**Association between socio-demographic factors and COVID-19 incidence and mortality in California**

**Background**

COVID-19 has taken the lives of over 70,000 Californians according to the New York Times. The pandemic has disproportionately affected various populations within California (CA). From February through July of 2020 the elderly, Hispanic, male, immigrant, and poorly educated were the most targeted demographics by the disease[[1](#_ENREF_1)]. A better understanding of how COVID-19 disproportionately affects different parts of the population could better inform efforts to curb its spread and mortality rates by identifying areas where the most improvement could be made. The past has shown that quick and targeted action is the key to stopping the spread of disease[[2](#_ENREF_2)].

Here, we will focus on the association between socio-demographic characteristics and COVID-19 incidence and related mortality. Although several studies have shown this association[[1](#_ENREF_1), [3](#_ENREF_3)], this study aims to investigate factors associated with county disparities in CA. In this regard, two studies at a county level in the US from 2020 found that factors such as socioeconomic status and percentage of essential workers correlate with higher COVID-19 mortalities[[4](#_ENREF_4), [5](#_ENREF_5)]. In this study, we will use updated data from 2021 and include the number of vaccinated people in our analysis. Our study could help better identify the most effective communities to target for both now and future pandemics.

**Problem Statement**

Which socio-demographic factors are associated with COVID-19 incidence and mortality rates at the county-level in CA?

**Data**

To address this question, we will use different datasets from open access databases with aggregated data at a county level. We will retrieve COVID-19 related data from the California Health and Human Services Open Data Portal. This website reports a “COVID-19 Time-Series Metrics by County and State” which contains documented COVID-19 case numbers, death numbers, test numbers, positive tests, and population density from the 58 counties in California. From this same website we will use the database “COVID-19 Vaccine Progress Dashboard Data” which summarizes vaccination data at the county level. Most recent available datasets for demographic characteristics at county level that we will use for this study are described in Table 1.

**Table 1: County-level Data Sets**

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Description** | **Source** |
| Statewide COVID-19 Cases, Deaths, and, Tests | Contains data such as COVID-19 case numbers, deaths numbers, tests numbers, positive test, and population density, per date. | <https://data.chhs.ca.gov/dataset/f333528b-4d38-4814-bebb-12db1f10f535/resource/046cdd2b-31e5-4d34-9ed3-b48cdbc4be7a/download/covid19cases_test.csv> (up to date) |
| County-Level Poverty Estimates | Poverty estimates from the U.S. Census Bureau. | <https://data.ers.usda.gov/reports.aspx?ID=17826>(2019) |
| County-Level Unemployment, and Median Household Income Estimates | Unemployment rates from the Bureau of Labor Statistics, and median household income estimates from U.S. Census Bureau. | <https://data.ers.usda.gov/reports.aspx?ID=17828>(2019) |
| Rural-Urban Continuum Codes | Classification scheme that distinguishes metropolitan counties by the population size of their metro area | <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/> (2013) |
| Highest level of educational attainment | Education (2015-19, 5-yr average latest) | <https://data.ers.usda.gov/reports.aspx?ID=17829>(2019) |
| Statewide COVID-19 Vaccines Administered By County | Partial and total number of vaccinated people, and vaccine type, per date. | <https://data.chhs.ca.gov/dataset/e283ee5a-cf18-4f20-a92c-ee94a2866ccd/resource/130d7ba2-b6eb-438d-a412-741bde207e1c/download/covid19vaccinesbycounty.csv> (up to date) |

**Methods**

Study design

This will be a cross-sectional study to compare COVID-19 mortality rates across different counties in California, in a defined timeframe. The final goal is to investigate demographic factors associated with county variations in COVID-19 mortality.

Data processing

We will merge the datasets from different sources to generate a master dataset containing the outcome of interest, mortality and cases of COVID-19, and demographics characteristics for the 58 counties in California. The data will be organized into a data frame by way of google sheets.

Statistical analysis

We will use a multi-predictor linear regression model to test the association of different variables with the outcome of COVID-19 mortality rate per county. COVID-19 Mortality rates will be defined as the number of deaths per total county population per 100,000. Other possible outcomes of interest are case fatality rate (number of deaths per 100 COVID-19 cases) and incidence (number of cases per total population per 100,000). Variables of interest to fit in the model will include median household income, number of vaccinated people, number of tests, and percentage of poverty, unemployment, and low education level. Population density and rural estimates per county will be additional covariates as these vary between counties and are potential confounders. One possibility is stratifying the analysis in rural and non-rural counties. Number of vaccinated people and tests will be normalized by the county population. All analysis will be performed using the programming language R. Overall the results will give a quantifiable observation as to which communities suffer strongly in the midst of the Covid-19 pandemic.

**References**

1. Garcia, E., et al., *COVID-19 mortality in California based on death certificates: disproportionate impacts across racial/ethnic groups and nativity.* Ann Epidemiol, 2021. **58**: p. 69-75.

2. Kaur, S. and H. Bherwani, *Understanding COVID-19 transmission, health impacts and mitigation: timely social distancing is the key.* 2020: p. 1-17.

3. Vaughan, L., et al., *Relationship of socio-demographics, comorbidities, symptoms and healthcare access with early COVID-19 presentation and disease severity.* BMC Infectious Diseases, 2021. **21**(1): p. 40.

4. Fielding-Miller, R.K., M.E. Sundaram, and K. Brouwer, *Social determinants of COVID-19 mortality at the county level.* PLoS One, 2020. **15**(10): p. e0240151.

5. Karmakar, M., P.M. Lantz, and R. Tipirneni, *Association of Social and Demographic Factors With COVID-19 Incidence and Death Rates in the US.* JAMA Network Open, 2021. **4**(1): p. e2036462-e2036462.

Grade Breakdown:

Background – 28/30

Reasoning: Need citation. No justification for why COVID-19 incidence and mortality are the outcomes.

Problem Statement – 10/10

Reasoning: Problem statement identifies a reasonable problem in 1 sentence.

Data – 25/25

Reasoning: Data sources are appropriate for the task with relevant features for the problem.

Methods – 30/35

Reasoning: Analysis timeframe unclear. Needs to modify language to make it clear they are also analyzing COVID-19 incidence. Needs to perform univariate statistical tests to determine feature importance in addition to linear regression coefficient values.

Total: 93/100