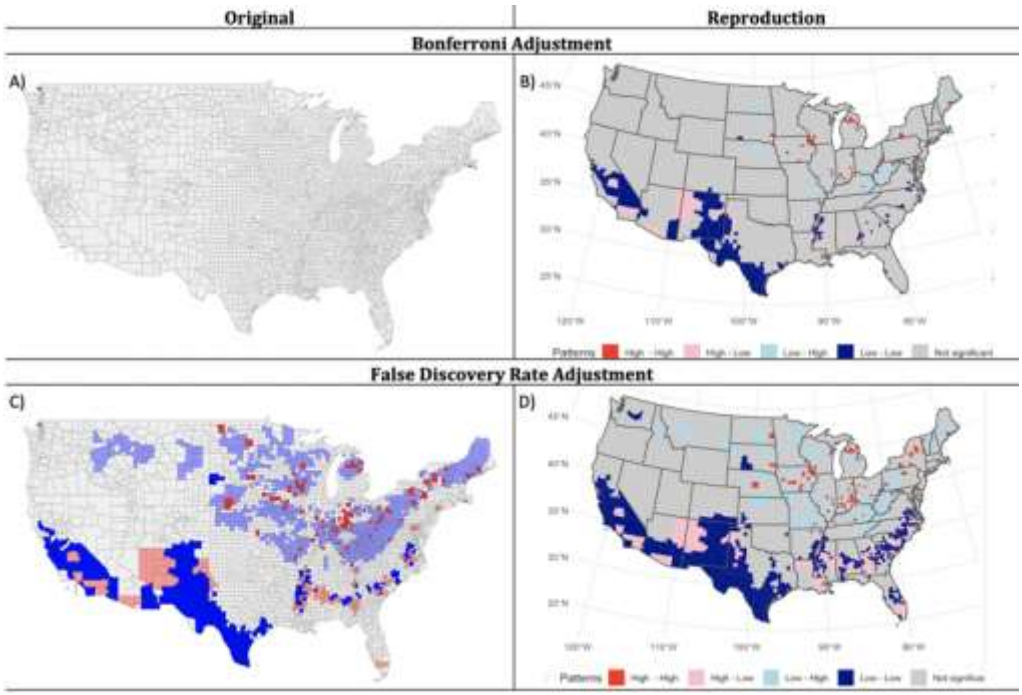


Figure 18. Results from adjusted bivariate local Moran's I for death rate vs non-Hispanic white in original analysis and reproduction analysis

Unplanned Deviations from the Protocol

We deviated from our original protocol in several ways during the reproduction process. First, the original authors used a sample size of 3,142 counties (including Hawaii and Alaska) when reporting the mean, median, and IQR for each variable. However, the global bivariate and univariate Moran's I and all local Moran's I analyses used a smaller sample size that omitted Hawaii and Alaska. While a note accompanying the original Table 2 indicates that the descriptive statistics used the larger set of counties, it wasn't immediately obvious that the global bivariate/univariate Moran's I statistics were calculated using the smaller sample size. As a result, we tested both county sets when defining the weight matrix for our spatial statistical analyses and determined the appropriate sample sizes for each analysis by comparing the results with those in the original paper. We determined that the original study excluded US territories, Alaska, and Hawaii from all analyses except the calculation of the means, medians and IQRs presented in Table 2 in the original paper.

Second, the published data file does not include the PCP variable. We were able to collect this data from the County Health Rankings 2020, but the variable contains missing values. Saffary et al. provided no explanation of how they addressed these missing data. Our investigation of the mean, median, and IQR in Table 2 suggests that they omitted missing values.

Finally, the original authors did not provide enough information in their paper to exactly reproduce their analyses. The authors did not provide the specific R packages used nor the parameters of their analyses. Without this information, we were forced to test alternatives and make educated guesses as to their procedure. We used bivariate Moran's I syntax available through Rafael Pereira's github account, as there is no R package that currently implements this statistic. We were able to reproduce the majority of Table 2 global bivariate results and also the maps, suggesting that we used the correct procedure. Similarly, it appears that the original authors used a permutation approach as their inferential framework. However, this choice was not described in the paper.

Discussion

Procedural Concerns

We were able to partially reproduce the original analyses. While the authors did provide the majority of their data, they failed to make their code available which complicated our reproduction of their results. The decision to not share the full details of their analytical procedure led to many of the issues presented in this report (e.g., missing data handling, statistical significance procedures). One key instance worthy of additional emphasis is the authors' implementation of the Bonferroni and FDR adjustments in the supplemental materials. These adjustments are crucial to address the problem of selective inference. However, when we attempted to reproduce the Bonferroni adjustment we were unable to reproduce any of the results presented by the original authors. While the Bonferroni adjustment is more conservative than the FDR adjustment, we suspect that Saffary et al. may not have accurately implemented this adjustment, given the large departures between the two adjustment approaches in the original paper, and the inability to reproduce the Bonferroni adjustments using two separate approaches. Additionally, we found it odd that the authors chose not to age standardize their COVID case and death counts. This decision is problematic for inference since COVID-19 is known to have differential impacts across the age curve, with disproportionately greater risk and impact for older populations.

We likewise found the selection of some of the health factors puzzling. The authors chose to use BMI to represent obesity at the county scale. However, BMI is known to be a crude measure of obesity as it only accounts for an individual's height and weight. The ICU beds measure was similarly problematic. The

authors failed to standardize the count of ICU beds by county population. As a result, this variable may essentially be a measure of county population due to the strong positive correlation between these factors. The visual correspondence between the distribution maps of population centers, COVID cases and death, and ICU beds makes the issue clear. The measure of those uninsured in a county is also a concern. The measure used in the study was taken before the pandemic. However, as the pandemic began many individuals lost their jobs and simultaneously lost their insurance. It is unclear if the spatial pattern of this shift in the number of uninsured people would follow the pre-existing pattern of those uninsured. It is possible that change in those uninsured followed the pattern of job losses around the country. The extent to which this change would impact the analyses performed is unclear.

Finally, we expected the authors to express more uncertainty when presenting their results. The data used in this analysis covers the first 120 days of the COVID-19 outbreak in the United States. However, there was a well documented lack of COVID-19 testing infrastructure as the disease first emerged. In addition, we know that many COVID cases are asymptomatic. Both of these factors complicated detection throughout the pandemic, but were likely particularly impactful during the time frame under study. As a result, the actual measures of cases and deaths used as the basis for this analysis may be poor measures of actual disease prevalence and mortality. There is no ideal way to correct these issues. However, contextualizing inferences in the uncertainty inherent in the data is crucial for the interpretation of results.

Statistical and Inferential Concerns

Our primary concern with the credibility of the authors' original analysis stems from three issues - 1) the atomistic fallacy, 2) implementation of the Bi-variate Moran's I, and 3) the problem of selective inference.

First, the authors' inferences should be tempered because their research design may be subject to the atomistic fallacy. The motivations for variable selection presented by the authors are rooted in individual-level processes and relationships. However, the variables used in this analysis are measured at the county scale, which means those motivations are implicitly scaled to the group level. This scaling may be fallacious. Moreover, use of the Bivariate Moran's I to measure associations across space extends these assumptions across counties. However, it is not clear, for example, that the reasoning supporting why an individual person of color might be at a higher risk of contracting COVID would extend to all people of color in a county, or to all people of color in counties surrounding a county with COVID cases.

Second, the authors' implementation and interpretation of the Bi-variate Moran's I is inconsistent throughout the paper, which raises concerns about the inferences based on this statistic. Contrary to how many of the results are described in-text, Saffary et al. examined the extent to which rates of COVID-19 are spatially correlated with varying levels of health and demographic factors in surrounding counties. This analytic decision, while statistically valid, is opposite of how they describe their analysis. In the paper, the authors discuss COVID-19 rates as the outcome of interest and health and demographic factors as independent variables. Instead, the Moran's I statistic that they calculate is the product of each focal county's rate of COVID-19 and the spatial lag of adjacent counties' health and demographic compositions. Given the inconsistency with which Saffary describes their results, it would be easy for a reader to misinterpret the implications of this analysis.

Third, Selective inferences occur when researchers focus inferences on a subset of findings that were identified as interesting only after viewing the data. Observed selective inferences occur when researchers implement many statistical tests, fail to account for the effects of multiple testing, and then emphasize only a subset of their results. Saffary et al. did include adjustments for multiple testing as a supplement to their analysis. Put simply, we believe the supplemental materials should be the results presented in the paper. Making appropriate adjustments for the large amount of multiple testing done

during this analysis is key to making reliable inferences. Using a p=0.05 significance threshold, we would expect 156 'significant' results in a set of 3,105 tests even when no relationship existed in any county.

In our opinion, the results presented by Saffary et al. should be treated and presented as an exploratory spatial data analysis of potential bivariate associations contextualized by the numerous uncertainties that exist in the data and research design. Using this analysis for causal inference, prediction, or extrapolation across space/time would be an error. It is not clear if the COVID data and death counts reflect actual prevalence or mortality during the period studied. It is likewise not clear if many of the health-related variables measure the constructs they are intended to represent.

Presented as an exploratory analysis the inferences presented in the original paper may not be credible. An adjustment for multiple testing must be made, as was done in the supplemental materials, to avoid the pitfalls of selective inference. However, even adjusting for multiple testing the inferential logic of the original analysis may be flawed because the authors do not account for covariance with other potential predictors. For example, minority proportion is highly correlated with population density across the US. Many of these analyses may simply be explained by population densities and the extreme spatial concentration of COVID 19 early in the pandemic in a few large cities.

References

1) Pereira, R. "Bivariate LISA". Github. Accessed March 28, 2021 https://gist.github.com/rafapereirabr/5348193abf779625f5e8c5090776a228
2) U.S. Census Bureau. *Population Estimates 2018 (county)*. Accessed March 10, 2021. https://www.census.gov/data/developers/data-sets/popest-popproj/popest/2018.html
3) Robert Wood Johnson Foundation, County Health Rankings & Roadmaps. In: *Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute* 2020.