Preventive Medicine Reports 28 (2022) 101858

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

An optimized machine learning model for identifying socio-economic, demographic and health-related variables associated with low vaccination levels that vary across ZIP codes in California

George Avirappattu [a,*], Alfred Pach III [b], Clarence E. Locklear [c], Anthony Q. Briggs [d]

[a] Center for Data Analytics, School of Mathematical Sciences, Kean University, NJ, USA
[b] Department of Medical Sciences, Hackensack Meridian School of Medicine, Nutley, NJ, USA
[c] Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA
[d] NYU Langone Health, Grossman School of Medicine, New York University, New York, USA

ARTICLE INFO

Keywords:
COVID-19
Vaccination uptake
Sociodemographic determinants
Spatial assessment
Machine learning

ABSTRACT

There is an urgent need for an in-depth and systematic assessment of a wide range of predictive factors related to populations most at risk for delaying and refusing COVID-19 vaccination as cases of the disease surge across the United States. Many studies have assessed a limited number of general sociodemographic and health-related factors related to low vaccination rates. Machine learning methods were used to assess the association of 151 social and health-related risk factors derived from the American Community Survey 2019 and the Centers for Disease Control and Prevention (CDC) BRFSS with the response variables of vaccination rates and unvaccinated counts in 1,555 ZIP Codes in California. The performance of various analytical models was evaluated according to their ability to regress between predictive variables and vaccination levels. Machine learning modeling identified the Gradient Boosting Regressor (GBR) as the predictive model with a higher percentage of the explained variance than the variance identified through linear and generalized regression models. A set of 20 variables explained 72.90% of the variability of unvaccinated counts among ZIP Codes in California. ZIP Codes were shown to be a more meaningful geo-local unit of analysis than county-level assessments. Modeling vaccination rates was not as effective as modeling unvaccinated counts. The public health utility of this model provides for the analysis of state and local conditions related to COVID-19 vaccination use and future public health problems and pandemics.

1. Introduction

SARS-CoV-2 cases were first identified in the United States (US) in January of 2020. As of January 11, 2022, the US has the highest number of confirmed cases of COVID-19 (61,732,283) and deaths (837,274) among countries of the world (CDC, 2022). The US numbers of COVID-19 represent 25% of cases worldwide and 16% of deaths, although the US is only 4% of the world's population (S A, 2020). The United States has developed three vaccines against COVID-19 (i.e., Pfizer-BioNTech, Moderna, and Johnson and Johnson), with approval for emergency use throughout the country by April-May 2021. Although widely available, there has only been a moderate uptake of these vaccines, with 62.6% of the US population fully vaccinated (ages 5+, 1/11/22). The highly transmissible omicron variant of the coronavirus has increased the spread of the virus across the US, with a weekly average of 586,391 new cases a day of COVID-19 and 1,246 related deaths (CDC, 2022; Romano, 2022).

There is an urgent need to identify high-risk populations most likely to refuse or delay vaccination in order to focus on public health interventions to support the uptake of COVID vaccines, which reduce hospitalizations and deaths and, to an extent, neutralize transmission (Finney Rutten et al.). This study involves the development and application of the statistical tools of machine learning to provide a more powerful mode of analysis than standard statistical measures in order to evaluate a comprehensive range of socioeconomic, demographic, and health-related variables associated with the risk of not participating in a COVID-19 vaccination (Statistical, 2001; Carmichael and Marron, 2017). This analysis is combined with a comparative assessment of the

* Corresponding author at: Center for Data Analytics, The School of Mathematical Sciences, 1000 Morris Avenue, Kean University, NJ 07083, USA.
E-mail address: gavirapp@kean.edu (G. Avirappattu).

https://doi.org/10.1016/j.pmedr.2022.101858
Received 26 January 2022; Received in revised form 6 May 2022; Accepted 6 June 2022
Available online 10 June 2022

response measures of vaccination status (i.e., vaccination rates and unvaccinated counts) and geographic parameters (i.e., ZIP CODES and county-level) to be able to identify a large range of explanatory factors related to those at risk for not participating in a COVID-19 vaccination.

1.1. Issues in the uptake of vaccines

Studies have attributed much of the low use of COVID-19 vaccines to vaccine hesitancy, which the World Health Organization defines as involving a lack of confidence in the safety and effectiveness of vaccines, problems with access to them, and a low perceived risk of disease or need for a vaccine (MacDonald; Sallam, 2021). Several studies across the US have found that vaccine hesitancy is associated with socioeconomic, demographic, and health-related variables (SEDH). These factors involve issues related to race and ethnicity (Kricorian and Turner, 2021; Mm, 2021), age (Al-Mohaithef et al.), income (Wang et al.; Soares et al., 2021), low education (Williams et al.; Aw et al., 2021), gender (Woolf et al., 2021), and lack of health insurance (Lindemer et al., 2021).

Mounting evidence also indicates significant geographic patterns in the variation of COVID-19 vaccination rates. Locations designated by ZIP code (postal code used by the United States Postal Service to uniquely identify residences), county-level, and census track designations have shown disproportionate numbers of unvaccinated individuals (Mm, 2021; Bruckhaus et al., 2021). Geospatial analysis and mapping of socioeconomic, demographic, and health-related factors (SEDH) have demonstrated significant variation in their association with neighborhood and location-based rates of COVID-19 vaccination (Mollalo and Tatar, 2021; Kearney et al., 2021).

However, researchers have also observed that most studies of COVID-19 cases and vaccination status have involved a limited number of SEDH variables (Kearney et al., 2021; Karaye and Horney, 2020). In addition, many of the variables analyzed have involved general analytical categories (e.g., minority status, income). Such analyses cannot more specifically account for populations' diverse behavioral, social, and health-related characteristics. Related to this issue is the common focus on a scale of analysis typically at the county level, which often involves large and diverse populations (Mm, 2021; Bruckhaus et al., 2021). This compounds the difficulty of identifying and targeting interventions to the specific population segments and locales of those most at risk (Mollalo and Tatar, 2021). These approaches have led to gaps in our knowledge and understanding of barriers to vaccination.

We use machine learning methods to address these problems by assessing associations between a large set of SEDH predictive variables and COVID-19 vaccination levels (measured by vaccination rates and unvaccinated counts) in 1,555 ZIP codes in California. We assess the performance of various analytical methods according to their ability to regress between predictive variables and vaccination levels and choose the best method using cross-validation. We utilize ZIP codes as a more localized unit of analysis, in contrast to assessing vaccination behavior at the county level, demonstrating a more localized identification of diverse vaccination risk factors and rates of COVID-19 vaccination.

Utilizing a powerful statistical tool and developing a more meaningful and applicable methodological approach offers a means to improve public health information and interventions.

1.2. Research categories and analysis

1.2.1. Geographic level of analysis

We demonstrate that the analysis of SEDH measures related to vaccination rates per county provides a weak measure of vaccination prevalence. For example, one of the 50 counties, Los Angeles County, is home to 9,811,939 people out of 37,154,935 in California, the largest state in the US. It has an unvaccinated population of 4,525,164 residents as of July 12, 2021 (Health CDoP, 2021). This situation presents two concerns when county-level rates are used to measure the prevalence of COVID-19 vaccination. First, counties with such large populations as Los Angeles County have vastly diverse distributions of socioeconomic, demographic, and health-related determinants at neighborhood/location-based levels. It is, therefore, challenging to meaningfully associate influential social and health-related covariates of vaccination rates with the local segments of the population most at risk for being unvaccinated. Second, although the vaccination rate is 50% in Los Angeles County, that still leaves a substantial number (4,525,164) of people unvaccinated. Our ZIP Code-based analysis addresses these concerns as ZIP Code populations are smaller and more local than counties, reducing issues of scale in context evaluations for targeting interventions.

1.2.2. Vaccination risk variables

The most commonly utilized sociodemographic measures to identify barriers to COVID-19 vaccination involve the set of SEDH factors that make up the CDC's Social Vulnerability Index (Bruckhaus et al., 2021; Barry et al., 2021). The CDC created this Index to manage the social and health needs of communities experiencing disasters, including disease outbreaks (Barry et al., 2021; Flanagan et al., 2011). It includes 15 variables related to socioeconomic status, household composition, minority status and language, and housing type. A number of the variables or subcomponents within these thematic domains include general, non-specific social indicators (e.g., minority status), which are often used to explain situational barriers and disparities in COVID vaccination rates and counts of unvaccinated individuals (Mm, 2021; Mollalo and Tatar, 2021).

We expand and supplement the 15 SVI variables to 151 SEDH determinants as influencers on participation in COVID-19 vaccination programs (see Appendix for a complete list). This more extensive set improves our ability to explain differences in COVID-19 vaccination rates and counts between locations.

1.2.3. Data analytics

For data analytics, we primarily use machine learning methods. Machine learning methods, in general, can handle a large number of predictors, which is the case here, more than traditional statistical methods. Machine learning methods are algorithmic and operate differently on data sets than traditional statistical methods that are typically parametric model-based (Statistical, 2001; Carmichael and Marron, 2017). We do not make substantial assumptions like non-collinearity or normality on predictor variables in machine learning. Machine learning methods' performance is generally superior to traditional statistical models (Raschka, 2021).

In the following, the methods section starts with listing the data sources and exploratory statistical analysis on the variability of SEDH factors in the counties of California and their correlation with vaccination rates and the number of unvaccinated individuals. Then we detail the preprocessing, feature selection, modeling method selection, hyperparameter tuning, bootstrapping, and prediction stages of machine learning. The results section lists the chosen SEDH predictor variables, their correlations with the unvaccinated counts and vaccination rates, and an assessment of their collective ability to explain variation in response variables between ZIP codes. We also provide a ranking of the predictor variables according to their importance in the modeling. We conclude the paper with a discussion on the implications of our findings.

2. Methods

2.1. Data and sources

COVID-19 vaccination rates for each ZIP code in California are from the State Department of Health, CA (accessed on July 12, 2021) (Health CDoP, 2021). We obtained socioeconomic and demographic data for ZIP codes through a Python package uszipcode on October 19, 2020 (uszipcode, 2020). The primary source for the data is the US Census Bureau American Community Survey 2019 (ACS2019) (American, 2019) update for ZIP Code Tabulation Areas (ZCTA). Health

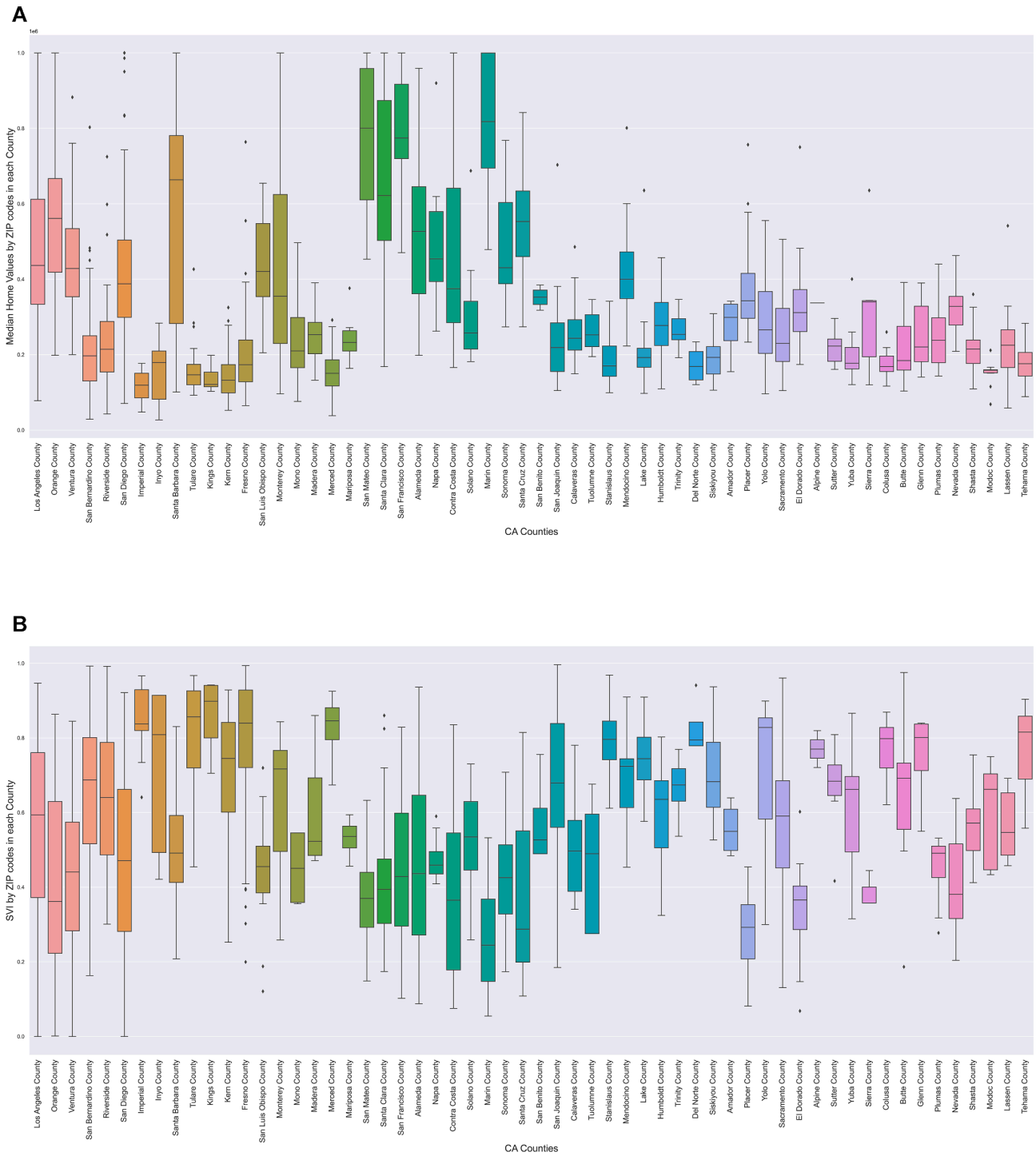

Fig. 1. A and B: Box plots illustrate within-county variations on median household income and the Social Vulnerability Index (SVI) within the counties listed on the x-axis. Significant variations of many of these variables, especially within large counties, make it very challenging to understand how they may play a role in vaccination prevalence if measured only by counties, as is done in much of the literature and media.

determinants data are based on the Behavioral Risk Factor Surveillance System (BRFSS) (BRFSS, 2020) obtained from the CDC on five chronic disease-related unhealthy behaviors, thirteen health outcomes, and nine on the use of preventive services. Both data sets contain model-based ZCTA estimates from sample surveys. Weighted counts are used for the region, region by age group, region by gender, and region by race and ethnicity (Bureau, 2022).

This paper is an observational study based on the vaccination data on the population aged 12 and over from the 1,555 residential ZIP codes of California. We obtained COVID-19 vaccination rates (percent of the population fully vaccinated) from the government website (Health CDoP, 2021) and calculated unvaccinated counts (i.e., counts of not fully vaccinated with two shots of Pfizer or Moderna, or one shot of Johnson and Johnson) using the population estimates from ACS2019 on each ZIP code. All data are publicly available in anonymized databases and determined to be exempt from additional ethical compliance measures by the Institutional Review Board of Kean University.

2.2. Exploratory analyses

From ACS2019 and BRFSS data, we select 151 SEDH variables (see the complete list in the Appendix) and begin with an analysis of variation within Zip codes among counties.

2.2.1. Box plots

Our exploratory analysis starts with a sample set of box plots to illustrate how the SEDH determinants in each ZIP code vastly vary within the 50 counties of California. Considerable variation within counties makes assessing the association of those determinants with county-based COVID-19 vaccination rates, the standard measure of vaccination prevalence, limited in targeting public health programs.

2.2.2. Correlations

We estimate the strength of the linear correlation of each select variable (features) to unvaccinated counts and vaccination rates by ZIP codes through Pearson correlation coefficients and p-values(significance level). Scatter plots demonstrate the strength of linear association between these variables with unvaccinated counts and vaccination rates.

2.3. Predictive modeling with machine learning methods

2.3.1. Preprocessing

We scale both predictor and target variables using the standard scalar. The standard scaler calculates the *z*-value, $\frac{\overline{x}-x^{(i)}}{s}$, where s is the sample standard deviation.

2.3.2. Model selection

To assess the performance of several machine learning methods and select the one with the best performance , we try these methods on the whole data set with 5-fold cross-validation according to "explained_variance" criteria, see Fig. 2 in results section

The following is an example of the data partition for 5-fold cross-validation.

1	2	3	4	5
Train Data	Train Data	Train Data	Cross-validation Data	Train Data

In general, for N-fold cross-validation, we partition the data into N-parts (Hastie et al., 2009). Let $\pi : [1, 2, 3, \cdots, N] \rightarrow 1, 2, 3, \cdots, K$ be an indexing function that randomly picks one of the k parts of the data not chosen yet. Let $\widehat{f}_{-k}(x)$ be the fitted function, computed with the k^{th} part of the removed. The cross-validation loss estimate $CV(\widehat{f})$ is the measure of error by which the prediction may differ from the actual values of y, the target variable:

$$CV(\widehat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i - \widehat{f}_{-\pi(i)}(x_i)\right)$$

The function L is the criteria, for us, of the variance explained ("explained_variance"), that we use to assess the performance of the fitted function to model the data.

The explained variance: $L(y - \widehat{y}) = 1 - \frac{Var(y-\widehat{y})}{Var(y)}$, where y and $\widehat{y}$, the actual and predicted values of the dependent variable, and $Var(x)$ is the variance of random variable x. The largest value of L is one (1), and the least is zero(0) when f is a constant function; the larger the L value, the better the capacity of the predictors in explaining the variability in the response variable. The machine learning method with the highest "explained_variance" is the best.

Table 1A
Correlations between predictor variables and unvaccinated counts.

Feature Variable: Proportion of	r	r^2	p-value
Males age 60–64	−0.53	0.28	0
Females age 20–24	0.35	0.12	0
Females age 15–19	0.31	0.09	0
Females age 10–14	0.30	0.09	0
residents in Nursing Homes	0.05	0.00	0
Whites	−0.42	0.18	0
Households with Annual Income > 200 K	−0.20	0.04	0
Households with Annual Income < 60 K	0.11	0.01	0
Proportion with Doctorate level Education	−0.18	0.03	0
Proportion Children in Private School	−0.18	0.03	0
Proportion Homes Vacant	−0.37	0.13	0
Homes Vacant for Sale	0.33	0.11	0
Vacant Other Reasons	−0.10	0.01	0
Rented but Unoccupied	0.11	0.01	0
Homes Vacant for Migrant Workers	−0.11	0.01	0
Part Time Workers	−0.08	0.01	0
car commuters	0.16	0.03	0
public commuters	0.15	0.02	0
30 – 40 Min to work commuters	0.07	0.01	0
homes built before 1930	−0.14	0.02	0

Table 1B
Correlations between predictor variables and vaccination rates.

Feature Variable: Proportion of	r	r^2	p-value
Females age _85_Plus	0.17	0.03	0
Residents in Nursing Homes	0.07	0.00	0.01
Asians	0.41	0.17	0
A.I. and Alaskan	−0.25	0.06	0
Hawaiian and Pacific Islanders	0.07	0.01	0.01
Homes with Value <25K	−0.32	0.10	0
Homes with Value <150K	−0.49	0.24	0
Homes with Value <200K	−0.49	0.24	0
Homes with Value <750K	0.46	0.21	0
Proportion with Professional Education	0.49	0.24	0
Homes Vacant	−0.12	0.01	0
Public commuters	0.26	0.07	0
Over 90Min to work commuters	−0.14	0.02	0
taxi commuters	0.10	0.01	0

2.3.3. Variable selection

Using an algorithmic forward selection scheme31, we examine the additive contributions of each of the 151 variables. This process helps us narrow the variable list to an optimal set of thirty-one (31) for predictive analytics for unvaccinated counts and vaccination rates. We then remove variables with high variance inflation factors (VIF > 3.5) to reduce variable counts to twenty (20) and fourteen (14), respectively.

2.3.4. Parameter tuning and testing

We find the optimal hyper-parameter settings through a grid search. With optimized hyper-parameters, two hundred models are constructed on bootstrapped train data (80%). Each model then predicts on the 20% set-aside test data. We bootstrap for two reasons: 1. Every ZIP code will be in the test set about forty (40 = 200*0.2) times; their average makes predictions robust, and 2. It eliminates reporting just the best test case scenario. For each of these 200 bootstrapped models, we also keep track of the feature importances and average them for robustness.

3. Results

3.1. Exploratory analysis

3.1.1. Within county variation of socioeconomic, structural, and health-related variables

Fig. 1A and 1B are collections of fifty box-and-whisker plots, one per county on a sample of two SEDH variables. The significant within-county variations illustrate the diminished potential for interpreting these SEDH variables' association with county-based vaccination rates or counts.

3.1.2. Correlation of socioeconomic, demographic, and health-related variables to vaccination counts and rates measured by ZIP codes in California

Tables 1A and B list all features we selected for modeling unvaccinated counts and vaccination rates together with linear correlation coefficients and p-values. There is an overlap with those in the Social Vulnerability Index but provides additional variables and more specified social categories.

The first scatter plot group shows correlations between the select feature variables (used in our modeling) and unvaccinated counts by ZIP codes. The second shows correlations between feature variables and vaccination rates by ZIP codes.

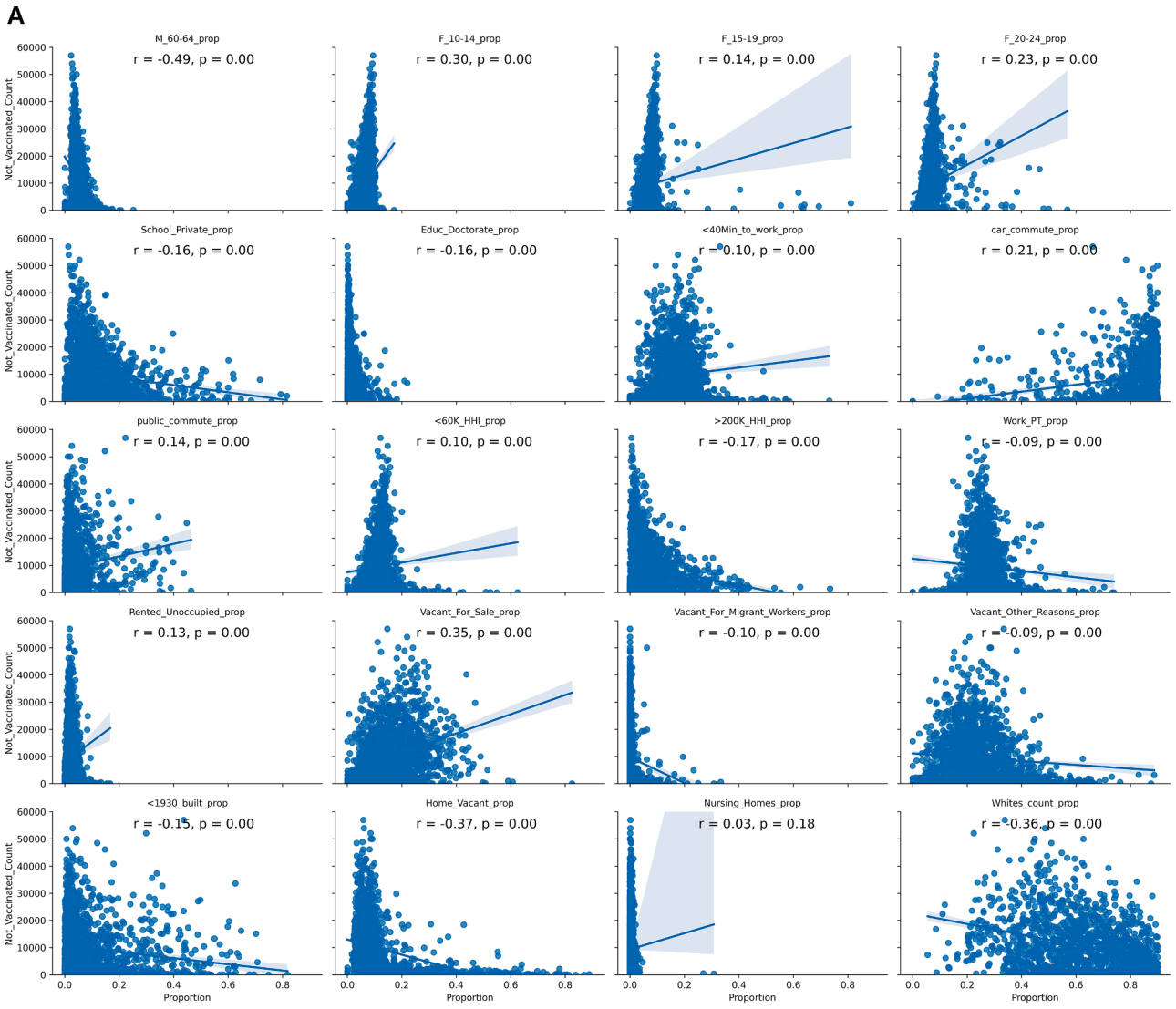

Fig. 2A. Correlations between proportions of each feature group in ZIP codes in California and unvaccinated counts. As shown in the table above, these feature variables correlate significantly (p = 0) to the vaccination counts.

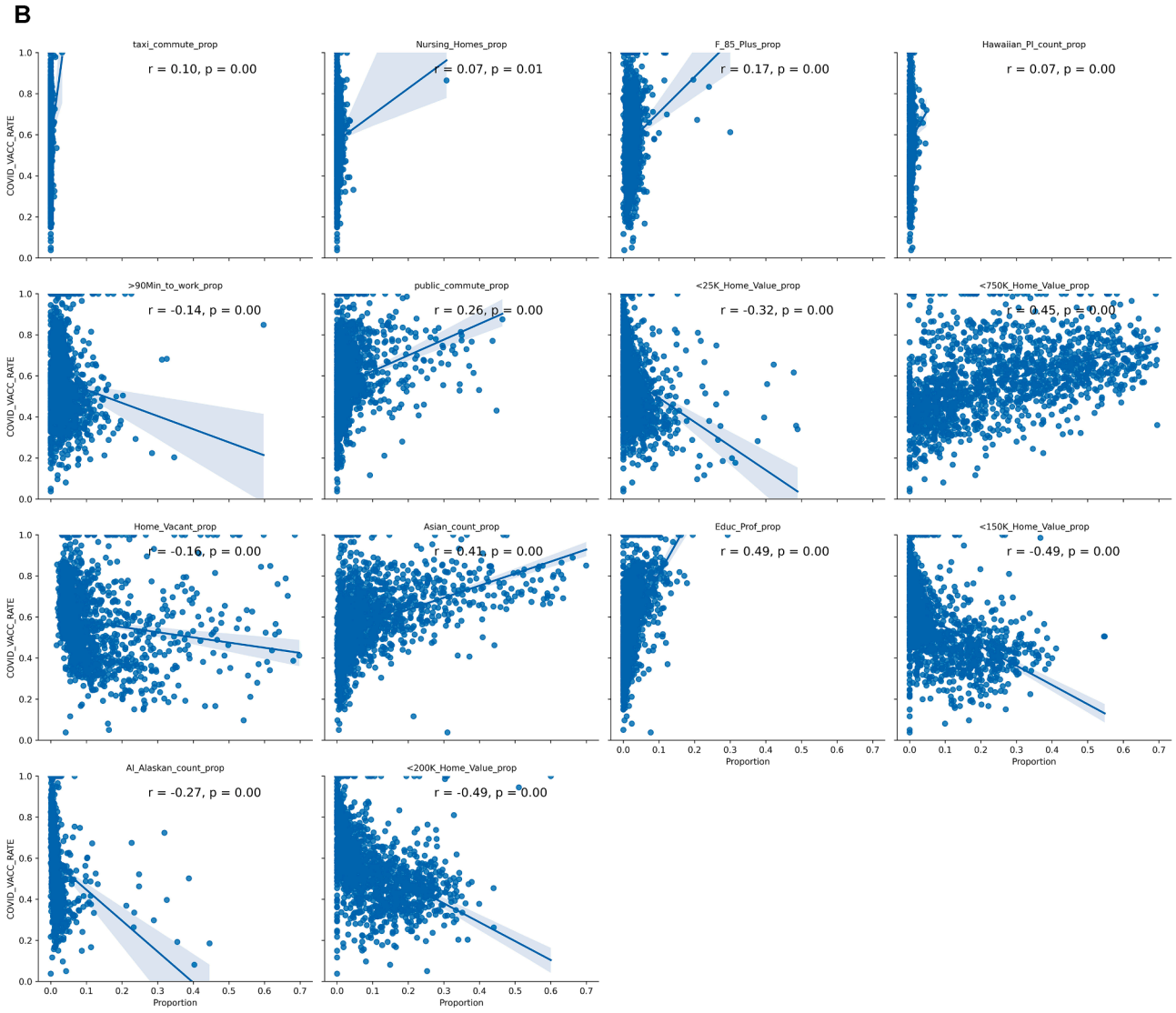

Fig. 2B. Correlations between proportions of each feature group in ZIP codes in California and vaccination rates. These feature variables correlate significantly to the vaccination rates, as shown in the table above.

A

Model	Error Metric	Variance Explained(+/- 2SD)
LinearRegression	CV-5 Variance Explained:	0.40 (+/- 0.08)
Lasso	CV-5 Variance Explained:	0.00 (+/- 0.00)
Ridge	CV-5 Variance Explained:	0.40 (+/- 0.07)
ElasticNet	CV-5 Variance Explained:	0.04 (+/- 0.05)
DecisionTreeRegressor	CV-5 Variance Explained:	0.08 (+/- 0.26)
SVR	CV-5 Variance Explained:	0.59 (+/- 0.25)
KNeighborsRegressor	CV-5 Variance Explained:	0.43 (+/- 0.07)
RandomForestRegressor	CV-5 Variance Explained:	0.62 (+/- 0.12)
GradientBoostingRegressor	CV-5 Variance Explained:	0.64 (+/- 0.12)

B

Model	Error Metric	Variance Explained(+/- 2SD)
LinearRegression	CV-5 Variance Explained:	0.37 (+/- 0.42)
Lasso	CV-5 Variance Explained:	0.00 (+/- 0.00)
Ridge	CV-5 Variance Explained:	0.49 (+/- 0.29)
ElasticNet	CV-5 Variance Explained:	0.00 (+/- 0.00)
DecisionTreeRegressor	CV-5 Variance Explained:	-0.05 (+/- 0.58)
SVR	CV-5 Variance Explained:	0.50 (+/- 0.25)
KNeighborsRegressor	CV-5 Variance Explained:	0.41 (+/- 0.32)
RandomForestRegressor	CV-5 Variance Explained:	0.48 (+/- 0.24)
GradientBoostingRegressor	CV-5 Variance Explained:	0.49 (+/- 0.24)

Fig. 3. A and B: ML Method selection - Modeling Unvaccinated Counts and Vaccination Rates in CA. These tables list different machine learning methods we tried on our data set before selecting the best one. We use 5-fold cross-validation to assess the performance of each method through "explained_variance" criteria. 20% of the data is set aside for each fold, and the model is trained on the other 80%. Then the trained model predicts on the 20% and calculates variance explained between the prediction $\hat{y}$ and the actual y values: $L(y-\hat{y}) = 1 - \frac{Var(y-\hat{y})}{Var(y)}$. The higher the variance explained, the better the model.

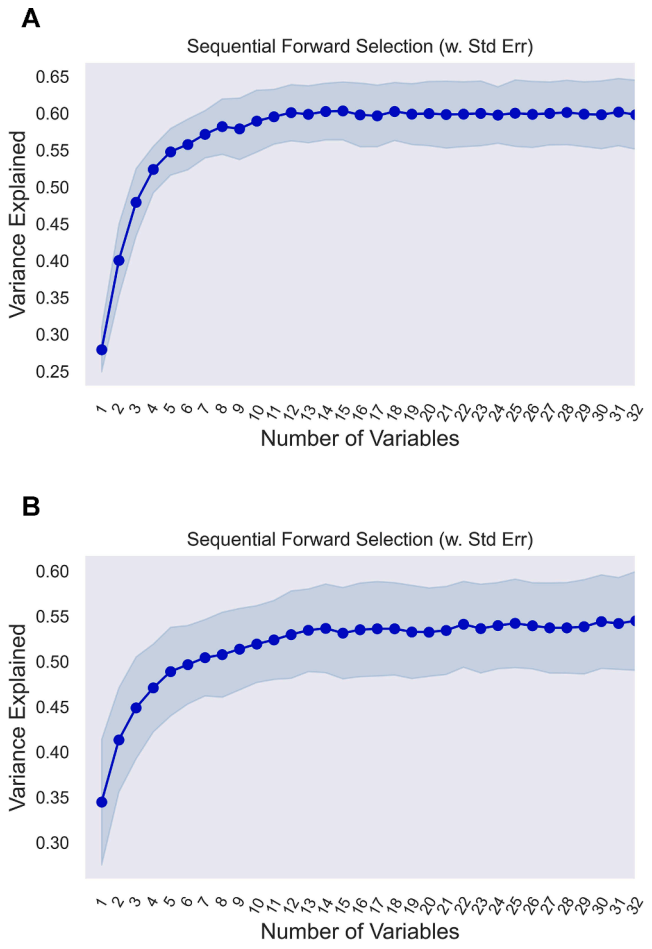

Fig. 4. A and B: Forward Sequential Selection considering variance explained by each variable starting from largest until a maximum is reached - for Unvaccinated Counts and Vaccination Rates.

3.2. Machine learning modeling: Selected SEDH variables explain a significant variation between ZIP codes on vaccination rates and unvaccinated counts

We consider machine learning methods including linear regression with and without regularization (e.g., lasso, ridge, and elastic net) , Decision Tree Regressor , Support Vector Regressor and Gradient Boosting Regressor (GBR) (Scikit-learn: Machine Learning in Python, 2011). See Fig. 3A and B for a complete list and their 5-fold cross-validation performance for unvaccinated counts and vaccination rates. GBR outperforms the other methods, with a variance explained value of 0.64 for vaccination rates, and we choose it for further modeling analysis. GBR is close to optimal for vaccination counts, and we keep it for consistency for comparing counts and rates.

MLxtend (Raschka, 2018), a sequential variable selection algorithm with GBR method, forward selection option, and "explained_variance" criteria, selects 31 variables from our list of 151 that explain about 60% of the variance in the unvaccinated counts among ZIP codes. For vaccination rates, 31 variables explain about 55%. See Fig. 4A and B.

By limiting variables to ones with a variance inflation factor (VIF) <3.5, we now narrow down the list of variables from 31 to 20 for unvaccinated counts. And for vaccination rates, we are left with 14 variables.

The averaged GBR model predictions (from about 40) for each ZIP code (dotted-red) of COVID-19 unvaccinated counts versus the actual (green) are plotted in Fig. 5A. The average unvaccinated count predictions show the same trend as actual counts. Fig. 5B is for vaccination rates.

Table 2
A and B: Predictor variables and their VIF. Twenty variables were selected from the total 31 for modeling unvaccinated counts to keep VIF less or equal to 3.5. Fourteen were selected for modeling vaccination rates.

Feature Variables(counts): Proportions of	VIF
Males age 60–64	3.1690
Females age 20–24	2.6542
Females age 15–19	2.1112
Females age 10–14	2.5018
Residents in Nursing Homes	1.0508
Whites	2.3752
Households with Annual Income > 200 K	2.6160
Households with Annual Income < 60 K	1.2494
Doctorate level Education	2.0046
Children in Private School	1.6698
Homes Vacant	2.2624
Homes Vacant for Sale	1.6960
Vacant Other Reasons	1.3404
Rented but Unoccupied	1.2711
Homes Vacant for Migrant Workers	1.1313
Part Time Workers	1.2718
Car commuters	2.8178
Public commuters	2.9791
30 – 40 Min to work commuters	1.1858
Homes built before 1930	1.7722
Feature Variables(rates): Proportions of	**VIF**
Females age _85_Plus	3.0167
Residents in Nursing Homes	1.2365
Asians	2.3768
A.I. and Alaskan	1.2849
Hawaiian and Pacific Islanders	1.6505
Homes with Value <25K	1.5487
Homes with Value <150K	2.9728
Homes with Value <200K	3.3837
Homes with Value <750K	3.0288
Professional Education	2.0481
Homes Vacant	1.9450
Public commuters	1.6492
Over 90Min to work commuters	1.9392
Taxi commuters	1.1358

The scatter plot in Fig. 6A between predicted and actual unvaccinated counts further illustrates the strength of the predictive capability of the 20 variables when modeled with the GBR method. Explained variance by our fitted GBR model for unvaccinated counts is $0.854^2 = 0.729$. Fig. 6B is vaccination rates and explained variance by our fitted GBR model using 14 variables is $0.73^2 = 0.5329$.

The average of the importances, when tested on the test sets of the 200 models built on bootstrapped data, is on the left, Fig. 7A. The box plots of the importances from these two hundred (200) models are on the right. Note that the importances only highlight the relative significance of each in keeping the predictions closer to the actual unvaccinated counts and not necessarily increasing or decreasing the counts themselves. Fig. 7B is for feature importances of the 14 variables for vaccination counts.

4. Discussion

Advanced statistical methods can effectively and rapidly identify predictors related to high-risk populations likely to be unvaccinated. This information can support interventions to increase the uptake of COVID vaccines in the US. This study provides an expanded and powerful statistical assessment tool and a methodological format for providing a more comprehensive and locally-focused evaluation of socioeconomic, demographic, and health-related (SEDH) variables associated with those at risk for refusing or delaying use of a COVID-19 vaccination.

We use machine learning methods to understand how SEDH determinants are associated with COVID-19 vaccination rates and unvaccinated individual counts among ZIP codes in California. Our modeling

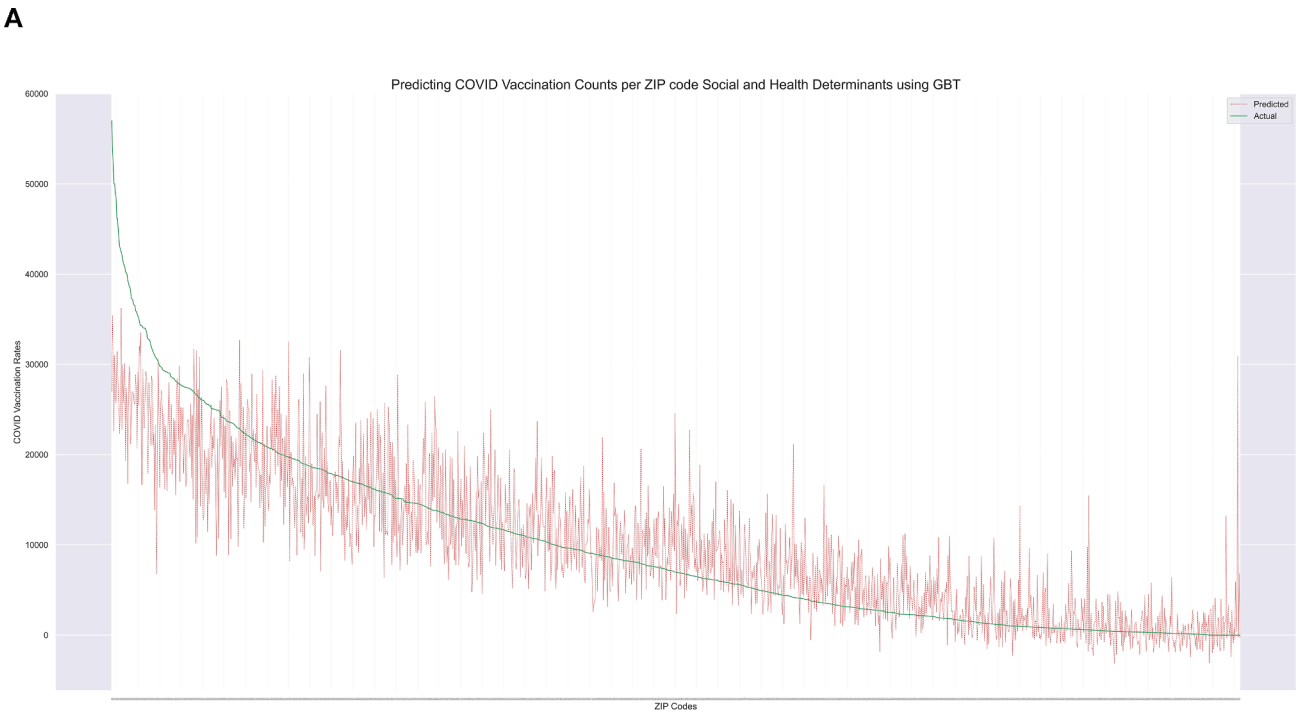

Fig. 5. A and B: Modeling Unvaccinated Counts and Vaccination Rates by ZIP codes in CA. The green lines represent the actual values in each ZIP code, and the dotted red predicted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

explores various algorithms and selects the Gradient Boosting Regressor (GBR) to model vaccination rates and counts. This approach allows us to expand on the 15 commonly used SVI risk measures (Flanagan et al., 2011; CDC1, 2016), to explore 151 possible SEDH explanatory variables using various selection criteria. Consequently, we end up with an optimal list of 20 variables for modeling COVID-19 vaccination counts and 14 variables for vaccination rates. These select sets of explanatory variables, when modeled with GBR, account for over 72% of the variance in vaccination counts among Californian ZIP codes and over 53% of vaccination rates. We also rank the explanatory variables according to their importance in the modeling.

Many recent studies of COVID-19 vaccine rates apply standard or generalized linear regression models to state and national datasets (Mm, 2021; Mollalo and Tatar, 2021). The methodology utilized in this paper is bolstered by machine learning, which via the Gradient Boosting Regressor (GBR), illustrates a higher percentage of the explained variance than linear regression when modeling unvaccinated counts in CA (see Fig. 3) (Scikit-learn: Machine Learning in Python, 2011). This highlights the value of employing this approach to achieve higher levels of the explanatory power of differences in the use and non-use of COVID-19 vaccines on a multi-level population scale (Carmichael and Marron, 2017).

The vaccination rate is a standard metric for analyzing vaccine use amongst a given population. However, vaccination rates over large population groups such as counties lack local specificity due to the significant within-county differences in population characteristics. We address this in two ways: 1) we demonstrated the extreme variability of influential variables across ZIP Codes within counties in California; 2)

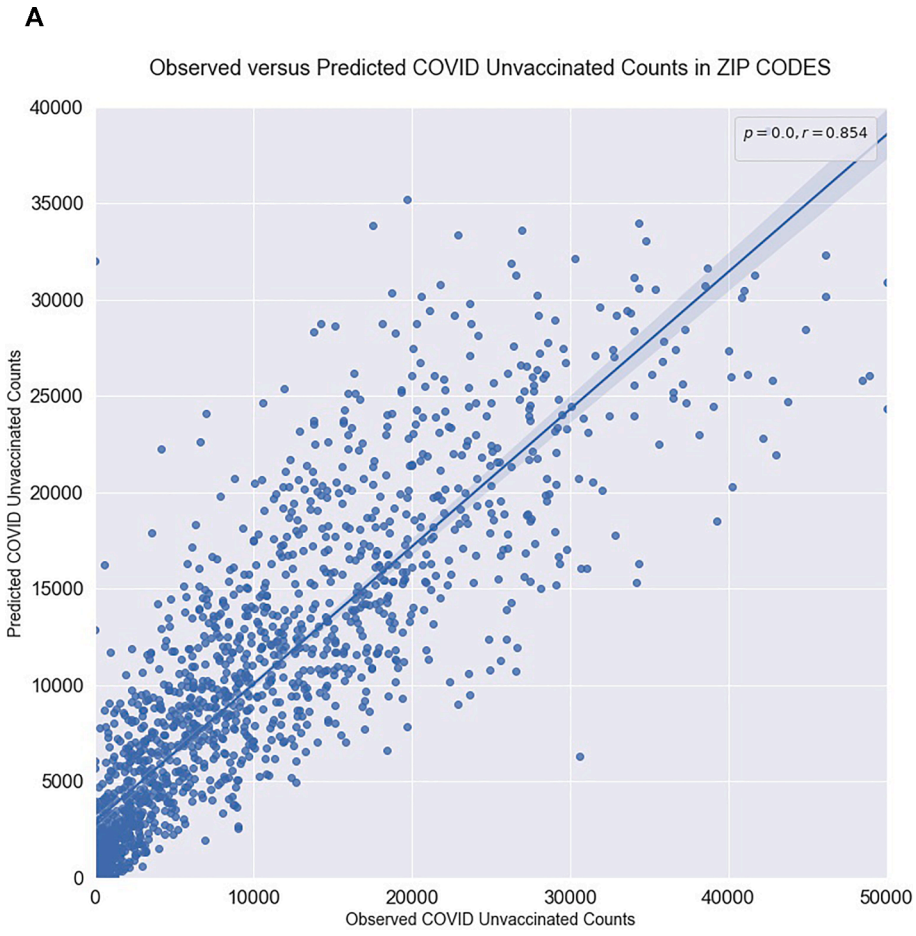

Fig. 6. A and B: Modeling Unvaccinated Counts and Vaccination Rates by ZIP codes in CA. These scatter plots illustrate the correlation between actual values (on the x-axis) and predicted.

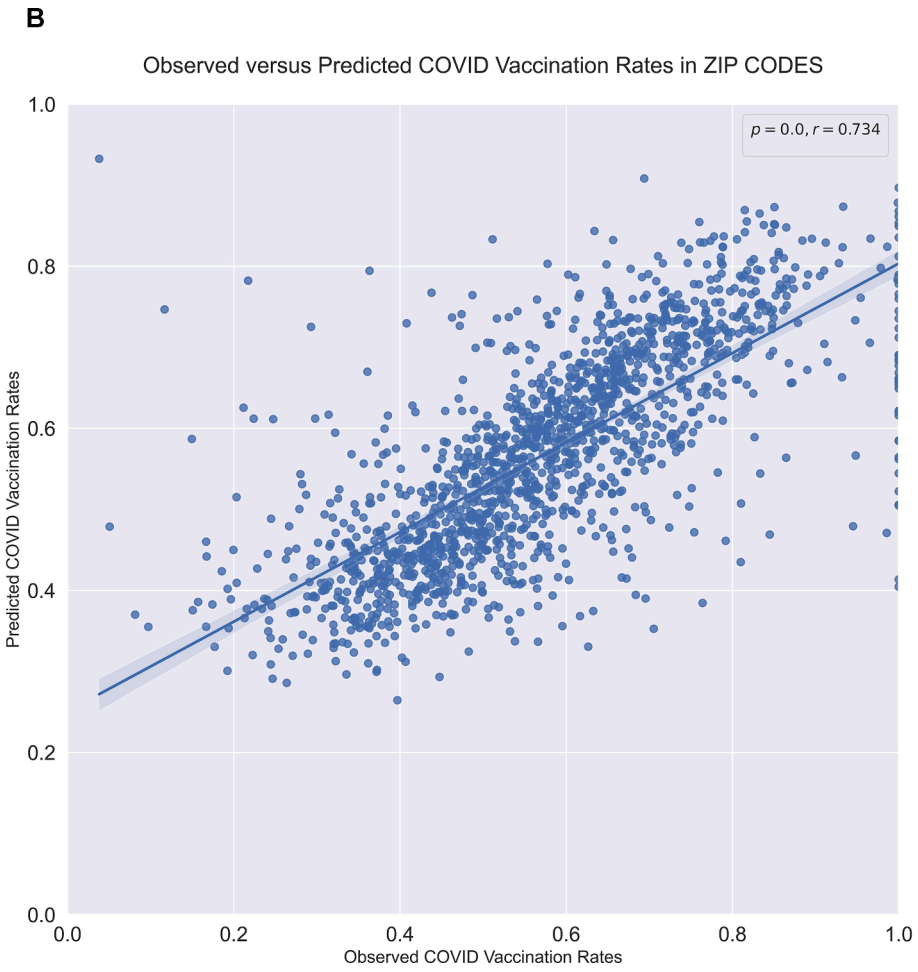

Fig. 6. (*continued*).

we compared unvaccinated counts and vaccination rates for ZIP code-based population groups. When we compared the response variables of vaccination rates and unvaccinated counts with advanced machine learning methods such as the GBR, it showed greater variability in the association between SEDH variables and unvaccinated count differences among ZIP codes (see Fig. 6).

Our study illustrates the strength of utilizing unvaccinated counts as a metric for understanding trends in vaccination prevalence. When applied in conjunction with vaccination rates, within the parameters of ZIP Code levels of analysis, these metrics provide a robust analytical device for determining what populations are most at risk and, most importantly, what ecological factors must be considered to mitigate disparities in vaccination uptake.

4.1. *Further expanding analytic variables*

A range of 20 social and economic influential factors was identified that explain vaccination counts (see Table 2A). These variables mirror the commonly assessed Social Vulnerability Index individual variables. Yet, there are notable differences as we were able to assess a much larger number of social and health-related factors not considered in the SVI. These findings show that the identified influential variables overlap, complement, and extend the SVI measures. This study meets the well-documented observation of the need to assess an extensive and holistic range of influential variables to understand and address social determinants of health and risk behavior (Carmichael and Marron, 2017).

4.2. *Limitations and future work*

This study may be limited by utilizing data from CA alone based on differences in environmental factors that may not arise across all states and on a national scale. The study also examines COVID-19 pandemic responses at one point in time, although COVID-19 is a dynamic pandemic and coronavirus, a highly mutable disease pathogen. However, the findings of this study provide a framework for optimized modeling strategies, which can be employed to understand the risk for non-participation in COVID-19 vaccination programs at multiple population levels and at different time points. A key purpose of this study is to provide a methodology that can identify local at-risk populations and their social demographic and health-related behavioral and contextual variables associated with delaying and remaining unvaccinated (Raschka, 2021). This approach can assist in targeting public health interventions to the most at-risk groups (Finney Rutten et al.). The next step to alleviate disparities in vaccination coverage is to apply this methodology to identify specific sociodemographic and geo-local populations and formulate meaningful policies and programs to support the use of a COVID-19 vaccine (Barry et al., 2021). The overarching goal of this effort is to provide a powerful and adaptable analytical tool to identify at-risk populations in terms of a larger range and more specific set of covariates than has been typically used to identify those most at risk (Wang et al.; Williams et al.).

4.3. *Conclusion*

Our machine learning model can consolidate prominent predictor

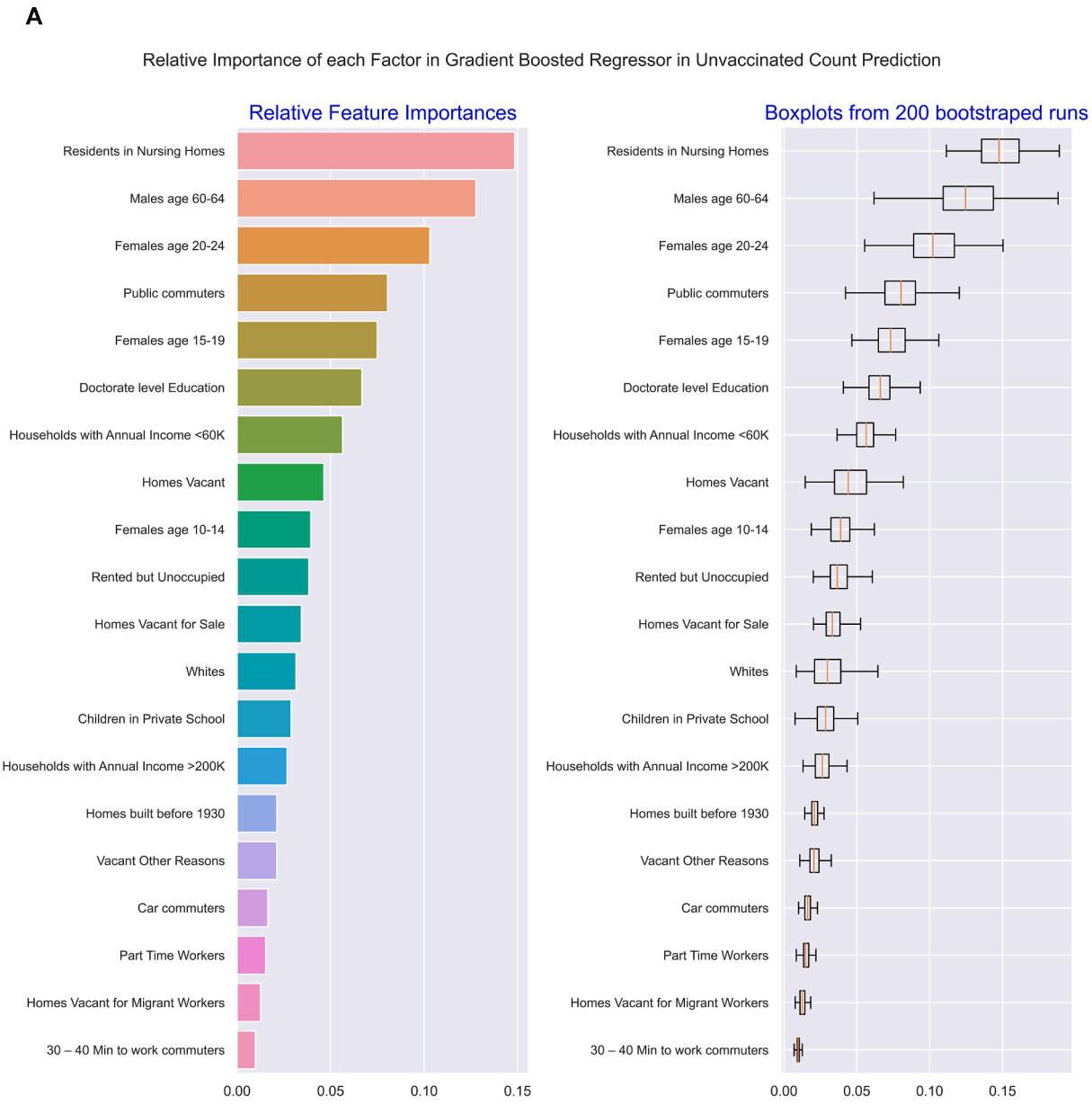

Fig. 7. A and B: Feature Importances in predicting Unvaccinated Counts and Vaccination Rates using GBR in CA ZIP codes. The importances give us a sense of each variable's contribution in bringing the prediction as close as possible to the actual values.

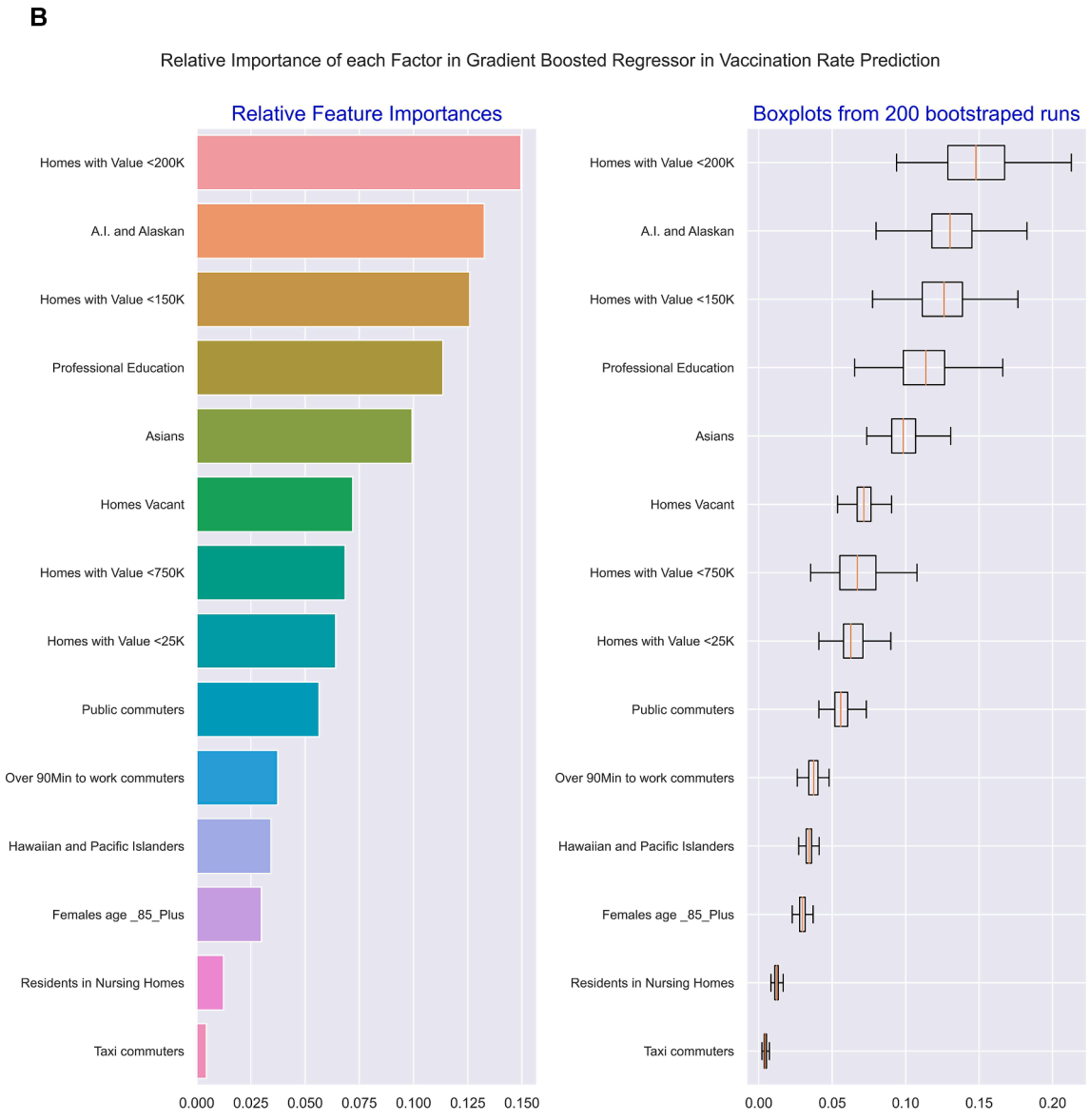

Fig. 7. (*continued*).

variables into subcomponents and expand on the predictor variables commonly used to assess risk factors associated with refusal and delay in accepting a COVID-19 vaccination. The GBR model has the potential to create a vaccination-specific index for future outbreaks and pandemics. This demonstration of the public health utility of applying machine learning methodology to current public health conditions provides a means for expanding and adapting this approach to analyzing an array of future public health concerns.

Funding

This work was partially supported by Kean University through Release Time for Research award for Dr. George Avirappattu.

CRediT authorship contribution statement

George Avirappattu: Conceptualization, Data curation, Formal analysis, Methodology, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Alfred Pach III:** Conceptualization, Investigation, Resources, Writing – review & editing. **Clarence E. Locklear:** Resources, Writing – review & editing. **Anthony Q. Briggs:** Resources Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Appendix

A. Complete predictor variable List

Social and Structural variables:

M_Under_5_prop, M_5-9_prop, M_10-14_prop, M_15-19_prop, M_20-24_prop, M_25-29_prop, M_30-34_prop, M_35-39_prop, M_40-44_prop, M_45-49_prop, M_50-54_prop, M_55-59_prop, M_60-64_prop, M_65-69_prop, M_70-74_prop, M_75-79_prop, M_80-84_prop, M_85_Plus_prop,

F_Under_5_prop, F_5-9_prop, F_10-14_prop, F_15-19_prop, F_20-24_prop, F_25-29_prop, F_30-34_prop, F_35-39_prop, F_40-44_prop, F_45-49_prop, F_50-54_prop, F_55-59_prop, F_60-64_prop, F_65-69_prop, F_70-74_prop, F_75-79_prop, F_80-84_prop, F_85_Plus_prop,

School_Public_prop, School_Private_prop, School_None_prop,

Educ_<HS_prop, Educ_HS_prop, Educ_AA_prop, Educ_BA_prop, Educ_Msters_prop, Educ_Prof_prop, Educ_Doctorate_prop,

<10Min_to_work_prop, <20Min_to_work_prop, <30Min_to_work_prop, <40Min_to_work_prop,

<50Min_to_work_prop, <60Min_to_work_prop, <90Min_to_work_prop, >90Min_to_work_prop,

car_commute_prop, public_commute_prop, taxi_commute_prop, motorcycle_commute_prop, walk_commute_prop,

work_at_home_prop,

<25K_HHI_prop, <45K_HHI_prop, <60K_HHI_prop, <100K_HHI_prop, <150K_HHI_prop,

<200K_HHI_prop, >200K_HHI_prop,

Work_FT_prop, Work_PT_prop, Work_None_prop,

Studio_Count_prop, 1Bdrm_Count_prop, 2Bdrm_Count_prop, 3Bdrm_Count_prop,

<25K_Home_Value_prop, <50K_Home_Value_prop, <100k_Home_Value_prop, <150K_Home_Value_prop,

<200K_Home_Value_prop, <400K_Home_Value_prop, <750K_Home_Value_prop, >750K_Home_Value_prop,

Vacant_For_Rent_prop, Rented_Unoccupied_prop, Vacant_For_Sale_prop, Vacant_Sold_Unoccupied_prop,

Vacant_Recreational_Occasional_prop, Vacant_For_Migrant_Workers_prop, Vacant_Other_Reasons_prop,

<1930_built_prop, 1940_built_prop, 1950_built_prop, 1960_built_prop, 1970_built_prop, 1980_built_prop,

1990_built_prop, 2000_built_prop, 2010_built_prop,

Home_Owned_Mortgaged_prop, Home_Owned_prop, Home_Rented_prop, Home_Vacant_prop,

In Occupied Housing Units_prop, Correctional_prop, Juvenile_prop, Nursing_Homes_prop, Institutional_prop,

College_prop, Military_prop, Noninstitutional_prop, Husband_Wife_Family_prop,

Single_Parent_Family_prop, Single_Family_prop, Single_w_roommate_Family_prop,

Whites_count_prop, AA_count_prop, AI_Alaskan_count_prop, Asian_count_prop, Hawaiian_PI_count_prop,

Others_count_prop, Multi_count_prop,

Median_home_value, Median_household_income.

Health-related variables:

All teeth lost among adults aged >=65 years.

Arthritis among adults aged >=18 years.

Binge drinking among adults aged >=18 years.

Cancer (excluding skin cancer) among adults aged >=18 years.

Cervical cancer screening among adult women aged 21–65 years.

Cholesterol screening among adults aged >=18 years.

Chronic kidney disease among adults aged >=18 years.

Chronic obstructive pulmonary disease among adults aged >=18 years.

Coronary heart disease among adults aged >=18 years.

Current asthma among adults aged >=18 years.

Current lack of health insurance among adults aged 18–64 years.

Current smoking among adults aged >=18 years.

Diagnosed diabetes among adults aged >=18 years.

Fecal occult blood test; sigmoidoscopy; or colonoscopy among adults aged 50–75 years.

High blood pressure among adults aged >=18 years.

High cholesterol among adults aged >=18 years who have been screened in the past 5 years.

Mammography use among women aged 50–74 years.

Mental health not good for >=14 days among adults aged >=18 years.

No leisure-time physical activity among adults aged >=18 years.

Obesity among adults aged >=18 years.

Older adult men aged >=65 years who are up to date on a core set of clinical preventive services:

Flu shot past year.

PPV shot ever.

Colorectal cancer screening.

Older adult women aged >=65 years who are up to date on a core set of clinical preventive services:

Flu shot past year.

PPV shot ever.

Colorectal cancer screening or Mammogram past 2 years.

Physical health not good for >=14 days among adults aged >=18 years.

Sleeping<7 h among adults aged >=18 years.

Stroke among adults aged >=18 years.

Taking medicine for high blood pressure control among adults aged >=18 years with high blood pressure.

Visits to dentist or dental clinic among adults aged >=18 years.

Visits to doctor for routine checkup within the past year among adults aged >=18 years.

B. Social Vulnerability Factors and their Domains

Factor	SVI Domains 1–4
Below Poverty	Socioeconomic Status
Unemployed	
Income	
No High School Diploma	
Age 65+	Household Composition & Disability
Age 17 under	
Older than 5 with disability	
Single-Parent Household	
Minority	Minority Status and Language
Speak English "Less than Well"	
Multiunit Structures	Housing & Transportation
Mobile Homes	
Crowding	
No Vehicle	
Group Quarters	

Source: Flanagan et al (2011), A Social Vulnerability Index for Disaster Management.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2022.101858.

References

Al-Mohaithef, M., Padhi, B.K., Ennaceur, S. Socio-Demographics Correlate of COVID-19 Vaccine Hesitancy During the Second Wave of COVID-19 Pandemic: A Cross-Sectional Web-Based Survey in Saudi Arabia. (2296-2565 (Electronic)).

Bureau USC. American Community Survey, 2019 American Community Survey 5-Year Estimates. US Census Bureau. Accessed October 19, 2020. https://data.census.gov/mdat/#/search?ds=ACSPUMS5Y2019.

Aw, J., Seng, J.J.B., Seah, S.S.Y., Low, L.L., 2021. COVID-19 Vaccine Hesitancy—A Scoping Review of Literature in High-Income Countries. Vaccines 9 (8), 900.

Barry V Fau - Dasgupta S, Dasgupta S Fau - Weller DL, Weller Dl Fau - Kriss JL, et al. Patterns in COVID-19 Vaccination Coverage, by Social Vulnerability and Urbanicity - United States, December 14, 2020-May 1, 2021. (1545-861X (Electronic)).

BRFSS. PLACES: Local Data for Better Health, ZCTA Data 2020 release. https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-ZCTA-Data-2020/qnzd-25i4.

Bruckhaus, A.A., Abedi, A., Salehi, S., et al., 2021. COVID-19 Vaccination Dynamics in the US: Coverage Velocity and Carrying Capacity Based on Sociodemographic Vulnerability Indices in California, 2021/11/19 J. Immig. Minority Health. https://doi.org/10.1007/s10903-021-01308-2.

Bureau UC. American Community Survey. US Census Bureau. Accessed April 29, 2022. https://www2.census.gov/programs-surveys/acs/.

Carmichael, I., Marron, J.S., 2017. Data Science vs. Statistics: Two Cultures? arXiv: 1801.00371. Accessed December 01, 2017. https://ui.adsabs.harvard.edu/abs/2018arXiv180100371C.

CDC, 2022. COVID Data Tracker Weekly Review. Accessed January 11. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html#more-info.

CDC1. Centers for Disease Control and Prevention. CDC SVI 2016 Documentation. 2019. Accessed 7/31/2020. https://svi.cdc.gov/Documents/Data/2016_SVI_Data/SVI2016Documentation.pdf.

Finney Rutten, L.J., Zhu, X., Leppin, A.L., et al. Evidence-Based Strategies for Clinical Organizations to Address COVID-19 Vaccine Hesitancy. (1942-5546 (Electronic)).

Flanagan, B.E., Gregory, E.W., Hallisey, E.J., Heitgerd, J.L., Lewis, B., 2011. A Social Vulnerability Index for Disaster Management. J. Homeland Security Emerg. Manag. 8(1). doi:10.2202/1547-7355.1792.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Health CDoP. Data from: CA COVID-19 Vaccine Data by ZIP Code. 2021. https://covid19.ca.gov/vaccines/. Deposited April 21, 2021.

Karaye, I.M., Horney, J.A., Sep 2020. The Impact of Social Vulnerability on COVID-19 in the US: An Analysis of Spatially Varying Relationships. Am. J. Prev. Med. 59 (3), 317–325. https://doi.org/10.1016/j.amepre.2020.06.006.

Kearney, G., Jones, K., Park, Y.M., et al., 2021. COVID-19: A Vaccine Priority Index Mapping Tool for Rapidly Assessing Priority Populations in North Carolina. Online J. Public Health Informat. 12/24 13(3)doi:10.5210/ojphi.v13i3.11617.

Kricorian, K., Turner, K., 2021. COVID-19 Vaccine Acceptance and Beliefs among Black and Hispanic Americans. PLOS ONE 16 (8), e0256122. https://doi.org/10.1371/journal.pone.0256122.

Lindemer, E., Choudhary, M., Donadio, G., Pawlowski, C., Soundararajan, V., 2021. Counties with lower insurance coverage are associated with both slower vaccine rollout and higher COVID-19 incidence across the United States. medRxiv. 2021.03.24.21254270. doi:10.1101/2021.03.24.21254270.

MacDonald, N.E. Vaccine hesitancy: Definition, scope and determinants. (1873-2518 (Electronic)).

MM, H., A, W., MK, G., et al., 2021. County-Level COVID-19 Vaccination Coverage and Social Vulnerability — United States, December 14, 2020–March 1, 2021. MMWR Morbid. Mortal. Weekly Rep. Rep 2021(70):431-436.

Mollalo, A., Tatar, M., 2021. Spatial Modeling of COVID-19 Vaccine Hesitancy in the United States. Int. J. Environ. Res. Public Health 18 (18), 9488.

Raschka, S., 2018. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J. Open Sour. Softw. 24. Open J.; 2018;3. doi:10.21105 /joss. 00638 Accessed 7/2/2021. http://rasbt.github.io/mlxtend/.

Raschka, S., 2021. What are the main differences between statistical modeling and machine learning? Accessed July 15, 2021. https://sebastianraschka.com/faq/docs/statistical-modeling-vs-machine-learning.html.

Romano A. US sets new COVID hospitalization record, signaling Omicron surge could be less mild than experts hoped. Accessed January 11, 2022. https://news.yahoo.com/us-sets-new-covid-hospitalization-record-signaling-omicron-surge-could-be-less-mild-than-experts-hoped-100032439.html.

S A. US has 4% of world population but 25% of its corona virus cases. CNN. Accessed June 30, 2020. https://www.cnn.com/2020/06/30/health/us-coronavirus-toll-in-numbers-june-trnd/index.html.

Sallam, M., 221. COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates. Vaccines 9(2). doi:10.3390/vaccines9020160.

Scikit-learn: Machine Learning in Python. 2011. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Soares, P., Rocha, J.V., Moniz, M., et al., 2021. Factors Associated with COVID-19 Vaccine Hesitancy. Vaccines 9 (3), 300. https://doi.org/10.3390/vaccines9030300.

Breiman, L., 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat. Sci. 08;16(3), 199–231.

uszipcode, 2020. Version 1.0.1. GitHub; 2020. Accessed October 19, 2020. https://uszipcode.readthedocs.io/index.html.

Wang, S.X., Bell-Rogers, N., Dillard, D., Harrington, M.A. COVID-19 Vaccine Hesitancy in Delaware's Underserved Communities. (2639-6378 (Electronic)).

Williams, A.M., Clayton, H.B., Singleton, J.A. Racial and Ethnic Disparities in COVID-19 Vaccination Coverage: The Contribution of Socioeconomic and Demographic Factors. LID - S0749-3797(21)00565-1 [pii] LID - 10.1016/j.amepre.2021.10.008 [doi]. (1873-2607 (Electronic)).

Woolf, K., McManus, I.C., Martin, C.A., et al., 2021. Ethnic differences in SARS-CoV-2 vaccine hesitancy in United Kingdom healthcare workers: Results from the UK-REACH prospective nationwide cohort study. Lancet Regional Health - Europe. 2021/10/01/, 9:100180. doi:https://doi.org/10.1016/j.lanepe.2021.100180.