**A retrospective spatial-epidemiologic and socio-demographic analysis of reported COVID-19 cases and testing in a large population center (Orange County) in Southern California, U.S.A. (March through June 2020)**

**Daniel M. Parker1,2; … Vladimir Minin; Veronica Vieira; Scott Bartell??; Bernadette Boden-Albala1,2; Tim Bruckner1**

**[authorship order may change based on contributions]**

1. **Population Health and Disease Prevention; University of California, Irvine**
2. **Epidemiology; University of California, Irvine**

**INTRODUCTION**

The United States is currently experiencing an epidemic of respiratory disease (coronavirus disease 19 or COVID-19), caused by a novel coronavirus (SARS-CoV-2) that has reached pandemic proportions. As of the time of this report there are a total of 3,698,209 confirmed cases and 139,960 reported deaths in the U.S. [1,2].

As the disease has swept the globe, it has manifested in different ways across social, economic, and demographic groups [1–3], with regard to apparent risk of infection, disease severity, and mortality. Co-morbidities and age both appear to play a major role in risk of severe disease, with hospitalization rates increasing in older age groups and the majority of hospitalized patients in the U.S. having comorbidities [4]. Many hospitalized patients require ventilators [5], and there is a general high mortality rate among those who are hospitalized [6].

Current case fatality statistics are based on questionable data and sparse immunological surveys. This circumstance has called into question the accuracy of reported case fatality rates. Availability of tests and operational barriers remain limiting factors for diagnosis in many places, including the U.S. [7]. In some places (e.g. Lombardi Region of Italy, New York City in the U.S.), healthcare facilities have been overwhelmed by the surge in cases and have fallen short of needed ventilators and ICU beds, resulting in massive casualties [8,9].

In California, where social distancing measures have been mandated by law since March 19, 2020 [10] [3], exceptions exist for individuals working in “essential jobs”, which have been relatively vaguely defined but include medical professionals, food providers, delivery agencies, public officials, contractors, building laborers, landscapers, etc. [who else??] [11]. The social, demographic, and economic attributes of individuals working in the construction, maintenance, food and delivery industries tends to differ from those who are more capable of staying at home during the social distancing campaign. This stratification of occupation by socioeconomic level may lead to increased risk of infection among subsets of the population that are unable to practice social distancing. Furthermore, this population may face disproportionate barriers to accessing healthcare, including COVID-19 testing and treatment.

Here we present results of a spatial epidemiological analysis of COVID-19 in Orange County, California, which ranks second highest in population density for all state counties. An estimated 3.19 million people live in approximately 800 square miles of land. The Orange County Health Care Agency (OCHCA) reveals large disparities in income, education, housing, safety, and health outcomes [12]. Almost half of Orange County residents over the age of five speak a language other than English at home. Additionally, many within the Hispanic/Latino and Asian communities of Orange County live below the poverty level (17.9% and 16%, respectively) and face challenges in education, household income, access to healthcare, obesity levels, and life expectancy [12,13]. The relatively small land area, high population density and diverse population of Orange County provides a unique opportunity to explore potentially important social, economic, and demographic predictors of COVID-19 test positivity and their contribution to spatial patterns in testing and reported cases.

**METHODS**

***Data.*** Case data were provided by OCHCA and consisted of individual level records of both negative and positive case reports based on results of polymerase chain reaction (PCR) diagnostic tests conducted from March through June 30, 2020. OCHCA receives the testing data from the California Reportable Disease Information Exchange (CalREDIE), an infectious disease surveillance system implemented by the California Department of Public Health (CDPH) [14]. The available data, which include information on diagnosis date, age, gender, race, ethnicity, and zip code of the individual taking the test, was obtained from reported lab testing if provided by the lab [15]. For individuals who were repeatedly tested after testing positive, only the first positive diagnosis is included. The total number of case reports over the study period is ##. Information on analytic variables show less than XX% with missing or implausible values.

***Spatial analysis.*** Reported cases and number of tests were aggregated at the zip code level, by week. There were XXX zip codes included in the analysis. The outbreak in Orange County began in the XX week. For mapping, cases were further aggregated into months, corresponding to the final three weeks in March, all of April, May, and June. Case incidence was mapped as positive cases per 100,000 population per week. Testing intensity was mapped as total number of tests per 100,000 people per week. Here a positive case is defined as any case reported to the California State health authorities and the total number of tests were defined as any test result that was reported to the California State authorities.

Formal testing of spatial autocorrelation was done using the global Moran’s I statistic and the local version (Local Indicators of Spatial Autocorrelation or LISA) for both the March, April, and May datasets and both for case incidence and testing intensity. The global Moran’s I gives an indication of the absence or presence of spatial clustering whereas the LISA statistics give an indication of local clustering (which can be mapped to indicate where clustering is occurring).

***Analysis of test positivity.*** Median household income (HHI) by zip code was included as a potential predictor of case positivity. The data were extracted from <http://www.ochealthiertogether.org/> and come from the last relevant American Community Survey of 2018. We plotted median income at the zip code level for all reported tests, by test result, and over time.

We then used a logistic regression analysis to explore demographic and economic predictors of the odds of testing positive. Independent variables in our model included age group, gender, race, median HHI at the zip code level, zip code population density, and testing month. We calculated adjusted odds ratios (AORs) and 95% confidence intervals (CI) for covariates and included zip code as a random intercept in the model in order to account for potential baseline differences in risk of testing positive (e.g. different testing intensities) across zip code areas. Descriptive reports at the outset of the epidemic in Orange County noted that test availability appeared to concentrate in higher-income neighborhoods. For this reason, an interaction term between month and median HHI at the zip code level was included in the regression analysis.

***Software.*** Maps were created using QGIS version 3.4.9. Tests for spatial autocorrelation were done using GeoDa version 1.14.0. All other analyses were conducted using R statistical software version 3.5.2.

**RESULTS**

There was a total of XXX tests reported to OCHCA between March XX and June 30, 2020.

***Spatial patterns in reported COVID-19 case and testing intensity***

The highest reported case incidence in March occurred in communities along the central coast of Orange County (**Figure 1**). By May of 2020 the highest reported COVID-19 case intensities had shifted to North-Central part of the county. Testing intensity similarly shifted geographically over time. The highest reported testing intensities in March were in the Central Coastal and Southern part of the county. By May, however, overall testing intensity was much higher throughout the county.

Spatial clustering of reported cases was strongest in May (**Supplementary Figure 2**) whereas spatial clustering of testing intensity was strongest in March (**Supplementary Figure 3**). The results of the LISA statistics (**Figure 2**) indicate hot and cold-spots in both reported COVID-19 case intensity and testing intensity. In March reported COVID-19 case intensity was highest in the Central Coastal areas, with cold spots occurring in North- Central areas. These hot and cold spots largely overlap with the LISA results from testing intensity in March. A hotspot of COVID-19 cases began developing in April in the North-Central part of the County, which steadily progressed into a relatively larger cluster by May.

***Economic correlates of COVID-19 testing and test positivity.***

Two general patterns are apparent in testing and positivity by zip code from which patients were identified. The first general pattern is that individuals who received tests at the beginning of the COVID-19 epidemic in Orange County (regardless of testing outcome) resided in zip codes with higher overall median HHI. For example, in the second week of March, the median HHI in the zip codes of all patients was $97,459; by contrast, in the fourth week of May, it fell to $80,218.

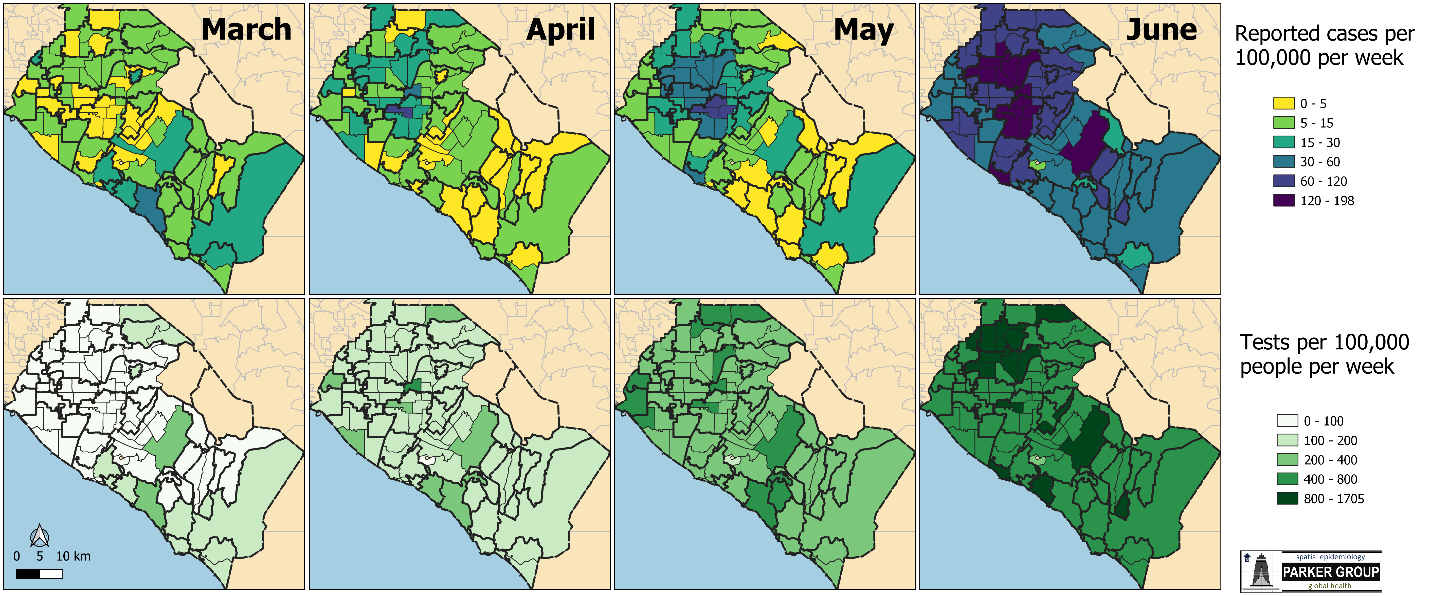
The second apparent pattern is the widening over time in the difference in median HHI between those who tested positive and those who tested negative (Figure 3). For example, individuals who tested positive in the fourth week of May lived in zip codes with a median HHI of $65,645; by contrast, individuals who tested negative in the fourth week of May lived in zip codes with a median HHI of $81,347.

***Logistic regression for odds of being diagnosed with COVID-19.***

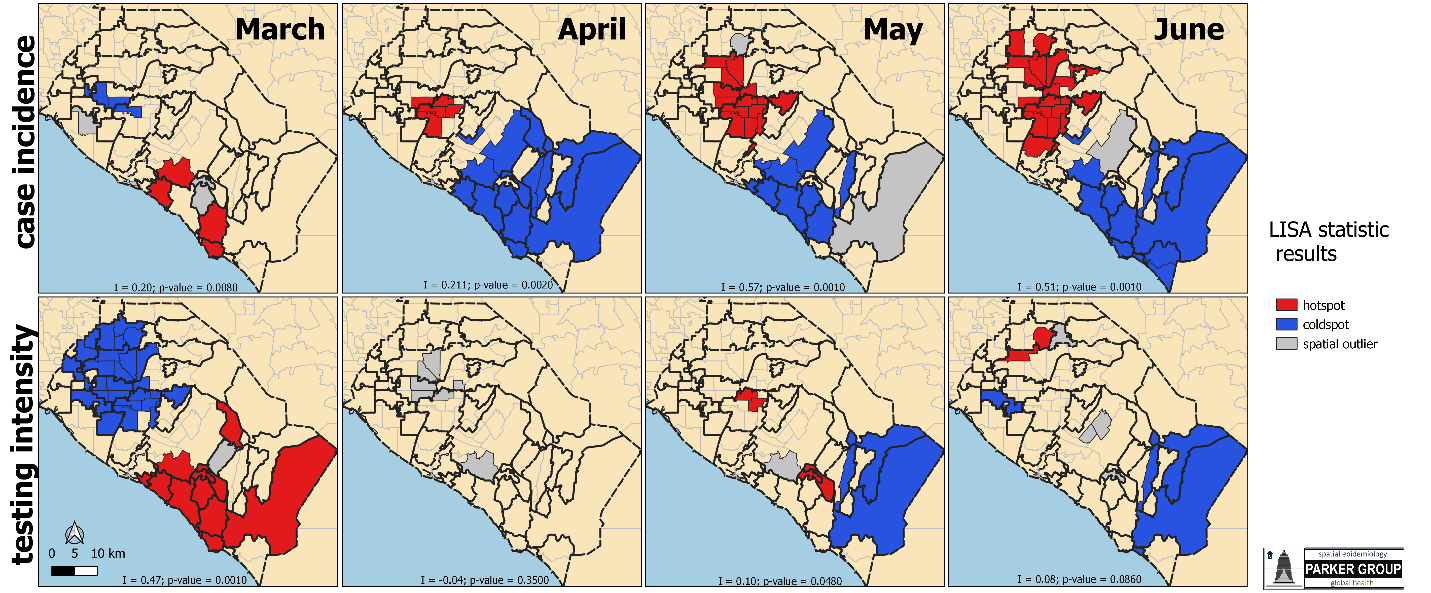
The regression results appear in **Table 1**. There is a general pattern of increasing odds of being diagnosed with COVID-19 by age, with the highest predicted odds of being diagnosed with COVID-19 occurring among 20- 24 year olds (AOR: 5.29; CI: 4.30 – 6.51). Males have higher odds of being diagnosed with COVID-19 (**Table 1**). For example, when looking at the entire data set (**M1** in **Table 1**), males have 36% higher odds of being diagnosed with COVID-19 than females. Individuals who self-identified as being Hispanic or Latino had higher odds of being diagnosed with COVID-19. In comparison to people who identified as White, they had 30% increased odds of being diagnosed with COVID-19 (AOR: 1.30, CI: 1.21 – 1.40).

Median HHI at the zip code level was a statistically significant predictor of being diagnosed with COVID-19. In the baseline model (**M1** of **Table 1**), the adjusted odds ratio suggests that for every $10k increase in zip code level median HHI there is a 11% decrease in the odds of being diagnosed with COVID-19. From **Figure 3** it was apparent that this effect changed over time, with higher zip-code level household income for positive cases in early March. By late March, the HHI for positive cases had dropped below that for negative cases. In **M2** (from **Table 1**) an interaction term between median HHI and month was included. Using March as the comparison month, an increase in median HHI in all subsequent months was associated with decreased odds of testing positive for COVID-19 (22% decrease in April, 27% decrease in May, and 17% decrease in June).

**Figure 1: Reported COVID-19 case incidence and COVID-19 testing intensity in Orange County zip codes in March, April, and May of 2020.** Supplementary Figure 1 indicates city locations (corresponding to dashed black borders).

****

**Figure 2: Results of tests for spatial clustering of reported COVID-19 case intensity (top row) and COVID-19 testing intensity (bottom row).** Classifications come from LISA (local indicators of spatial autocorrelation) statistics and only statistically significant outcomes are shown. Hotspots are indicated in red, coldspots in blue, and grey areas are statistically significant outliers (either high values surrounded by low values or low values surrounded by high values). Global Moran’s I statistics and pseudo p-values are included in each panel.



**Figure 3: Median household income (from zip code estimates) by test result over time.**

**Table 1: Logistic regression results predicting the odds of being diagnosed positive for COVID-19 among all tests.** M1 is the baseline regression model. M2 includes an interaction term for household income and month since this effect varies over time. Model results are presented as the model adjusted odds ratio (AOR) and 95% confidence intervals (95% CI).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **covariate** | **COVID-19 +** | **total** | **M1** | **M2** |
|  |  | **n** | **n** | **AOR (95% CI)** | **AOR (95% CI)** |
| age | *0to4* | 120 | 1224 | Referent | Referent |
| *5to9* | 111 | 700 | 1.76 (1.33 - 2.34) | 1.75 (1.31 - 2.32) |
| *10to14* | 191 | 858 | 2.95 (2.29 - 3.80) | 2.96 (2.30 - 3.83) |
| *15to19* | 435 | 1549 | 4.30 (3.44 - 5.38) | 4.30 (3.43 - 5.38) |
| *20to24* | 1046 | 3191 | 5.29 (4.30 - 6.51) | 5.25 (4.27 - 6.46) |
| *25to29* | 1199 | 4122 | 4.33 (3.53 - 5.32) | 4.35 (3.55 - 5.35) |
| *30to34* | 996 | 4420 | 3.33 (2.71 - 4.09) | 3.38 (2.75 - 4.16) |
| *35to39* | 854 | 3911 | 3.39 (2.75 - 4.17) | 3.41 (2.77 - 4.20) |
| *40to49* | 1630 | 6845 | 3.81 (3.12 - 4.66) | 3.82 (3.13 - 4.68) |
| *50Plus* | 3363 | 27099 | 1.85 (1.52 - 2.25) | 1.87 (1.54 - 2.28) |
| gender | *Female* | 4882 | 29394 | Referent | Referent |
| *Male* | 5063 | 24525 | 1.36 (1.30 - 1.43) | 1.36 (1.30 - 1.42) |
| race | *Non-Hispanic White* | 7219 | 38816 | Referent | Referent |
| *American Indian or Alaska Native* | 37 | 181 | 0.98 (0.68 - 1.43) | 1.00 (0.69 - 1.46) |
| *Hispanic or Latino* | 1476 | 4487 | 1.30 (1.21 - 1.40) | 1.30 (1.21 - 1.40) |
| *Native Hawaiian or Other Pac. Islander* | 59 | 1076 | 0.21 (0.16 - 0.27) | 0.21 (0.16 - 0.27) |
| *Asian* | 960 | 8093 | 0.54 (0.50 - 0.58) | 0.54 (0.50 - 0.59) |
| *Black or African American* | 194 | 1266 | 0.60 (0.51 - 0.71) | 0.61 (0.52 - 0.72) |
|  | *median house income (per $10k)* |  |  | 0.89 (0.86 - 0.92) | 1.09 (1.04 - 1.14) |
| Month | *March* | 639 | 2226 | Referent | Referent |
| *April* | 1669 | 8913 | 0.41 (0.37 - 0.46) | 3.74 (2.47 - 5.65) |
| *May* | 2795 | 17971 | 0.30 (0.27 - 0.34) | 4.77 (3.21 - 7.09) |
| *June* | 4842 | 24809 | 0.40 (0.36 - 0.44) | 2.35 (1.62 - 3.40) |
|  | *Population density* |  |  | 1.00 (0.97 - 1.03) | 1.00 (0.97 - 1.03) |
| interaction | *household income \* March* |  |  |  | Referent |
| *household income \* April* |  |  |  | 0.78 (0.75 - 0.82) |
| *household income \* May* |  |  |  | 0.73 (0.70 - 0.76) |
| *household income \* June* |  |  |  | 0.83 (0.80 - 0.86) |

**DISCUSSION**

Orange County is the 6th most populous county in the US and contains a mix of dense urban and suburban areas. We explored spatial and temporal patterns in both overall test availability and test positivity for COVID-19. At the initial stage, reported COVID-19 case incidence and overall test availability in Orange County was centered in wealthy and affluent areas along the Central Coast. Both hot-spots and cold-spots in case incidence in March overlapped geographically with testing intensity (**Figure 1** and **Figure 2**). Testing was strongly geographically clustered at the beginning of the epidemic. After the first month, however, a large cluster of reported cases formed in North-Central Orange County. This cluster began in April, grew in size in May, and persists to the present. Testing capacity has also expanded to less affluent areas. In addition, after the first month of the outbreak in Orange County, testing intensity is no longer a strong geographic correlate of case detection.

Over time, the gap in zip code level HHI between those diagnosed with and without COVID-19 has significantly widened. When the epidemic began in Orange County there was relatively little difference with regard to zip code level median HHI between those who were and those who were not diagnosed with COVID-19. In addition, most people receiving tests in March lived in zip codes with median HHIs greater than $90,000 per year (**Figure 3**). In May, individuals from lower income zip codes were increasingly diagnosed with COVID-19. As the total number of COVID-19 cases continues to grow, neighborhoods with relatively lower HHIs account for a disproportionate share of COVID-19 cases in the County.

Young adult males and people who identify as Hispanic or Latino were also more likely to be diagnosed with COVID-19. Social determinants of health, "conditions in which people are born, grow, work, live, age, and the wider set of forces and systems,”5 play a critical role in the creation of disparities related to morbidity, mortality, and quality of life6. These social determinants include poverty, wealth, educational quality, neighborhood conditions, childhood experience, social support, access, etc. Several speculative explanations have been proposed for these sociodemographic patterns.

As the state and local shelter in place and social distancing policies were mandated, it is also likely that individuals who are independently wealthy or who work in occupations where working from home is a viable option, were more capable of practicing social distancing. People from population dense, low socioeconomic status (SES) areas may have less ability to practice social distancing. Our finding that COVID-19 cases in Orange County increased linearly in zip codes associated with higher housing density, and lower educational attainment, underscore the importance of understanding contextual factors surrounding infectious disease outbreaks. Data from the H1N1 in 2009 suggest that individuals living in disadvantaged high density communities may be at greater risk of infectious diseases10. During H1N1,a significant association between severity and lower educational attainment (individual and community) was observed in a study examining factors associated with poorer outcomes 18. Other studies suggest that neighborhoods defined as disadvantaged were associated with higher influenza-related deaths, hospitalizations, and poorer vaccination coverage9-11,19. Similarly, for COVID-19 the interaction of social determinants at multiple levels may lead to a "perfect storm" that disproportionately affects underserved populations by negatively impacting both rates of infection and outcomes.

We hypothesize that underserved populations living in zip-codes associated with lower-income are experiencing a spike in COVID-19 transmission due to increased housing density, employment in "essential/frontline" job sectors, lower educational attainment, and limited health care access. For health care workers, this trend has been seen across the U.S. and has been attributed to their “frontline” status with exposure to COVID-19 patients20. Further, it’s possible that the lower income and minority populations demonstrate a similar increase over time may also be experiencing this “frontline” spike due to barriers to social distancing.

In Orange County, Hispanic/Latino residents may disproportionately work in sectors in which social distancing or “working from home” is not an option. These occupations may include work in the food industry, construction, and health care support. Second Hispanics/Latinos in Orange County, on average, live in denser household arrangements and thus may have greater opportunity for exposure due to greater volume of family contacts. Third, males appear to show slightly greater vulnerability to COVID-19 which may arise due to their slightly greater prevalence of co-morbid conditions and generally greater morbidity risk at each adult age group (relative to women). These explanations, however, require further refinement and testing.

The reasons for these spatial, socio-demographic, and economic patterns we discovered are likely complex. Travel among wealthy individuals between North American and European reservoirs of the disease represents one possibility. Some of these reservoirs, at the outset of wider spread of COVID-19 in March 2020, include Italy, New York City, and a ski festival in Colorado. Another possibility is that access to diagnosis and test kits was facilitated in part by wealth and greater health care access. As the social distancing mandate was enacted, moreover, individuals who are independently wealthy or who have the option of working from home may have been more capable of practicing social distancing than did less wealthy individuals. Adherence to social policies is likely difficult for populations that do not have and cannot afford remote work options. We suspect that additional data from contact tracing protocols, combined with subsequent analyses such as those we provide here, will assist in elucidating which of these explanations best fit the spatial and temporal patterns of COVID-19 cases and testing in Orange County, California.

***Limitations.***

The data used for this analysis were collected via passive case detection and are therefore an under-representation of the true incidence of COVID-19 in Orange County. The data are also aggregated at the zip code level. Zip codes are unlikely to adequately represent important spatial units that would be preferable to analyze (i.e. neighborhoods, communities, etc.). While collection of data by zip code allows for some socio-economic differentiation, zip code data lacks sensitivity to understand community level differences in social determinants. Further, missing data including race does not allow us to take a finer look at temporal parameters among specific vulnerable populations.

***Strengths.***

Despite the spatial resolution of the testing data, individual-level data were available for our temporal analysis. In addition, the diversity of Orange County in terms of socioeconomic and demographic predictors allowed us to examine important social determinants with sufficient power. California was also one of the first states to issue an executive order for residents to stay home, providing data for several months when only essential workers were permitted to work outside the home. Our analyses were able to identify temporal shifts in the demographics of COVID-19 test positivity that likely reflect disparities related to occupation type that are further amplified by household characteristics.

**REFERENCES**

1. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. PNAS. 2020 [cited 23 Apr 2020]. doi:10.1073/pnas.2004911117

2. Dorn A van, Cooney RE, Sabin ML. COVID-19 exacerbating inequalities in the US. The Lancet. 2020;395: 1243–1244. doi:10.1016/S0140-6736(20)30893-X

3. Yancy CW. COVID-19 and African Americans. JAMA. 2020 [cited 23 Apr 2020]. doi:10.1001/jama.2020.6548

4. Garg S. Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14 States, March 1–30, 2020. MMWR Morb Mortal Wkly Rep. 2020;69. doi:10.15585/mmwr.mm6915e3

5. Dondorp AM, Hayat M, Aryal D, Beane A, Schultz MJ. Respiratory Support in Novel Coronavirus Disease (COVID-19) Patients, with a Focus on Resource-Limited Settings. The American Journal of Tropical Medicine and Hygiene. 2020 [cited 23 Apr 2020]. doi:10.4269/ajtmh.20-0283

6. Bhatraju PK, Ghassemieh BJ, Nichols M, Kim R, Jerome KR, Nalla AK, et al. Covid-19 in Critically Ill Patients in the Seattle Region — Case Series. New England Journal of Medicine. 2020;0: null. doi:10.1056/NEJMoa2004500

7. Maxmen A. Thousands of coronavirus tests are going unused in US labs. Nature. 2020;580: 312–313. doi:10.1038/d41586-020-01068-3

8. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. The Lancet Respiratory Medicine. 2020;0. doi:10.1016/S2213-2600(20)30079-5

9. Grasselli G, Pesenti A, Cecconi M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. JAMA. 2020 [cited 23 Apr 2020]. doi:10.1001/jama.2020.4031

10. <https://www.gov.ca.gov/wp-content/uploads/2020/03/3.19.20-attested-EO-N-33-20-COVID-19-HEALTH-ORDER.pdf>

11. <https://covid19.ca.gov/essential-workforce/>

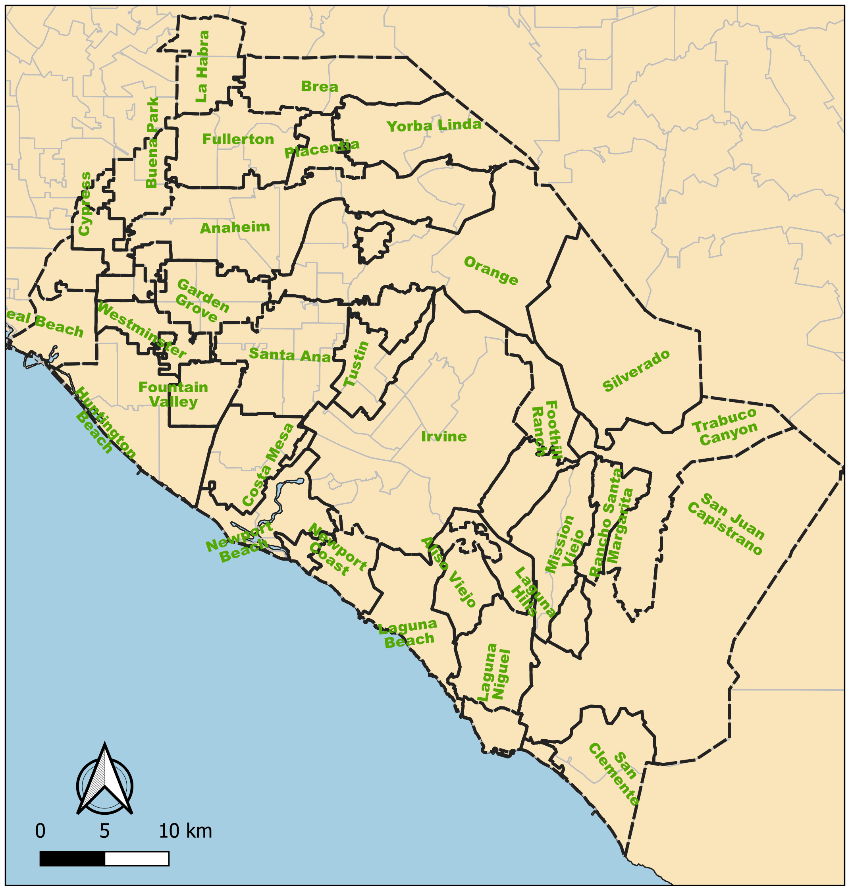
12. Orange County’s Healthier Together :: Indicators :: OC Dashboard. <http://www.ochealthiertogether.org/indicators/index/dashboard?alias=ocdashboard&localeId=267&page=2&card=1>

13. Vo, Linda Trinh; Hom LD. Transforming Orange County: Assets and Needs of Asian Americans & Native Hawaiians and Pacific Islanders. Orange County, CA; 2018.

14. <https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/CalREDIE.aspx>

15. <https://occovid19.ochealthinfo.com/oc-health-care-agency-covid-19-data-frequently-asked-questions>

**Supplementary Figure 1**: **Map of Orange County with cities labelled**. Analyses in this research are done at the zip code level (light grey lines indicate zip code boundaries while dashed dark black lines indicate city boundaries).



**Supplementary Figure 2**: Spatial correlogram indicating degree of clustering (along y-axis) in reported COVID-19 case incidence by distance (meters) away from neighboring zip codes in Orange County. Clustering in March is indicated by the dark red line, April by the light pink line, May by the bold red line, June by the blue line. The strongest clustering pattern (highest spatial autocorrelation) in reported cases occurred in May and June, corresponding to a large cluster in the North-Central part of the County (**Figure 2**).

**Supplementary Figure 3**: Spatial correlogram indicating degree of clustering (along y-axis) in COVID-19 testing intensity by distance (meters) away from neighboring zip codes in Orange County. Clustering in March is indicated by the dark blue color, April by the light grey color, May by the orange color, and June by the maroon color. The strongest clustering pattern (highest spatial autocorrelation) in testing occurred in March, at the beginning of the Orange County epidemic.

**Updated logistic regression results, now stratified by month**



**Logistic regression results for odds of death, among those who test positive**



R script

###########################################################################

# Load the full data

patdata\_OC<-read.csv("C:/Users/Daniel Parker/Dropbox/COVID\_19/SpatialEpiOC/AllELR\_PCR\_tests200709\_DropNA.csv",header=TRUE)

#drop repeated patients

patdata = subset(patdata\_OC, Series==1)

###########################################################################

# Pop data

PopData\_OC<-read.csv("C:/Users/Daniel Parker/Documents/COVID19/OC/OC\_Data/ZipPop.csv",header=TRUE)

# Area data

Area\_OC <-read.csv("C:/Users/Daniel Parker/Documents/COVID19/OC/OC\_Data/ZipArea.csv",header=TRUE)

PopArea\_OC = merge(PopData\_OC, Area\_OC, by="Zip", keep.x="TRUE")

PopArea\_OC$PopDens = PopArea\_OC$Population / PopArea\_OC$AreaKm

###########################################################################

# Test data

#testdata\_OC<-read.csv("C:/Users/Daniel Parker/Documents/COVID19/OC/OC\_Data/FromOCHCA/PatientLevel/NEG.test.results.csv",header=TRUE)

ZipPop = merge(patdata, PopArea\_OC, by.x="Zip", by.y="Zip", keep.x="TRUE")

#### income data coming from: http://www.ochealthiertogether.org/indicators/index/view?indicatorId=315&localeTypeId=3&periodId=1581

Income\_OC<-read.csv("C:/Users/Daniel Parker/Documents/COVID19/OC/OC\_Data/IncomeByZip2.csv",header=TRUE)

ZipIncome3 = merge(ZipPop, Income\_OC, by="Zip", keep.x="TRUE")

#### education data coming from: http://www.ochealthiertogether.org/indicators/index/view?indicatorId=315&localeTypeId=3&periodId=1581

Edu\_OC<-read.csv("C:/Users/Daniel Parker/Documents/COVID19/OC/OC\_Data/EducationByZip.csv",header=TRUE)

ZipIncome2 = merge(ZipIncome3, Edu\_OC, by="Zip", keep.x="TRUE")

#### insurance data coming from: http://www.ochealthiertogether.org/indicators/index/view?indicatorId=315&localeTypeId=3&periodId=1581

Insur\_OC<-read.csv("C:/Users/Daniel Parker/Documents/COVID19/OC/OC\_Data/InsuranceByZip.csv",header=TRUE)

ZipIncome1 = merge(ZipIncome2, Insur\_OC, by="Zip", keep.x="TRUE")

####create dataset without missing values

### for summary stats this isn't necessary but it is for the mixed effects regression

#vars to keep

myvars <- c("Zip", "COVIDbin", "Age", "Sex2",

"Race3", "IncomeMed", "PopDens", "Month",

"PercentBach", "PercentInsured")

####vars to keep for March

#myvars <- c("Zip", "COVIDbin", "Age", "Sex2",

#"Race4", "IncomeMed", "PopDens", "Month",

#"PercentBach", "PercentInsured")

#create dataset with only those vars

ZipIncome <- ZipIncome1[myvars]

#create age group variables for the ages

attach(ZipIncome)

ZipIncome$AgeGroup[Age < 5] <- "0to4"

ZipIncome$AgeGroup[Age > 4 & Age < 10 ] <- "5to9"

ZipIncome$AgeGroup[Age > 9 & Age < 15 ] <- "10to14"

ZipIncome$AgeGroup[Age > 14 & Age < 20 ] <- "15to19"

ZipIncome$AgeGroup[Age > 19 & Age < 25 ] <- "20to24"

ZipIncome$AgeGroup[Age > 24 & Age < 30 ] <- "25to29"

ZipIncome$AgeGroup[Age > 29 & Age < 35 ] <- "30to34"

ZipIncome$AgeGroup[Age > 34 & Age < 40 ] <- "35to39"

ZipIncome$AgeGroup[Age > 39 & Age < 50 ] <- "40to49"

ZipIncome$AgeGroup[Age > 49 & Age < 60] <- "50to59"

ZipIncome$AgeGroup[Age > 59 & Age < 70] <- "60to69"

ZipIncome$AgeGroup[Age > 69 & Age < 80] <- "70to79"

ZipIncome$AgeGroup[Age > 79] <- "80Plus"

ZipIncome$AgeGroup2[Age < 15] <- "0to14"

ZipIncome$AgeGroup2[Age > 14 & Age < 20 ] <- "15to19"

ZipIncome$AgeGroup2[Age > 19 & Age < 25 ] <- "20to24"

ZipIncome$AgeGroup2[Age > 24 & Age < 30 ] <- "25to29"

ZipIncome$AgeGroup2[Age > 29 & Age < 35 ] <- "30to34"

ZipIncome$AgeGroup2[Age > 34 & Age < 40 ] <- "35to39"

ZipIncome$AgeGroup2[Age > 39 & Age < 50 ] <- "40to49"

ZipIncome$AgeGroup2[Age > 49 & Age < 60] <- "50to59"

ZipIncome$AgeGroup2[Age > 59 & Age < 70] <- "60to69"

ZipIncome$AgeGroup2[Age > 69 & Age < 80] <- "70to79"

ZipIncome$AgeGroup2[Age > 79] <- "80Plus"

ZipIncome$Month2[Month==3] <- "March"

ZipIncome$Month2[Month==4] <- "April"

ZipIncome$Month2[Month==5] <- "May"

ZipIncome$Month2[Month==6] <- "June"

detach(ZipIncome)

ZipIncome$IncomStand = ZipIncome$IncomeMed/10000

ZipIncome$Zip = as.factor(ZipIncome$Zip)

ZipIncome$Pop = ZipIncome$Population/1000

###standardized zipcode level variables

ZipIncome$PDensStand = scale(ZipIncome$PopDens, center=TRUE, scale=TRUE)

ZipIncome$IncomeStand = scale(ZipIncome$IncomeMed, center=TRUE, scale=TRUE)

ZipIncome$EduStand = scale(ZipIncome$PercentBach, center=TRUE, scale=TRUE)

ZipIncome$InsuredStand = scale(ZipIncome$PercentInsured, center=TRUE, scale=TRUE)

#categorize zipcode level variables into quartiles

ZipIncome$IncomeQ <- with(ZipIncome, cut(IncomeMed,

breaks=quantile(IncomeMed, probs=seq(0,1, by=0.25), na.rm=TRUE),

include.lowest=TRUE))

ZipIncome$EduQ <- with(ZipIncome, cut(PercentBach,

breaks=quantile(PercentBach, probs=seq(0,1, by=0.25), na.rm=TRUE),

include.lowest=TRUE))

ZipIncome$InsuredQ <- with(ZipIncome, cut(PercentInsured,

breaks=quantile(PercentInsured, probs=seq(0,1, by=0.25), na.rm=TRUE),

include.lowest=TRUE))

ZipIncome$Race3 = factor(ZipIncome$Race3,

levels = c("White", "American Indian or Alaska Native",

"Hispanic or Latino",

"Native Hawaiian or Other Pacific Islander",

"Asian", "Black or African American", "Unknown"))

#This is only for when setting up the March data set

#ZipIncome$Race4 = factor(ZipIncome$Race4,

#levels = c("White",

#"Hispanic or Latino",

#"Native Hawaiian or Other Pacific Islander",

#"Asian", "Black or African American", "Unknown"))

##providing an order for the months

ZipIncome$Month2 = factor(ZipIncome$Month2,

levels = c("March", "April", "May", "June"))

#creating separate dataset for each month

##later I will run models for each month

March = subset(ZipIncome, Month2 == "March")

April = subset(ZipIncome, Month2 == "April")

May = subset(ZipIncome, Month2 == "May")

June = subset(ZipIncome, Month2 == "June")

#looking at some summary stats

table(ZipIncome$AgeGroup)

aggregate(IncomeMed ~ WeekOfYear, data=ZipIncome, FUN=summary)

aggregate(IncomeMed ~ COVIDbin + WeekOfYear, data=ZipIncome, FUN=summary)

aggregate(IncomeMed ~ WeekOfYearm + COVIDbin, data=ZipIncome, FUN=summary)

library(lme4)

library(beepr)

library(tictoc)

##models for each month, using quartiles for zipcode level vars

tic()

#march needs a different dataset because of different age and race/ethnicity bins

MarchM = glmer(COVIDbin ~ AgeGroup2 + Sex2 + Race4 + IncomeQ + EduQ + InsuredQ + PDensStand + (1|Zip),

data=March, family=binomial,

control = glmerControl(optimizer ="bobyqa", optCtrl=list(maxfun=100000)))

AprilM = glmer(COVIDbin ~ AgeGroup + Sex2 + Race3 + IncomeQ + EduQ + InsuredQ + PDensStand + (1|Zip),

data=April, family=binomial,

control = glmerControl(optimizer ="bobyqa", optCtrl=list(maxfun=100000)))

MayM = glmer(COVIDbin ~ AgeGroup + Sex2 + Race3 + IncomeQ + EduQ + InsuredQ + PDensStand + (1|Zip),

data=May, family=binomial,

control = glmerControl(optimizer ="bobyqa", optCtrl=list(maxfun=100000)))

JuneM = glmer(COVIDbin ~ AgeGroup + Sex2 + Race3 + IncomeQ + EduQ + InsuredQ + PDensStand + (1|Zip),

data=June, family=binomial,

control = glmerControl(optimizer ="bobyqa", optCtrl=list(maxfun=100000)))

toc()

beep("mario")