Racial disparities in Covid-19 incidence associated with Socio-Economic Factors in the United States

MSc Research Project
Data Analytics

Ken Wheatley
Student ID: x16103785

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ken Wheatley
Student ID:	x16103785
Programme:	Data Analytics **Year:** 2023
Module:	MSc Research Project
Supervisor:	Vikas Sahni
Submission Due Date:	14th August 2023
Project Title:	Racial disparities in Covid-19 incidence associated with Socio-Economic Factors in the United States
Word Count:	8341 **Page Count** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ken Wheatley……

Date: 11th August, 2023……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Racial disparities in Covid-19 incidence associated with Socio-Economic Factors in the United States

Ken Wheatley
x16103785

Abstract

The global Covid-19 pandemic has revealed several major risks to various population groups. One of these is racial health outcome disparity. In the US, this has long been a policy concern but has been exacerbated by the Covid-19 outbreak. A main driver in the disparity in health outcomes are socio-demographic factors. The United States maintains comprehensive sources of publicly available data on social determinants of health (SDH) which include socio-economic and environmental factors.

In this study, it is hypothesized that racial disparities in Covid-19 incidence are observed and various SDH factors are examined to explain these observations Both fixed effects and mixed effects models were considered to take in account possible between-state variance in infection rates. Results showed that between state difference were small and that a fixed effects logistic regression model would be better choice because of its relative simplicity. SDH factors relating mostly to ethnic minority status were identified. Relative to Whites, Black and Asian populations have the least likelihood of Covid-19 infections.

1 Introduction

The outbreak of Covid-19 was declared a pandemic on 11th of March 2020 by World Health Organisation (WHO) and has emerged as a global health crisis devastating lives and economies. In the United States alone, as of June 1st, 2023, around 103 million confirmed cases (WHO Covid-19 Dashboard)[1] have been reported and over 1,127,000 deaths. On the 5th of March 2023, the head of WHO declared an end to Covid-19 as a public health emergency, but warned the virus is still a global threat and that the risk remains of new variants emerging that cause new surges in cases and deaths. With limited resources available, a priority of health departments throughout the US is to protect and assist the most vulnerable in their communities to avert severe Covid-19 impacts. Vulnerable groups include racial minorities whom through various community health, socio-economic and other environmental determinants are at disproportionately high-risk of adverse outcomes.

Socio-economic and environmental health factors have long been recognised by researchers as the most important drivers of health inequities within socially vulnerable groups (Braveman & Gottlieb, 2014; Shortreed et al., 2021; Zoungrana et al., 2022). Although the associations are widely

[1] https://covid19.who.int/region/amro/country/us

acknowledged, consensus is still lacking on precisely how each factor bears influence. In his paper (Kelly, 2021) argues that "..the mechanisms linking the social factors and disease outcomes are not well understood,,". As far back as 1986, a landmark report was issued by the US Health Secretary Margaret M. Heckler (Heckler, 1986). The report called attention to the longstanding and persistent burden of death, disease and disability experienced by those of black, Hispanic, Native American and Asian/Pacific Islander heritage in the United States". The Covid-19 pandemic has highlighted and exacerbated these disparities (World Health Organisation, 2021), (Lopez et al., 2021) and has spawned renewed research discussed in the next section.

A comprehensive review of published research on racial/ethnic disparities due to socio-economic status was carried out by (Khanijahani et al., 2021). They systematically reviewed 52 papers published between December 2019 and March 2021. Most of the studies, which included 37 from the United States, showed that racial/ethnic minority populations had higher risks of Covid-19 infection, confirmed cases, hospitalisation, and deaths. Although they acknowledged limitations in their review such as incongruity between definitions of race/ethnicity among studies, they were able to conclude that racial/ethnic disparity was evident due to several socio-economic factors. These included living in overcrowded house-holds, low household income, poverty, low education, and inability to speak a language other than their own native tongue. The review also identified gaps such as the potential impact from lack of insurance, which needed further study.

Another review by Mackey et al (Mackey et al., 2021) found that lack of health care access and increased exposure risk may be driving higher infection and mortality rates amongst Black and Hispanic communities. The US public health agency, Centers for Disease Control and Prevention (CDC) provides summary data showing the high relative risk of racial or ethnic minorities compared to white (non-Hispanic) persons of Covid-19 infection, death, and hospitalisation (National Center for Immunization and Respiratory Diseases (NCIRD), 2021).

The CDC cites a number (CDC SVI Documentation, 2020)[2] of socio-economic factors amongst racial and ethnic groups that give rise to exposure risks which include overcrowded housing conditions and occupations requiring close public contact. The University of Wisconsin Population Health Institute (UW Population Health Institute, n.d.)[3] maintains a county health rankings model covering US states. The model includes a broad range of population health factors and associated publicly available datasets.

This study focuses on the socio-economic and community health determinants of racial Covid-19 health disparities. The socio-economic measures used are based on the publicly available Social Vulnerability Index (SVI) data. The SVI is maintained by the Geospatial Research, Analysis, and Services Program (GRASP) at the CDC and Agency for Toxic Substances and Disease Registry (ASTDR). The SVI has a hierarchical structure which comprises of 4 main themes comprising several socio-economic factors. Population health data is also available from the UW Population Health Institute. Community Resilience and Gini Index data was obtained from the US Census Bureau website. One of the challenges of the study was obtaining suitable data broken down by racial or ethnic group. Some US states do not report this data as a policy decision. There appears to be a lack of uniformity, completeness, and transparency in racial and ethnic Covid-19 data collected across the US. This was noted by the "The Covid Racial Data Tracker" website[4], a collaborative project between Boston University and the Center for Antiracist Research which ceased on 7th March 2021. In February 2021, the CDC began collecting deidentified patient case data

[2] https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/pdf/SVI2020Documentation_08.05.22.pdf

[3] https://www.countyhealthrankings.org/explore-health-rankings

[4] https://covidtracking.com/race

broken down by geography, race/ethnicity and various personal attributes such as underlying medical conditions, hospitalisation and mortality indicators. The dataset is regularly updated and contains a large number of case records, but significant gaps remain in some attributes.

This study focuses on a sample of 16 US States, intended to be representative of the entire United States. The tools used and inferences made are intended to be applicable to any area of the US where SVI scores and UW health data are published.

1.1 Research Question

Three related questions regarding the Covid-19 outbreak in the US within the study period are the focus of this study. They are based on the supposition that racial disparities in coronavirus infection rates exist. And furthermore, that these disparities can be explained in terms of socio-economic and community health factors. This contrasts with the null hypothesis which may be stated as infection rates are independent of socio-economic and community health factors. And as such, infection rates among racial and ethnic minorities, who are disproportionately represented in areas of low socio-economic status would be expected to have statistically similar rates of infection to the most privileged communities. The goal of this study is to uncover evidence of the alternate hypothesis against the null, and identify which factors are most important in driving the disparities.
Research questions:

1. Which racial groups are most disadvantaged in terms of Covid-19 case rates?
2. What are the associations between racial group case rate disparities and socio-economic inequity and social health determinants?
3. What is the extent of between-state variation in racial group Covid-19 prevalence can between-state differences be explained State level socio-economic and health environment factors??

Objectives and Contribution

1. Provide quantitative supportive evidence to demonstrate where racial disparities exist.
2. Highlight those SVI and SDH factors which most impact observed racial disparities.
3. By using a standard measure of SVI and SDH definitions, make analysis scalable and readily applied to other US States.
4. The model should be easy to interpret so that so that associations are transparent. The model output could then be used by planners to better target vulnerable communities needing resource.

The contribution this study aims to achieve to build an explanatory or effects model based on actual observed case rates by race together with environmental indicators. In addition, the model should be easy to interpret so that so that associations are transparent. The model output could then be used by planners to better target vulnerable communities needing resource. The advantage of using SVI and UW health factor data is that is it publicly available and widely used.
The rest of the study is structured as follows: Section 2 discusses related work on racial health disparities and socio-economic deprivation; Section 3 and 4 discusses the study methodology. In section 5 implementation is described and results are presented. Sections 6 and 7 cover evaluation and discussion of results. Finally, in Section 8 the conclusion and future work section is presented.

2 Related Work

The Covid-19 pandemic has highlighted racial and ethnic disparities in health outcomes. This has spawned considerable new research using an array of traditional and modern machine learning techniques. Of the more traditional methods used are sophisticated epidemiological SIR based models. SIR is Susceptible-Infected-Recovered (or Removed). For time-series forecasting, state-of-the-art methods such as the LSTM (long-short term memory) recurrent neural network have been used.
In this study, the requirement is for a multivariate model with transparency on how predictive features explain the outcome variable. Regression models, including hierarchical mixed-effects regression models are ideally suited for this purpose and are the focus of this literature review.

2.1 Regression based Studies.

In their study based on Missouri residents (Karen E. Joynt Maddox et al, 2022), reported case and mortality rates significantly higher among non-Hispanic (NH) Black and NH Other/Unknown races than among White NH cases, accounting for various SDH factors. From raw patient-level data, overall race/ethnicity group characteristics were determined, and cumulative case and mortality rates were analysed. They used logistic and hierarchical models to model SDH associations to explain disparities. Patient-level analysis was also undertaken. The study used private hospital data not made available, thereby creating a barrier to reproduce study results independently.
In their study, Nayak et al (Nayak et al., 2020) investigated temporal Trends in the Association of Social Vulnerability and Race with County-Level Covid-19 Incidence and Outcomes in the United States. The CDC SVI metrics were used to measure social vulnerability. It is important to note however that this study is based on inferences from total infection and death rates and the proportion of racial population in each county. It is not based on actual infection and death rate data within each racial group. 3091 Counties where more than 50 Covid-19 cases by March 6th, 2021, were observed. Negative binomial mixed effects models were used with offsets using total population in each county. Racial and ethnic groups studied were Black, White, and Hispanic. The study found that although higher SVI was indicative of greater social vulnerability, that the influence of SVI changed over time. During periods where SVI was a significant influence, incident, and death rates within each county was found to be worse among the racial group with proportionally the highest population. Where Hispanics made up the largest population group, outcomes were the worst.
Karaye et al. (Karaye & Horney, 2020) used multiple linear regression to model Covid-19 case counts per 100,000 using the publicly available county-level SVI scores. For each county, the case counts per 100k were obtained by dividing the cumulative counts (between January 21, 2020, and May 12, 2020) of confirmed Covid-19 cases by the county total population multiplied by 100,000. Testing data was also included. However, their study did not adjust for race and ethnicity. The broad study examined 48 US states. The authors fitted local models for each county recognising the impact for any given SVI category may change from county to county. The study found that overall, SVI and minority status and language were predictive of increased Covid-19 case counts.
A hierarchical linear mixed-effect model was used by (Hawkins et al., 2020)to investigate Covid-19 related cases across the United States. Their study included 1,089,999 cases and 62,298 deaths in 3127 counties and used a metric called Distressed Communities Index (DCI) to measure socio-economic status. They found that racial disparities were evident and

identified lower education levels and the percent of black populations strongly associated with infection cases. This study made inferences based on race population data, rather than race stratified raw data on cases.

In their study, Oates et al (Oates et al., 2021) examined associations between neighbour-hood social vulnerability and Covid-19 Incidence in Alabama and Louisiana. They used the SVI as their measures of vulnerability. Covid-19 testing data was also considered. Although the authors acknowledged stark racial disparities in Covid-19 cases across the US, it did not adjust for this in their models. The study used negative binomial regressions for their analysis. Their results show a positive and significant association between all measures of social vulnerability and Covid-19 incidence in both states.
The paper by (Rozenfeld et al., 2020) studied the socio-economic, clinical, and epidemiological risk factors associated with Covid-19 infection. They used privately held patient level data together with publicly available data from the American Community Survey (2018 data). A multivariate logistic regression model was used to predict the risk of initial infection in the community. A bivariate analysis was used to assess significant features in the outcome. The study concluded thatCovid-19 infection is higher among groups already affected by health disparities across age, race, ethnicity, language, income, and living conditions. Odds ratios with 95% CI were presented.
Abedi et al (Abedi et al., 2021) investigated racial and ethnic disparities in Covid-19 infection in the United States. An aim of their study was to examine associations between infections and mortality. A total of 7 states and 369 counties were examined. Although the study made use of race and ethnicity data when available, proportions of population was mainly used to make inferences regarding race and ethnicity disparities. They used bivariate linear regression and correlation analysis to examine associations. The study concluded that racial, economic, and health disparities were present in the population studied. They also found that risk factors for infection and mortality are different.
Tiana N. Rogers et al (Rogers et al., 2020) investigated racial disparities in Covid-19 mortality among America's essential workers. This category of social vulnerability is important as public-facing occupations place workers at higher risk of infection. In their analyses, race and ethnicity groups were defined as Non-Hispanic (NH) White, NH Black, Hispanic, NH Asian (including Native Hawaiians and Pacific Islanders), and NH Other (including American Indians/Alaska Natives and multiracial individuals). The central hypothesis was that Covid-19 mortality was higher among NH Blacks compared with NH Whites because NH Blacks hold more essential-worker positions. The study found Covid-19 mortality was higher among NH Blacks compared with NH Whites, due to more NH Blacks holding essential worker positions. Vulnerability to coronavirus exposure was increased among NH Blacks, who dis-proportionately occupied the top nine essential occupations. As Covid-19 death rates continue to rise, existing structural inequalities continue to shape racial disparities in this pandemic. Policies mandating the disaggregation of state-level data by race/ethnicity are vital to ensure equitable and evidence-based response and recovery efforts. The analysis was descriptive using Spearman rank-order correlations.
A CDC MMWR paper (Morbidity and Mortality Weekly Report (MMWR)) (Barry et al., 2021), reported on a study of disparities in vaccination coverage throughout the United States by social vulnerability, defined as social and structural factors associated with adverse health outcomes. The period covered was 14 December 2020 to 1 May 2021. The authors note that as vaccine eligibility and availability continue to expand, ensuring equitable coverage for disproportionately affected communities remains a priority. The CDC SVI index was used to measure social vulnerability. Also measured was urbanicity defined by the CDC's own NCHS Urban-Rural Classification Scheme for Counties. Trends in vaccination coverage were

evaluated by epidemiologic week for SVI quartile, stratified by urbanicity. Generalized estimating equation models using binomial regressions were used to estimate vaccination coverage by SVI metrics, overall and by urbanicity. The study found that disparities in county-level vaccination coverage by social vulnerability have increased as vaccine eligibility has expanded, especially in large fringe metropolitan areas surrounding large cities (e.g., suburban), and non-metropolitan counties. By May 1, 2021, vaccination coverage among adults was lower among those living in counties with lower socioeconomic status and with higher percentages of households with children, single parents, and persons with disabilities.

Another study incorporating the CDC SVI was conducted (Biggs et al., 2021). They conducted an ecological (correlational) study that looked at the relationships between Covid-19 incidence and social vulnerability at the census tract level in the State of Louisiana. Using Choropleth maps, census tracts with high social vulnerability (SVI scores) and high Covid-19 incidence were identified. Negative binomial regression with random intercepts was used to compare the relationship between incidence and population exposure as measured by all 15 SVI variables. The study found areas of higher social vulnerability were associated with higher Covid-19 incidence. Although this study did not focus on racial or ethnic disparities specifically, beyond the minorities community's component of SVI, it did find that the SVI was a useful tool.

2.2 Identified Gaps

Although much research has been done on both racial disparities and on social vulnerabilities related to Covid-19, studies incorporating both with data stratified by race appear to be relatively few.

3 Methodology

The aim of this study is to build an effects model to infer key SDH factors associated with disparities in racial group case rates. Reducing case rates in turn helps to reduce consequent adverse outcomes such as hospitalisations. Another objective is that the model should be parsimonious and easy to use and interpret.

Based on the literature review and exploration of the data, a logistic regression model is proposed. Count-based regression methods are common in epidemiological research of which logistic regression is one, Cross-sectional data is obtained by aggregating monthly Covid-19 incidence time-series data taken from January 2020 into a cumulative snapshot as of June 2023, Data is grouped by state, county, and race and then divided by respective county race populations to obtain racial group case rate proportions for each county. These proportions then lend themselves to logistic regression analysis.

This study broadly follows the Cross Industry Standard Process for Data Mining, commonly referred to as CRISP-DM. This widely used open standard methodology provides a structured framework to data science projects. The methodology comprises 6 hierarchical high-levels process phases (Chapman, 1999). in the following order: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The methodology is iterative allowing continuous model enhancements.

How each phase has been adapted for the specific needs of this study are described below.

3.1 Business Understanding

In this phase the purpose of the study is defined. Study goals are formulated with specific objectives on how to meet them. This is an iterative process which requires an understanding

of availability and constraints on data resources needed to support objectives and then putting a plan in place. The research goals of determining an association between racial disparities in Covid-19 infection rates and social vulnerability were shaped by a broad investigation. The literature review formed a major part of this.
In addition to reading formal academic work, various sources of grey literature were used to keep timely track of the status of the pandemic in general and the impacts on racial and ethnic minorities. Sources included the John Hopkins (University of Medicine) Coronavirus Resource Center (Johns Hopkins University of Medicine, 2022)[5], the World Health Organisation (WHO) Covid-19 Dashboard, (WHO Covid-19 Dashboard, 2020) and the CDC Covid Data Tracker (CDC COVID Data Tracker, 2022).

3.2 Data Understanding

To support and refine study goals, required data is identified, gathered, and analysed. Familiarity with the data is gained and insights discovered.
Incidence data: Monthly time-series data of new cases was downloaded from the CDC "Covid-19 Case Surveillance Public Use Data with Geography" website[6].
Incidence data records all laboratory-confirmed or probable cases. Both are used in the analysis and treated as cases. Data covers all US mainland states and counties. Incidence data also contains gender and age-group indicators for each patient.
Race. CDC defines 6 races categories:

- American Indian/Alaska Native
- Asian
- Black
- Multiple/Other
- Native Hawaiian/Other Pacific Islander
- White

Geographic Data. 16 US states were selected for the study. Selection was based on 2 main criteria:

- Geographic spread to best obtain a representative sample. At least one state in each of the 4 US Census Regions (Table 1.)
- States with the largest populations for every race were selected to maximise statistical inference capabilities.

ANSI Code	State	US Census Region		ANSI Code	State	US Census Region
AZ	Arizona	West		NC	North Carolina	South
CA	California	West		OH	Ohio	Midwest
CO	Colorado	West		OR	Oregon	West
FL	Florida	South		PA	Pennsylvania	Northeast
IL	Illinois	Midwest		TX	Texas	South
NV	Nevada	West		UT	Utah	West
NJ	New Jersey	Northeast		VA	Virginia	South
NY	New York	Northeast		WA	Washington	West

Table 1. US States sampled in study.
Population. Population estimates broken down by race were obtained from the CDC website[7] for "Wide-ranging Online Data for Epidemiologic Research" (WONDER). This data is used to transform case counts to rates per population and calculate population densities.

[5] https://coronavirus.jhu.edu/
[6] https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4
[7] https://wonder.cdc.gov/single-race-v2021.html

Socio-economic and Environmental Health Factors
These factors form part of a broader class of non-medical factors that influence health outcomes, known collectively as social determinants of health (SDH). In the US, publicly available SDH data is provided for many geographical areas including both county and state levels. Candidate SDH factors used in this study are described under their geographic level.

- **County Level**
  a. Social Vulnerability
  Social vulnerability refers to a community's ability to prevent human suffering and financial loss in a disaster, such as the coronavirus outbreak. Social vulnerability is driven by many socioeconomic and health factors. The CDC maintains a social vulnerability index (SVI) for which data and documentation is downloadable from its website[8]. The data is sourced from US Census tracts.
  b. Community Health.
  The University of Wisconsin Population Health Institute publishes a "County Health" rankings model for every US state and all counties[9]. Counties within each state are ranked according to SDH factors.
    - Length of Life
    - Quality of Life
    - Health Behaviours
    - Clinical Care
    - Social & Economic Factors
    - Physical Environment
- **State Level**
  c. Community Resilience (CRE).
  This is the capacity of individuals and households within a community to absorb the external stresses of a disaster. Data is provided by the US Census Bureau[10]. The CRE model considers 10 risk factors based on socio-economic factors. The risk factor either pertains to the individual directly (e.g., No health insurance coverage) or indirectly vis the household in which they reside (e.g., Households without a vehicle). For each state, estimates of the number of individuals per population are assessed as having zero, 1 or 2, and 3 or more risk factors are provided. Although, there is some minor overlap with SVI, SVI data is not available at the state level.

Gini Index Data. The Gini Index is a commonly used measure of income or welfare inequality within a group of people. State-level data was obtained from the US Census Bureau[11].

3.3 Data Cleaning and Preparation

This phase is used to finally prepare the data for modelling input. Required features are selected or derived, data is cleaned, transformed, and integrated. Covid-19 case data was downloaded from the CDC website as a .CSV file. The file contained monthly time-series of Covid-19 incidence broken down by state, county, and race. Both confirmed cases and

[8] https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html
[9] https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation
[10] https://www.census.gov/programs-surveys/community-resilience-estimates/data/datasets.html
[11] https://data.census.gov/table?q=Gini&g=010XX00US,$0400000_040XX00US01,02,04,05,06,08,09,10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,44,45,46,47,48,49,50,51,53,54,55,56,72&tid=ACSDT1Y2021.B19083&moe=false&tp=true

probable cases were recorded, and these were all included in this study. The data extracted was for the period from January 2020 up until June 2023. The data comprised around 63 million rows where each row of data within a month represented a deidentified individual infection case. Individuals that share the same state, county and race were indistinguishable from one another.
Time series entries where race data was missing (sometimes suppressed for data protection reasons), were necessarily excluded from the study as no corresponding population data could be applied. Around 3 million rows were dropped. The remaining 60 million rows were aggregated by state, county, and month to produce a cumulative total of cases (latest snapshot or cross-section) for each racial group. The aggregated table contained 2578 rows. The aggregated data was then merged with race population data to derive racial group case rate proportions within each county. A small number of UW County Health rankings data was missing for some categories. These were imputed by using median category values in the state for the missing data. The CDC County SVI data was merged to create a master table of county level data. Finally, the candidate model features were extracted from the master table and standardised ready for model input. State level features are added when required at a later stage in the modelling process.

3.4 Modelling

3.4.1 Chi-Squared Tests

To provide preliminary quantitative evidence of racial group disparities in covid-19 infection, a goodness-of-fit Chi-squared test was designed using a two-way crosstabulation of racial group proportions by state. This test was used to test the null hypothesis that independent of racial group, covid-19 infection rates followed a distribution based solely on population ratios.

3.4.2 Modelling Techniques

Logistic regression modelling has been selected to address the research objectives. The following provides context and rationale for the modelling techniques to be used.
The model data reveals a nested hierarchical structure where county-level racial groups are clustered within US states. The are 6 races, 16 states and their constituent counties modelled. A hypothesis of this study is that infection rates among racial groups residing in the same state (cluster) are likely more correlated to each other (because they share the same geographic area), than to those in neighbouring states.
One way to model the clustered data is by hierarchical mixed effects modelling which treats between-cluster variance as a flexible useful feature which allows cluster–level features to be added to explain the variance. Another advantage of mixed effects modelling is that the effects of unobserved cluster-level variables can also be taken into account (random effects).

In the context of this study, a mixed effect model would treat states as a random effect (randomly drawn sample from a larger population of states), and inferences related to states would apply to out of sample states. On the other, treating states as a fixed effect means that any inferences drawn from a model would not extend to out of sample states.
Another way to model clustered data is to treat it as a nuisance. Correlations within a cluster lead to standard errors being underestimated. A technique known as “robust standard errors”

is commonly used to correct this. Between cluster variance is ignored. Overdispersion can occur in a logistic model when the response proportions exhibit more variance than the theoretical binomial distribution variance. A reason for this may be omitted explanatory variable. Overdispersion is another source of underestimated standard errors and can be accounted for correct standard errors.

Based on the foregoing. The following logistic regressions models will be built and compared.

- State as fixed effect. Full model with no corrections to standard errors to allow for dispersion and clustering.
- State as fixed effect. Full model with corrections to standard errors to allow for dispersion and clustering.
- State as random effect. Null model to investigate extent of between-state variance.
- State as random effect. Fixed effect full model adapted with intercept only allowed to vary by state.
- State as random effect. Full fixed effect model state as a random effect on racial group slope. No state-level variables added.
- State as random effect, Full fixed effect model state as a random effect on racial group slope and state-level variable(s) added.

3.5 Evaluation

As the primary goal is to build an explanatory model, emphasis will be placed on goodness of fit measures. Predictor variables are standardised. As each variable is standardized, interaction effects can be better considered, and it is easier to see which variable has the greatest effect on the response variable. Model comparisons will be carried out using the Akaike information criterion (AIC). This is a penalised log-likelihood ratio measure designed to select the simplest (least number of parameters) model. Therefore, it is aligned to project goals. Model fit will also be assessed using McFadden's Pseudo-R2 statistic, and p-value significance tests. Model output will include 95% confidence intervals for predicted coefficients.

Logistic Regression Assumptions: Six assumptions apply.

1. The Response Variable is Binary or a Proportion

Proportion derived from aggregating individual binary case counts. Conceptually, individuals that appeared in the CDC data can be assigned a value of one, Individuals that make up the remainder of the population assigned a value 0. CDC data may contain individuals with repeat infections, but this is assumed negligible compared to true first-time cases.

2. The Observations are Independent.

Logistic regression assumes that the observations in the dataset are independent of each other. That is, the observations should not be related to each other in any way.

It is assumed that assumption is violated sue to clustering of racial groups within states, but can be accounted for by mixed-effects model or corrections to standard errors.

3. There is No Multicollinearity Among Explanatory Variables

Multicollinearity will be checked for using the variance inflation factor (VIF) method. Variables with a high VIF will be systematically removed based on correlations and subject area knowledge. A VIF value less than 9 is acceptable, but ideally all below 5.

4. There are No Extreme Outliers

This will be checked at both the data cleaning and modelling stages, Outliers will be assessed on a case-by-case basis and removed or replaced by imputed values where possible. Extreme values deemed genuine will not be removed, but their impact assessed. Outliers will be investigated by inspection, automated tools, Extreme impacts on model coefficients will be done using R beta() and diffft() functions.

5. There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable

This was checked using visual inspection of scatter plots and partial residual plots (also known as component-plus-residual plot).

6. The Sample Size is Sufficiently Large

Sample size of the dataset needs to large enough to draw valid conclusions from the fitted logistic regression model. Only racial groups with population size greater than 500 analysed. Also, within a state, a racial group residing in less than 5 counties is removed from the dataset.

3.6 Deployment

Deployment in this study comprised the following deliverables: final report, configuration manual, code artifacts and presentation.

4 Design Specification

4.1 Modelling Design

Within each US state different environmental, socio-economic and community health factors are at play. Diverse government and health policies affect the resources, approach, and ability to respond to the coronavirus pandemic crisis. The extent to which each racial group is impacted by the pandemic depends on their community status. For example, possible factors include their access to healthcare (health insurance), senior age and poverty levels.

4.2 Design Architecture

The high-level design architecture, process flow and main tools used are discussed in the Configuration manual. Data is downloaded as excel or .csv flat files, from the source websites. These are then loaded into Rstudio or Tableau software for analysis and pre-processing. Model building is carried out using RStudio. The persistent data is held in .CSV files.

5 Implementation

5.1 Tools Used

The implementation process made use of the following tools.
data import files and data persistent output files:

- .excel files,
- .csv files

Model building and presentation:

- R and R Studio

5.2 Data Exploration

The plots below show the results of some initial high-level data exploration.

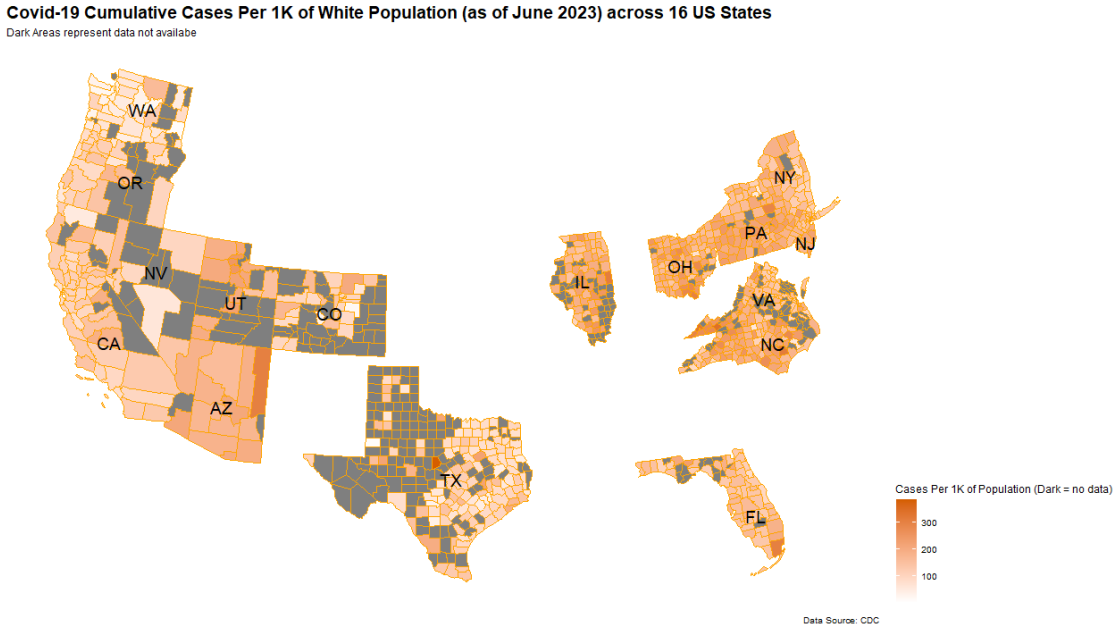

Fig 1. US map of states selected for study.

Fig 1 shows the 16 US states selected for the study. In the above example, a snapshot of cumulative case rates per 1k of population for the white population is shown. Not all counties in the US report their data to the CDC, and missing data is depicted by the grey areas.

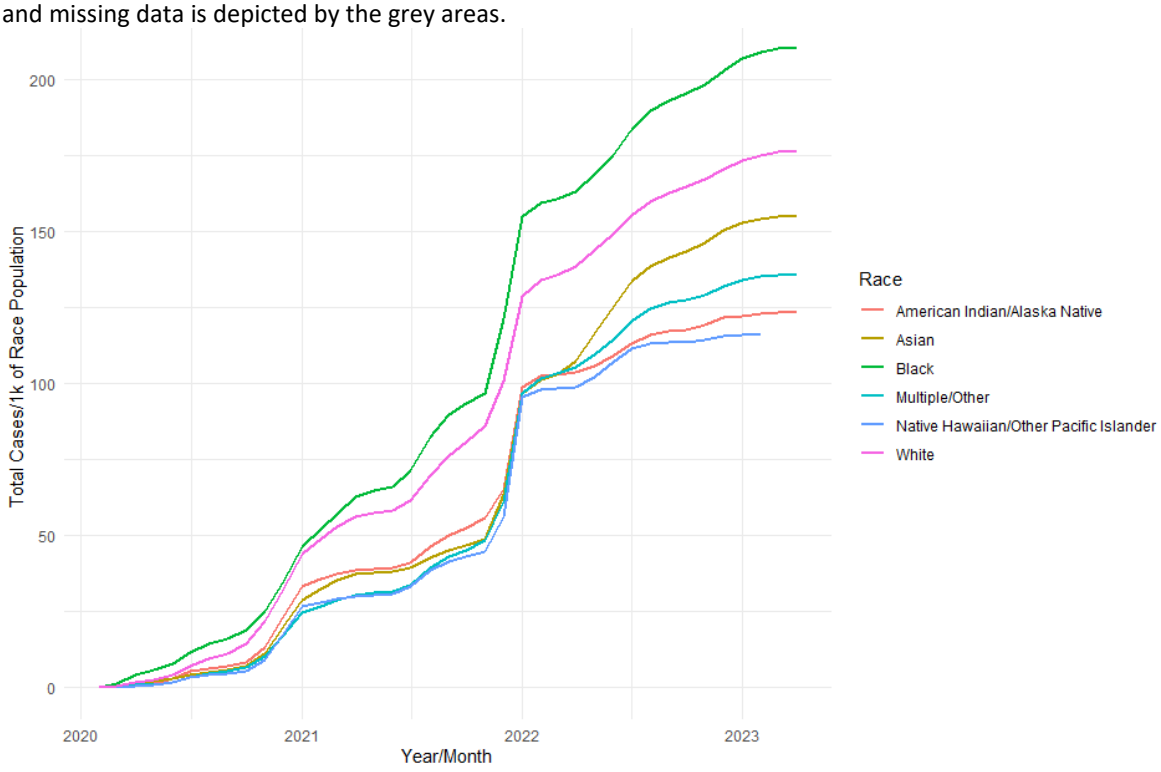

Fig 2 Cumulative cases per 1k of population

Fig 2 shows the cumulative monthly Covid-19 case rate curves for each race since Jan 2021. All curves show the same pattern with a massive spike ocurring around Dec 202/Jan 2022. Newspaper reports attributed this spike to the sudden emergence of the highly contagious Omicron strain of the Covid-19 virus.

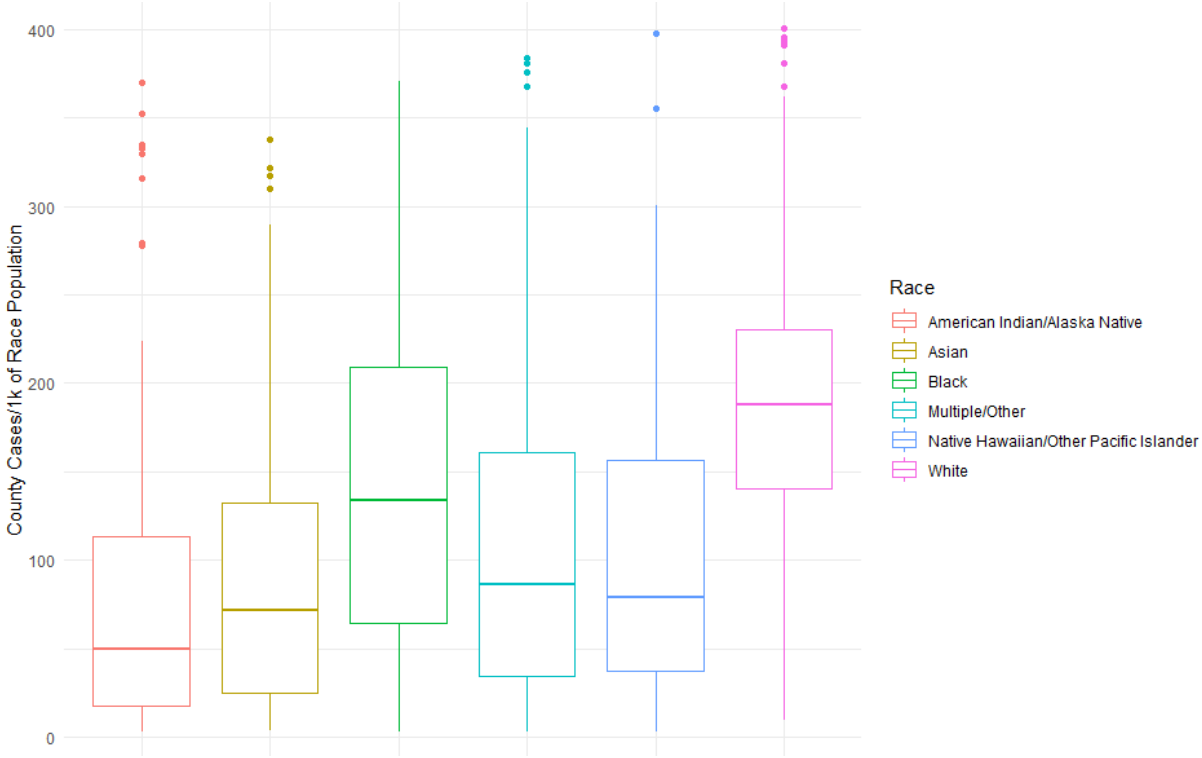

Fig 3 County cases/per 1k of Population in period Jan 2020 - Apr 2023

The boxplot in fig 3 depicts range of total cases, per 1k of population, for all counties within the 16 states. Evident are the large ranges and outlier flagged by the boxplots.

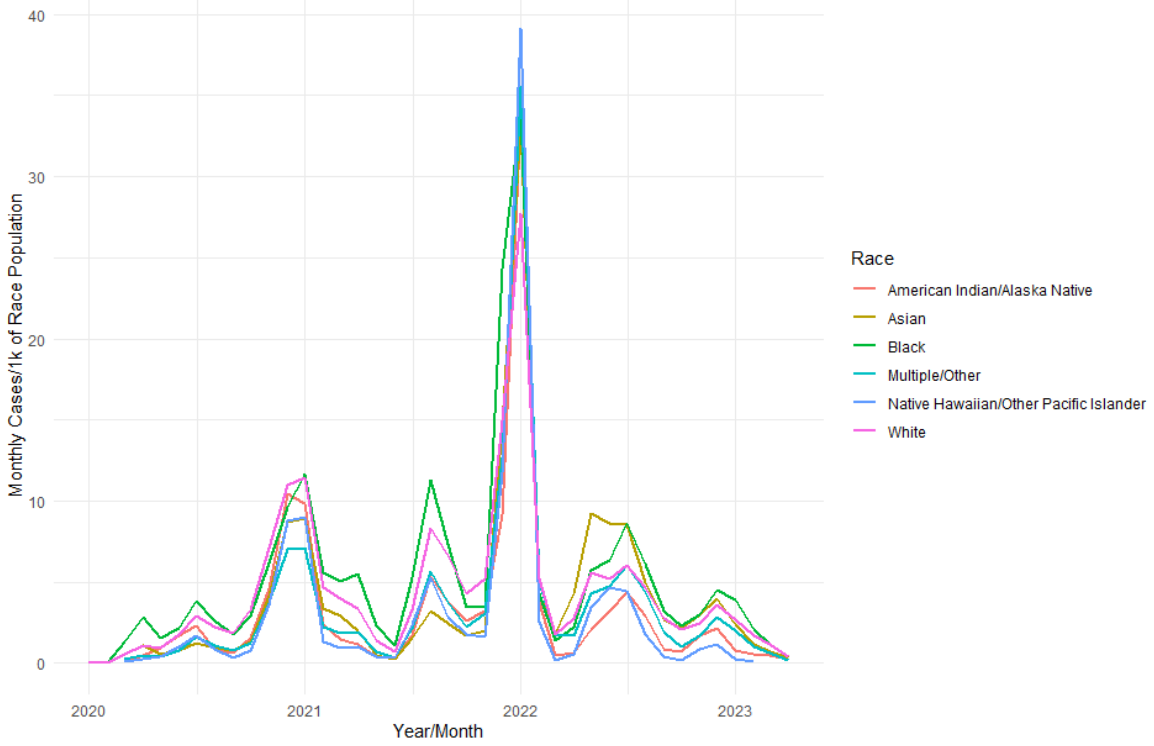

Fig 4. Monthly Cases/1k of population across selected States by Race

The monthly time-series of new cases shows very large volatility and a massive spike in and around Dec2020 coinciding with the arrival of the Omicron variant.

5.3 Feature Selection

Feature selection aims at removing redundant variables, retaining only those that explain the major part in the response variables variance. In this study, the primary focus is on inference therefore, it is crucially important that selected features show minimal multicollinearity and minimal correlation.

A total of 22 features were included in the study. Of these, 4 features applied only to the state level. Features were examined using a variety of plots, mostly scatter plots both at the identity and the logit scale for the dependent proportion variable of cases per population. The ggplot function from the ggplot2 package was used for most plot. Sometimes base R plot functions were used.

Correlation plots showed some features with statistically significant correlations. However, it was decided to defer any removal of features until the modelling stage, where correlations could be looked at in conjunction with VIF analysis and p-value significant tests. A cross table data frame of all correlations was built for easy future reference.

5.4 Chi-Square Test of Independence

In this research study, the null hypothesis is that the covid-19 incidence is independent of race. The alternate hypothesis is that there is an association and that infection rates differ between races. The table (Tab 2) below show the total number of cumulative cases by race across all 16 states in this study.

Covid-19 Infections	Asian	Black	White	Multiple /Other	American Indian/Alaska Native	Native Hawaiian/ Other Pacific Islander
Cases	2454796	5419656	27760153	872715	302442	44543
Non-Cases	13303317	20242182	129226196	5392686	2284644	510794
Case % of Population	15.58%	21.12%	17.68%	13.93%	11.69%	8.02%
Population	15758113	25661838	156986349	6265401	2587086	555337

Table 2. Chi-square test of independence for Covid-19 cases by race

Non-Cases are calculated by subtracting the cases from the total population. The Ch-square test indicates a strong association between race and covid prevalence (X-squared = 414947, 5 degrees of freedom, p-value < < 0.05). Hence the null hypothesis is rejected, and we can proceed with seeking a model and input features that help to explain the associations. The chi-squared test was carried out in R using the chisq.test function.

5.5 Model Development

In this section, development of the models described in section 3.4.2 (add internal link) are discussed. The models are built sequentially outputs. Often outputs such as narrowed list of candidate feature, lessons learned were fed into the next mode.

In addition to implementing the data cleaning and preparation steps outline sin 3.3, a deeper exploration of the data was carried out. Scatter plots showed extremely high variance and boxplots some extreme outliers.

To address these outliers, the time-series data was smoothed using the tsclean() function from the R forecast package. After the timeseries was cleaned, it was re-aggregated.

Prior to modelling, all numerical features were standardised.

5.5.1 Model 1. Logistic model with state as fixed effect. No corrections to standard errors

Starting with a full model of all county level features, the following automated feature selection algorithms were applied.
AICStep() Performs stepwise model selection by AIC (MASS Package)
Step(). Similar to AICStep(). Both forward and backward stepwise options tried
cv.glmnet(). Perform cross-validated lasso regression(glmnet package).
trainControl(). Cross-validation with 10-fold, repeated 100 time (Caret package).

Cross-validation and step-wise algorithms returned results showing all predictors highly significant. This was probably due to standard error estimates all being too small, due to clustering at least. Although lasso regression produced a reduced set of features,
The results were deemed unreliable due to very low standard errors.

5.5.2 Model 2. Logistic model with state as fixed effect. Corrections to standard errors and overdispersion.

Using a combination of vif output, aic and feature significance tests after allowing for overdispersion, a reduced set of features was obtained. A customised routine was written to perform both forward and backward stepwise aic tests and at the same time adjust for clustered errors. The goal was to obtain a set of features both optimised for minimal aic and maximal p-value significance using clustered errors corrections. Clustered errors corrections were carried using the coeff() function from the Sandwich package. Adjusting for overdispersion could have been done with quasibinomial regression, with the same results, but this method does not provide an aic value.
The method resulted in 6 county level features selected. This formed the final fixed effects logistic regression model. This model would be used as input to the mixed-effect logistic regression mode where state would become a random effect and the models would be compared.

5.5.3 Model 3. Intercept only mixed-effects Logistic model with state as random effect.

The goal of this model is simply to investigate the between state variance in the grand average proportion of Covid-19 cases across all states and races. This is known as the null model. Mixed efforts model uses package lme4.

5.5.4 Model 4. Mixed-effects Logistic model with state as random effect on intercept

Model compared with model 2.

5.5.5 Model 5. Mixed-effects Logistic model with state as a random effect and state level feature added.

This is Model 4 but with a state (level 2 hierarchical model) feature added.

5.5.6 Model 6. Mixed-effects Logistic model with state as a random effect against race slope.

Random slope model

6 Evaluation

6.1.1 Fixed Effects Models

Model 1, which did not account for overdispersion and independence of observations severely underestimated standard errors, gave rise to very small p-values for all features. For this reason, results were deemed unreliable, and the model no longer considered.

Model 2, took account of both overdispersion and dependencies on observations (clustering). With corrected standard errors, it was possible to prune original list of 18 county-level features down to 6.

The next step was to check the assumptions of logistic regression, described in 3.5.

1. The Response Variable is Binary or a Proportion
2. The Sample Size is Sufficiently Large
3. The Observations are Independent.
4. There is No Multicollinearity Among Explanatory Variables
5. There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable
6. There are No Extreme Outliers

Assumptions 1 and 2 are addressed by design. Case rates are expressed as a proportion of population (an estimate of mean probability of infection) and small sample sizes were excluded. Assumptions 3 and 4, were deemed to have not been met, but the model attempted to account for them.

To test linearity of logit response, an established method is to examine partial residual plots. Fig 5, below shows partial residual plots for the Model 2 variables.

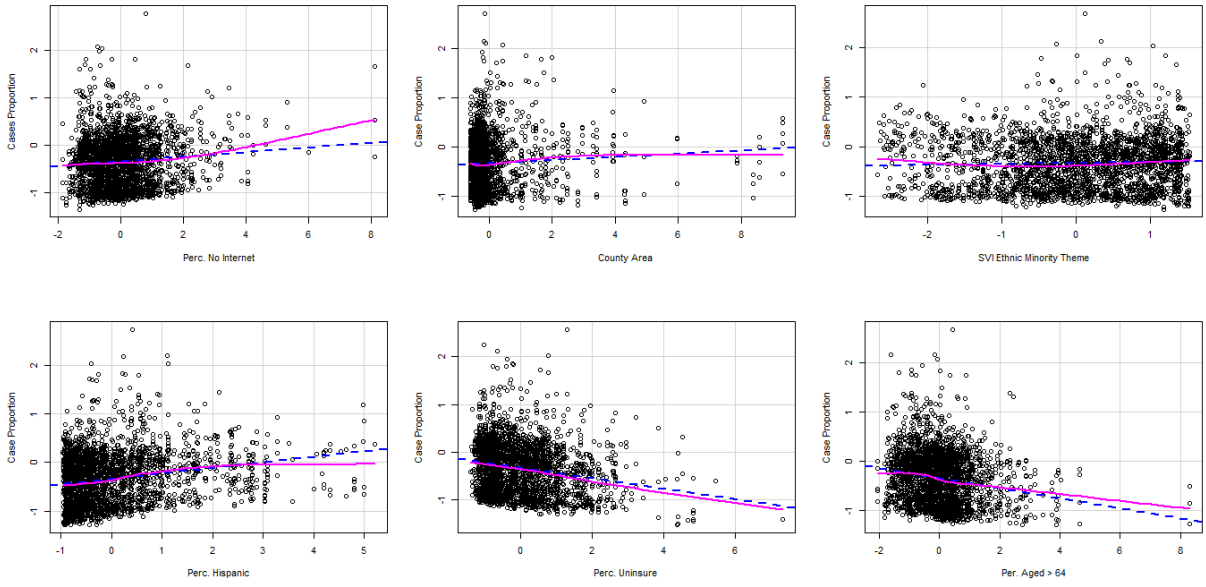

Fig 5 Partial Residual Plots

The pink link shows the actual model residual and the blue line the expected residuals if the relationship between the predictor feature and response was linear. In all cases the lines are close enough to provide evidence of a linear relationship.
The final assumption relates to extreme outliers. From previous data exploration, it appears in the nature of the Covid-19 data, that very extreme volatility is observed. To investigate influential outliers, the R function difft() was used. This measures the difference in fits, which arise when a particular observation is left out of the mode. Using difft(), 2 influential observations were found for white populations in California and Florida. However, these observations were deemed too important to be left out of the model/

6.1.2 Mixed Effects Models

Model 3 is known as the null model (intercept only) in mixed effects modelling. It tells us how much between-cluster variance there is. The figure for this model is 0.039, which means that just 3.9% of variance is accounted for by between-states in the model. This appears quite small. The table (Table 3) below shows the results of running models 4-6 with the best AIC obtained, McFaddens pseudo R2 does not apply to mixed models. The Nakagawa pseudo R2 was used to compare mixed effect models. Model 6 showed the lowest AIC, but gave a warning about failure to converge. Removing "Native Hawaiian/Other Pacific Islander" from the model eliminated the warning, but this was not an option. Therefore, this model was not considered further.

Model	**Type**	**Pseudo-R2**	**AIC**
Model2	Fixed Effect	McFadden 0.542	2115163
Model 4	Mixed Effect	Nakagawa 0.047	2115343
Model 5	Mixed Effect	Nakagawa 0.051	2115340
Model 6	Mixed Effect	Nakagawa 0.096	1829829.8 * Convergence warning

Table 3. Comparisons of model AIC metrics

The lowest AIC was for the fixed effects model, Model 2. On this basis and because it aligns with project goals as the simplest model, it was selected as the final model.
Tab 4 shows the summary output together with 95% confidence intervals using robust errors. Just the state California is shown for brevity.

County Feature	**Coefficient**	**Confidence Interval**		**Robust Standard Error**
		2.5%	**97.5%**	
California	-0.604	-0.677	-0.531	0.108
Asian	-0.156	-0.339	0.027	0.096
Black	-0.156	0.029	0.027	0.096
Multiple/Other	0.116	-0.168	0.204	0.044
American Indian/Alaska Native	0.055	-0.749	0.278	0.108
Native Hawaiian/Other Pacific Islander	-0.377	-0.299	-0.005	0.201
% Households without internet connection	-0.001	0.076	0.298	0.153
Area in square miles	0.098	0.006	0.120	0.030
SVI Racial & Ethnic Minority Status Theme	0.014	-0.127	0.022	0.008

% of Hispanic or Latino persons	-0.059	-0.010	0.009	0.059
% uninsured in 90th percentile	0.090	-0.148	0.190	0.046
% persons aged 65 or older	-0.094	-0.019	-0.039	0.035

Table 4. Summary for fixed-effects Model 2 using robust standard errors.

7 Discussion

The White race is set as the default category in the model The results in tab xx show that both American Indian/Alaska Native and the Multiple/Other racial groups have a positive log(odds). All other racial group have negative coefficients. For example, for the Black race group, log(odds) is -0.156, which exponentiated equates to odds of 0.855. In other words, controlling for all other variables, the odds for Blacks contracting \covid-19 are about 15% lower than Whites.

The explanatory features have a strong leaning towards ethnic minority status, which suggests this may be an important factor to consider in identifying communities vulnerable to Covid-19 infection. The overall SVI Racial and Ethnic Status theme shows a small positive log(odds) which suggests that communities higher rank in this theme are at more risk of Covid-19 than those lower. However, the log(odds) for percentage of persons aged 65 or over is negative, which appears suspect.

8 Conclusion and Future Work

The aim of this study was to build a simple effects model to establish if there was evidence for disparities in Covid-19 outcomes depending on racial group and if so, identify key SDH drivers to explain these disparities. The model appears to support the hypothesis that disparities do exist, backed up by the high-level Chi-squared test in 5.4. A logistic regression model was proposed as this seemed a natural choice to estimate probabilities of infection based on population size (exposure).

Given the data showed high volatility and extreme outliers, this created a challenge for using logistic regression modelling as dispersion was so high. But it is also possible that high dispersion can be partly attributed to key explanatory variables missed. Therefore, it is suggested that immediate future work would seek to improve the applicability of logistic regression. This would involve broadening the number of features considered, Perhaps, incorporating geo-spatial data such as neighboring counties Another possible candidate would be virus strains, as the timeseries data shows a massive spike which my be due to the arrival of the omicron strain. The emergence of different highly contagious creates time-varying changes in probabilities of infection which should be accounted for. Logistic regression could also be extended in other ways such as incorporating monthly time-series data as repeated measures. Interactions between features can also be explored

References

Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., Li, J., & Zand, R. (2021). Racial, Economic, and Health Inequality and COVID-19 Infection in the United States. *Journal of Racial and Ethnic Health Disparities*, *8*(3). https://doi.org/10.1007/s40615-020-00833-4

Barry, V., Dasgupta, S., Weller, D. L., Kriss, J. L., Cadwell, B. L., Rose, C., Pingali, C., Musial, T., Sharpe, J. D., Flores, S. A., Greenlund, K. J., Patel, A., Stewart, A., Qualters, J. R., Harris, L. T., Barbour, K. E., & Black, C. L. (2021). Patterns in COVID-19 Vaccination Coverage, by Social Vulnerability and Urbanicity — United States, December 14, 2020–May 1, 2021. *MMWR Recommendations and Reports*, *70*(22). https://doi.org/10.15585/mmwr.mm7022e1

Biggs, E. N., Maloney, P. M., Rung, A. L., Peters, E. S., & Robinson, W. T. (2021). The Relationship Between Social Vulnerability and COVID-19 Incidence Among Louisiana Census Tracts. *Frontiers in Public Health*. https://doi.org/10.3389/fpubh.2020.617976

Braveman, P., & Gottlieb, L. (2014). The social determinants of health: It's time to consider the causes of the causes. *Public Health Reports*, *129*(SUPPL. 2). https://doi.org/10.1177/00333549141291s206

Chapman, P. (1999). The CRISP-DM User Guide. *The CRISP-DM User Guide*.

Hawkins, R. B., Charles, E. J., & Mehaffey, J. H. (2020). Socio-economic status and COVID-19–related cases and fatalities. *Public Health*, *189*. https://doi.org/10.1016/j.puhe.2020.09.016

Heckler, M. M. (1986). Report of the Secretary's Task Force on Black and Minority Health. In *MMWR. Morbidity and mortality weekly report* (Vol. 35). US Department of Health and Human Services.

Karaye, I. M., & Horney, J. A. (2020). The Impact of Social Vulnerability on COVID-19 in the U.S.: An Analysis of Spatially Varying Relationships. *American Journal of Preventive Medicine*, *59*(3). https://doi.org/10.1016/j.amepre.2020.06.006

Kelly, M. P. (2021). The relation between the social and the biological and COVID-19. *Public Health*, *196*, 18–23. https://doi.org/10.1016/J.PUHE.2021.05.003

Khanijahani, A., Iezadi, S., Gholipour, K., Azami-Aghdash, S., & Naghibi, D. (2021). A systematic review of racial/ethnic and socioeconomic disparities in COVID-19. In *International Journal for Equity in Health* (Vol. 20, Issue 1). https://doi.org/10.1186/s12939-021-01582-4

Lopez, L., Hart, L. H., & Katz, M. H. (2021). Racial and Ethnic Health Disparities Related to COVID-19. In *JAMA - Journal of the American Medical Association* (Vol. 325, Issue 8). https://doi.org/10.1001/jama.2020.26443

Mackey, K., Ayers, C. K., Kondo, K. K., Saha, S., Advani, S. M., Young, S., Spencer, H., Rusek, M., Anderson, J., Veazie, S., Smith, M., & Kansagara, D. (2021). Racial and Ethnic Disparities in COVID-19-Related Infections, Hospitalizations, and Deaths : A Systematic Review. *Annals of Internal Medicine*, *174*(3), 362–373. https://doi.org/10.7326/M20-6306

National Center for Immunization and Respiratory Diseases (NCIRD), D. of V. D. (2021). *Risk for COVID-19 Infection, Hospitalization, and Death By Age Group | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html

Nayak, A., Islam, S. J., Mehta, A., Ko, Y. A., Patel, S. A., Goyal, A., Sullivan, S., Lewis, T. T., Vaccarino, V., Morris, A. A., & Quyyumi, A. A. (2020). Impact of social vulnerability on COVID-19 incidence and outcomes in the United States. In *medRxiv*. https://doi.org/10.1101/2020.04.10.20060962

Oates, G. R., Juarez, L. D., Horswell, R., Chu, S., Miele, L., Fouad, M. N., Curry, W. A., Fort, D., Hillegass, W. B., & Danos, D. M. (2021). The Association Between Neighborhood Social Vulnerability and COVID-19 Testing, Positivity, and Incidence in Alabama and Louisiana. *Journal of Community Health*, *46*(6). https://doi.org/10.1007/s10900-021-00998-x

Rogers, T. N., Rogers, C. R., VanSant-Webb, E., Gu, L. Y., Yan, B., & Qeadan, F. (2020). Racial Disparities in COVID-19 Mortality Among Essential Workers in the United States. *World Medical and Health Policy*, *12*(3). https://doi.org/10.1002/wmh3.358

Rozenfeld, Y., Beam, J., Maier, H., Haggerson, W., Boudreau, K., Carlson, J., & Medows, R. (2020). A model of disparities: Risk factors associated with COVID-19 infection. *International Journal for Equity in Health*, *19*(1). https://doi.org/10.1186/s12939-020-01242-z

Shortreed, S. M., Gray, R., Mary, ·, Akosile, A., Walker, R. L., Fuller, S., Temposky, L., Fortmann, S. P., Albertson-Junkans, L., Floyd, J. S., Bayliss, E. A., Harrington, L. B., Lee, M. H., & Dublin, S. (2021). Increased COVID-19 Infection Risk Drives Racial and Ethnic Disparities in Severe COVID-19 Outcomes. *Journal of Racial and Ethnic Health Disparities*. https://doi.org/10.1007/s40615-021-01205-2

Zoungrana, T. D., Yerbanga, A., & Ouoba, Y. (2022). Socio-economic and environmental factors in the global spread of COVID-19 outbreak. *Research in Economics*. https://doi.org/10.1016/J.RIE.2022.08.001