



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Using machine learning to develop a novel COVID-19 Vulnerability Index (C19VI)

Anuj Tiwari ^{a,*}, Arya V. Dadhania ^b, Vijay Avin Balaji Ragnathrao ^c, Edson R.A. Oliveira ^d

^a Discovery Partners Institute, UIC, Chicago, IL, USA

^b College of Medicine, UIC, Chicago, IL, USA

^c Department of Pharmacology, UIC, Chicago, IL, USA

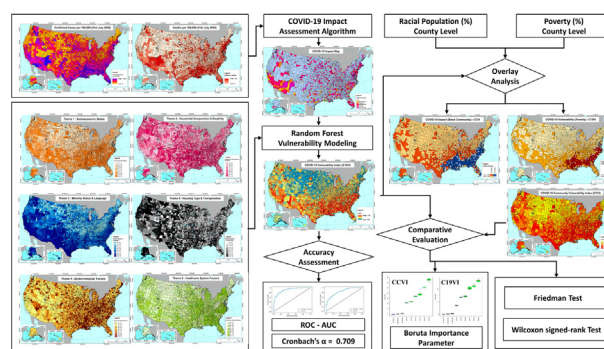
^d Department of Microbiology and Immunology, UIC, Chicago, IL, USA



HIGHLIGHTS

- C19VI was developed using machine learning on CDC socio-demographic and COVID-19 themes.
- 'COVID-19 Impact Assessment' algorithm was developed using homogeneity and trend assessment.
- C19VI provides an accurate and reliable way to measure county level vulnerabilities in the US.
- Racial minorities and economically poor are disproportionately impacted by the pandemic.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 29 September 2020

Received in revised form 30 January 2021

Accepted 31 January 2021

Available online 5 February 2021

Editor: Jay Gan

Keywords:

COVID-19

Vulnerability modeling

Machine learning

Racial minority

Disproportionate COVID-19

ABSTRACT

COVID-19 is now one of the most leading causes of death in the United States (US). Systemic health, social and economic disparities have put the minorities and economically poor communities at a higher risk than others. There is an immediate requirement to develop a reliable measure of county-level vulnerabilities that can capture the heterogeneity of vulnerable communities. This study reports a COVID-19 Vulnerability Index (C19VI) for identifying and mapping vulnerable counties. We proposed a Random Forest machine learning-based vulnerability model using CDC's sociodemographic and COVID-19-specific themes. An innovative 'COVID-19 Impact Assessment' algorithm was also developed for evaluating severity of the pandemic and to train the vulnerability model. Developed C19VI was statistically validated and compared with the CDC COVID-19 Community Vulnerability Index (CCVI). Finally, using C19VI and the census data, we explored racial inequalities and economic disparities in COVID-19 health outcomes. Our index indicates that 575 counties (45 million people) fall into the 'very high' vulnerability class, 765 counties (66 million people) in the 'high' vulnerability class, and 1435 counties (204 million people) in the 'moderate' or 'low' vulnerability class. Only 367 counties (20 million people) were found as 'very low' vulnerable areas. Furthermore, C19VI reveals that 524 counties with a racial minority population higher than 13% and 420 counties with poverty higher than 20% are in the 'very high' or 'high' vulnerability classes. The C19VI aims at helping public health officials and disaster management agencies to develop effective mitigation strategies especially for the disproportionately impacted communities.

© 2021 Elsevier B.V. All rights reserved.

* Corresponding author.

E-mail address: anujt@uic.edu (A. Tiwari).

1. Introduction

The novel coronavirus disease (COVID-19) has been recognized as the newest and biggest global public health crisis (Organization WH, 2020). In the first half of 2020, the COVID-19 has nearly killed half a million people worldwide, of which more than 25% have died in the United States of America (COVID C, 2020). Within three months after the first reported case in the United States in January 2020, coronavirus cases had been confirmed in all fifty states, including the District of Columbia and other inhabited US territories (COVID C, 2020). Shortly after, many state-imposed policies to curb the spread (Agency. FEM, 2020; Zhao et al., 2020). Despite these diffused efforts, a massive surge in the incidence and mortality began to recur in June 2020 that placed the US as the most affected country setting apart from other countries by a huge margin (COVID C, 2020; Oster et al., 2020). This has clearly indicated a lapse in effective COVID risk assessment and response at different levels. With the added concern of the second wave and disproportionate impact of the pandemic on minorities and economically poor, a reliable country-wide assessment of COVID-19 vulnerability is a matter of necessity and urgency (COVID I and Murray, 2020).

Identification of vulnerable areas is critical for public health departments to take the appropriate measures to increase preparedness against COVID-19. In order to identify these areas, the Centers for Disease Control and Prevention (CDC) initially used the social vulnerability index (SVI) (Flanagan et al., 2011), which is calculated based on census variables distributed in four distinct themes: i) socio-economic factors, ii) household composition, iii) minority status and language, and iv) access to housing and transportation. SVI failed to sufficiently determine vulnerability during this unprecedented situation, mainly because its primary objective aimed at addressing natural disaster crises during hurricanes, earthquakes, and forest fires (Karaye and Horney, 2020; Amram et al., 2020). Thus, the CDC and Surgo Foundation developed COVID-19 Community Vulnerability Index (CCVI) (Foundation S, 2020) by introducing two new variables, v) epidemiological risk factors and vi) public health system capacity, in the hope to rectify the shortcomings in the previous vulnerability assessment approach. Despite this optimization, CCVI and SVI are based on a statistical linear algorithm (Flanagan et al., 2011; Foundation S, 2020) that is unable to sufficiently account for the multiplicative, non-linear nature of vulnerability (Sambanis et al., 2019). Not only have the public health planners (Karaye and Horney, 2020; Kim and Bostwick, 2020; Sequist, 2020) and policy makers (Tai et al., 2020; Liu et al., 2020; Acharya and Porwal, 2020) recognized the need for a more nuanced methodology in the domain of vulnerability modeling, but they are also concerned about the highly dynamic nature of the pandemic that brings unique challenges to the existing methods of pandemic vulnerability modeling. The lack of a comprehensive and accurate COVID-19 Vulnerability Index impairs the preparedness of critical areas against the pandemic as they are blurred to public health measures. This scenario highlights the urgent need for improvements of the nationwide approaches to identify vulnerable areas amid COVID-19 pandemic.

Pandemic vulnerability modeling techniques thus far, including that of CDC's CCVI, only analyze some of the variables that introduce the variability, such as the COVID-19 impact (theme 5: epidemiological factors) and the counties' preparation and resources against the pandemic (theme 6: healthcare system factors) and, they do so in a linear statistical fashion. In the current study, we developed a more reliable assessment: the COVID-19 Vulnerability Index (C19VI) which quantifies the pandemic vulnerability of each county in the United States. This relative index processed the same six input variables as CCVI; however, instead of using a statistical linear algorithm, we utilized machine learning technique. We implemented Random Forest (RF) machine learning technique to calculate C19VI. An innovative 'COVID-19 Impact Assessment' algorithm was also developed using homogeneity analysis and temporal trend assessment techniques for training the RF model. Our 'COVID-19 Impact Assessment' algorithm, for the first time, introduce

the concept of analyzing temporal dynamics of confirmed cases, deaths and IFR in addition to analyzing the CDC's six themes in a non-parametric, non-linear machine learning-integrated method. Thus, our vulnerability modeling approach has a two-fold added advantage than the conventional methods. First, we assessed the additional variables that introduce variability in vulnerability modeling, i.e., temporal analysis of daily confirmed cases, deaths, and IFR data. Secondly, all of the variables were processed in a non-linear, non-parametric fashion by using RF machine learning techniques. Next, our C19VI index was compared with CDC's CCVI using advanced statistical measures and a machine learning model. We then tested the accuracy and checked the internal consistency of the C19VI.

Our vulnerability assessment methodology has allowed us to analyze the impact of COVID-19 that has been unequal and widespread across the nation (Tai et al., 2020; Moore, 2020; Dang et al., 2020; Finch and Hernández Finch, 2020). Besides, there are systemic socioeconomic inequalities that increase the susceptibility and exposure of the marginalized groups (Moore, 2020; Finch and Hernández Finch, 2020; Ahmed et al., 2020). Thus, in addition to conducting a nationwide analysis of COVID-19 vulnerability, C19VI has allowed us, for the first time, to explore the existing healthcare disparities in the realm of COVID-19 pandemic in great detail. This study may enhance the current techniques in vulnerability modeling, leveraging the preparedness of vulnerable counties to reduce the COVID-19 burden within the United States.

2. Data and methods

2.1. Input datasets

We used publicly available datasets from Johns Hopkins University (COVID C, 2020), Centers for Disease Control and Prevention (CDC) (Foundation S, 2020), United States Census Bureau (Bureau UC, 2018), and United States Department of Homeland Security (DHS, 2016) for impact assessment, vulnerability modeling, population-specific vulnerability analysis, and data visualization and mapping, respectively. The data for COVID-19 confirmed cases, including all reported infections and reinfections, and mortality in the United States from 22nd January 2020 to 31st July 2020 were obtained from Johns Hopkins University. Fig. 1 (A) and (B) presents the normalized (per 100,000) total confirmed cases and deaths dataset, up to 31st July 2020 for all United States counties, respectively. The four socio-demographic SVI indicators and two CCVI thematic indicators, referred to as input themes were obtained from CDC. This results in six input themes that determine the vulnerability of a region to COVID-19 pandemic (Table 1). Fig. 2 shows (A) Socioeconomic Status, (B) Household Composition & Disability, (C) Minority Status & Language, (D) Housing Type & Transportation, (E) Epidemiological Factors and (F) Healthcare System Factors maps for the United States. County wise total population, racial population and poverty breakdown were obtained from the United States Census Bureau. Regional boundary data of the United States was collected from Homeland infrastructure foundation-level data (DHS, 2016) in Geographic Information System (GIS) ready file format (ESRI shapefile (ESRI E, 1998)). This study did not require a review by the Institutional Review Board since publicly available, de-identified data was used.

2.2. 'COVID-19 Impact Assessment' algorithm

In order to understand the impact of COVID-19 pandemic in all 3142 counties in the United States, we have proposed a 'COVID-19 Impact Assessment' algorithm. This algorithm 'Scores' and 'Ranks' the impact of COVID-19 pandemic by evaluating the temporal changes in confirmed cases, deaths, and infection fatality rate (IFR) (Magnani et al., 2020) datasets using trend analysis (Mann Kendall (Mann, 1945; Kendall, 1955) & Theil and Sen Slope (Theil, 1950; Sen, 1968)) and homogeneity assessment (Pettitt's test (Pettitt, 1979)). Trend analysis characterizes

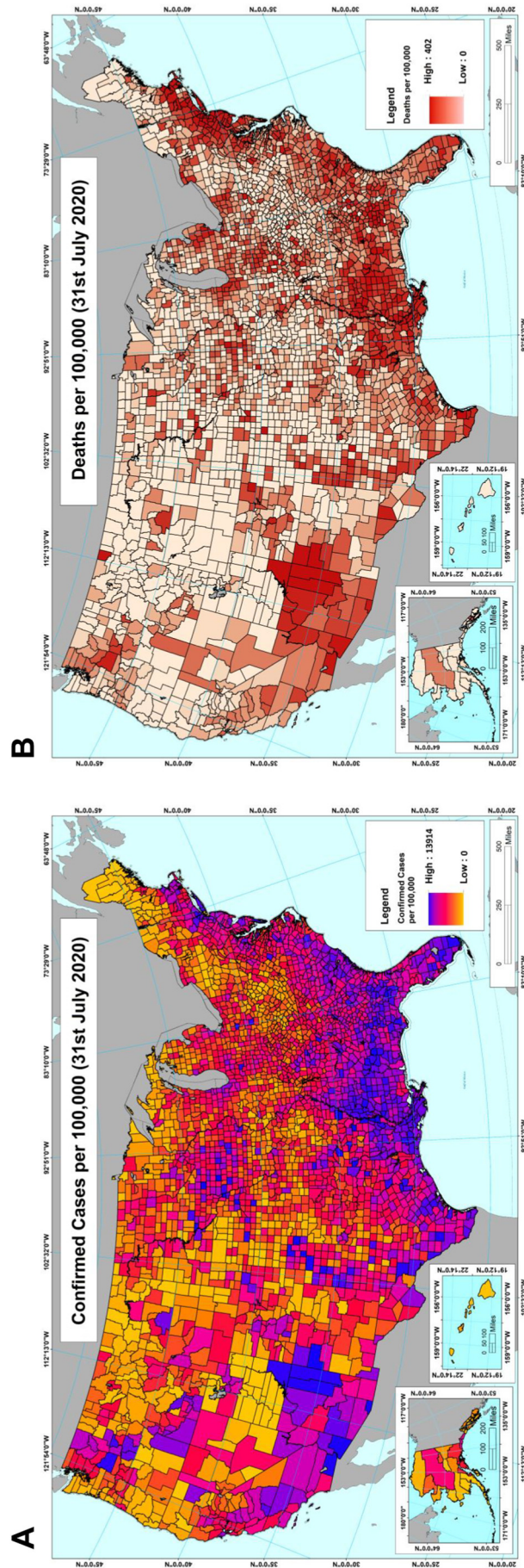


Fig. 1. Maps of the US counties representing confirmed COVID-19 cases and deaths normalized (per 100,000 people). (A) County-wise confirmed cases in the US up to 31st July 2020. (B) County-wise deaths in the US up to 31st July 2020.

Table 1
The CDC's CCVI theme indicators and corresponding variables.

Theme	Indicator	Type	Variable	Resolution
1	Socioeconomic Status	Social	Below poverty	Census tract
			Unemployed	Census tract
			Income	Census tract
			No high school diploma	Census tract
2	Household Composition & Disability	Social	Aged 65 or older	Census tract
			Aged 17 or younger	Census tract
			Older than age 5 with a disability	Census tract
			Minority	Census tract
3	Minority Status & Language	Social	Speaks English "less than well"	Census tract
			Multi-unit structures	Census tract
4	Housing Type & Transportation	Social	Mobile homes	Census tract
			Crowding	Census tract
			No vehicle	Census tract
			Cardiovascular conditions	County
5	Epidemiological Factors	COVID	Respiratory conditions	County
			Immuno-compromised	County
			Obesity	County
			Diabetes	County
			Population density	Census tract
			Influenza and pneumonia death rates	County
			Health system capacity	State/hospital region
6	Healthcare System Factors	COVID	Health system strength	State/county
			Health system preparedness	State/county

the overall pattern in daily-time series dataset and homogeneity assessment identifies abrupt changes in temporal trends (Mann, 1945; Kendall, 1955; Theil, 1950; Sen, 1968; Pettitt, 1979). Together, trend and homogeneity analyses make the algorithm more sensitive to daily changes in the epidemiological curve and recognize the subtle impacts of the health policies. Thus, the algorithm classifies each county in one of the six impact groups, 'very high' (Rank = 1), 'high' (Rank = 2), 'moderate' (Rank = 3), 'low' (Rank = 4), 'very low' (Rank = 5) and 'non-significant' (Rank = -999). See Supplementary material for the 'COVID-19 Impact Assessment' algorithm pseudocode. The algorithm functions in four steps:

1. Data import and pre-processing: County-wise, daily time-series data of the confirmed cases and deaths were obtained from the John Hopkins University as mentioned above (COVID C, 2020). Then, daily time-series data for IFR is calculated using the imported datasets.
2. Homogeneity analysis: Pettitt's test (Pettitt, 1979) was applied county-wise to check for the homogeneity in the time-series dataset of all three epidemiological parameters obtained after step 1. If the data was found to be non-homogeneous, pre and post-change point time series were computed and kept alongside the 'overall' dataset, which was the only populated data column in the cases of homogeneous datasets. This expanded the time-series dataset into three aspects, i.e., pre-change point, post-change point, and overall, for each of the three epidemiological parameters, i.e., confirmed cases, deaths, and IFR for each county.
3. Trend analysis: We applied Mann Kendall's test (Mann, 1945; Kendall, 1955) to assess the trend and its nature, i.e. increasing, decreasing, or no trend, in a given time-series. Next, the trend magnitude was quantified using the Theil and Sen slope estimator test (Theil, 1950; Sen, 1968). Mann Kendall's, and Theil and Sen slope estimator test was performed on all three time-series computed at the end of step 2 for all three epidemiological parameters in each county.
4. COVID-19 Impact 'Score' and 'Rank' determination: Impact Score was determined using the trend magnitude data obtained from the previous step. We used IFR as the most important parameter for assessing the impact of the COVID-19 pandemic in our algorithm (Magnani

et al., 2020; Meehan et al., 2020). In the instances where IFR did not show a significant trend in a given county, we first used the deaths (Meehan et al., 2020). If the deaths did not show a significant trend either, confirmed cases were used to evaluate the impact of the pandemic (Meehan et al., 2020). Thus, rank classification occurred in three stages, each further divided according to the homogeneity results:

- a. On the basis of the IFR:
 - i. In a homogeneous IFR time-series with an increasing 'overall' trend, the county was assigned Rank 1 and its impact Score was equal to the 'overall' trend magnitude.
 - ii. In a non-homogeneous IFR time-series with an increasing pre-change point trend, the scoring and ranking were specified based on the post-change point trend. Counties with increasing post-change point trends were classified as Rank 1, no post-change point trends as Rank 3, and decreasing post-change point trends as Rank 5. The Score of the counties with increasing (Rank 1) and no (Rank 3) post-change point trends were equal to the trend magnitude of 'post' and 'pre' time-series data, respectively. Finally, Score of the counties with decreasing post-change point data (Rank 5) was equal to the negative of the 'post' time-series trend magnitude.
- b. On the basis of the fatalities (deaths):
 - i. In a homogeneous death time-series with an increasing 'overall' trend, the county was assigned Rank 2 and its impact Score was equal to the 'overall' trend magnitude.
 - ii. In a non-homogeneous death time-series with an increasing pre-change point trend, the scoring and ranking were specified based on the post-change point trend. Counties with increasing post-change point trends were classified as Rank 2, no post-change point trends as Rank 3, and decreasing post-change point trends as Rank 5. The Score of the counties with increasing (Rank 2) and no (Rank 3) post-change point trends were equal to the trend magnitude of 'post' and 'pre' time-series data, respectively. Finally, Score of the counties with decreasing post-change point data (Rank 5) was equal to the negative of the 'post' time-series trend magnitude.
- c. On the basis of the confirmed cases:
 - i. In a homogeneous time-series of confirmed cases with an increasing 'overall' trend, the county was assigned Rank 4 and its impact Score was equal to the 'overall' trend magnitude.
 - ii. In a non-homogeneous time-series of confirmed cases with an increasing pre-change point trend, the scoring and ranking were specified based on the post-change point trend. Counties with increasing post-change point trends were classified as Rank 4, no post-change point trends as Rank 5, and decreasing post-change point trends as Rank 5. The Score of the counties with increasing (Rank 4) and no (Rank 5) post-change point trends were equal to the trend magnitude of 'post' and 'pre' time-series data, respectively. Finally, Score of the counties with decreasing post-change point data (Rank 5) was equal to the negative of the post time-series trend magnitude.

Every other county was classified as Rank -999 and Score -999. Finally, out of the three ranks, assigned to each county, based on the three epidemiological variables, the highest impact group (lowest rank) and its corresponding trend magnitude were decided as the final COVID-19 Impact Score and Rank for a given county.

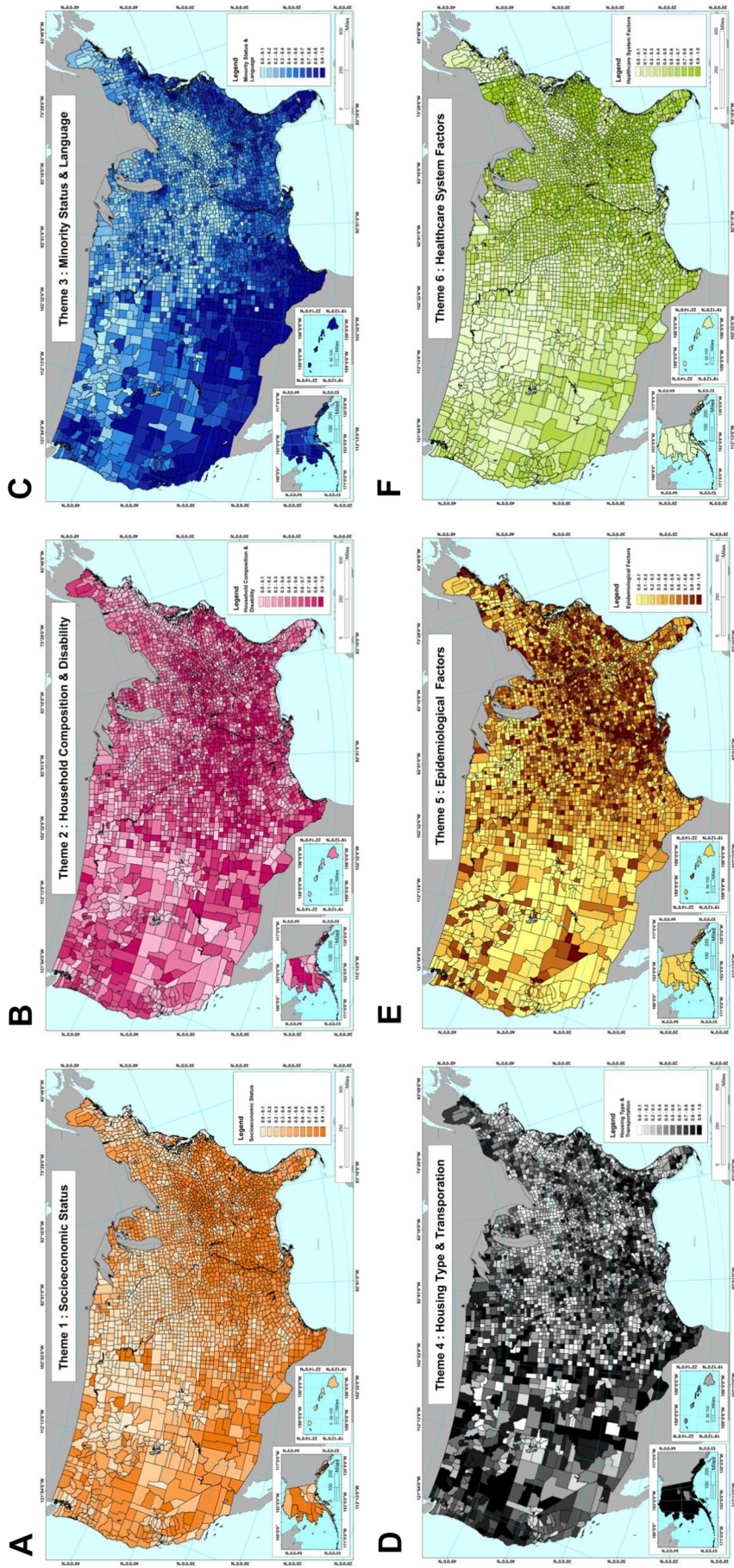


Fig. 2. Maps of the US counties representing the CDC's six COVID-19-specific input themes. Panels A–F shows the CDC's scoring of the respective sociodemographic variables in the US (A) Socioeconomic Status, (B) Household Composition & Disability, (C) Minority Status & Language, (D) Housing Type & Transportation, (E) Epidemiological Factors and (F) Healthcare System Factors maps for the United States.

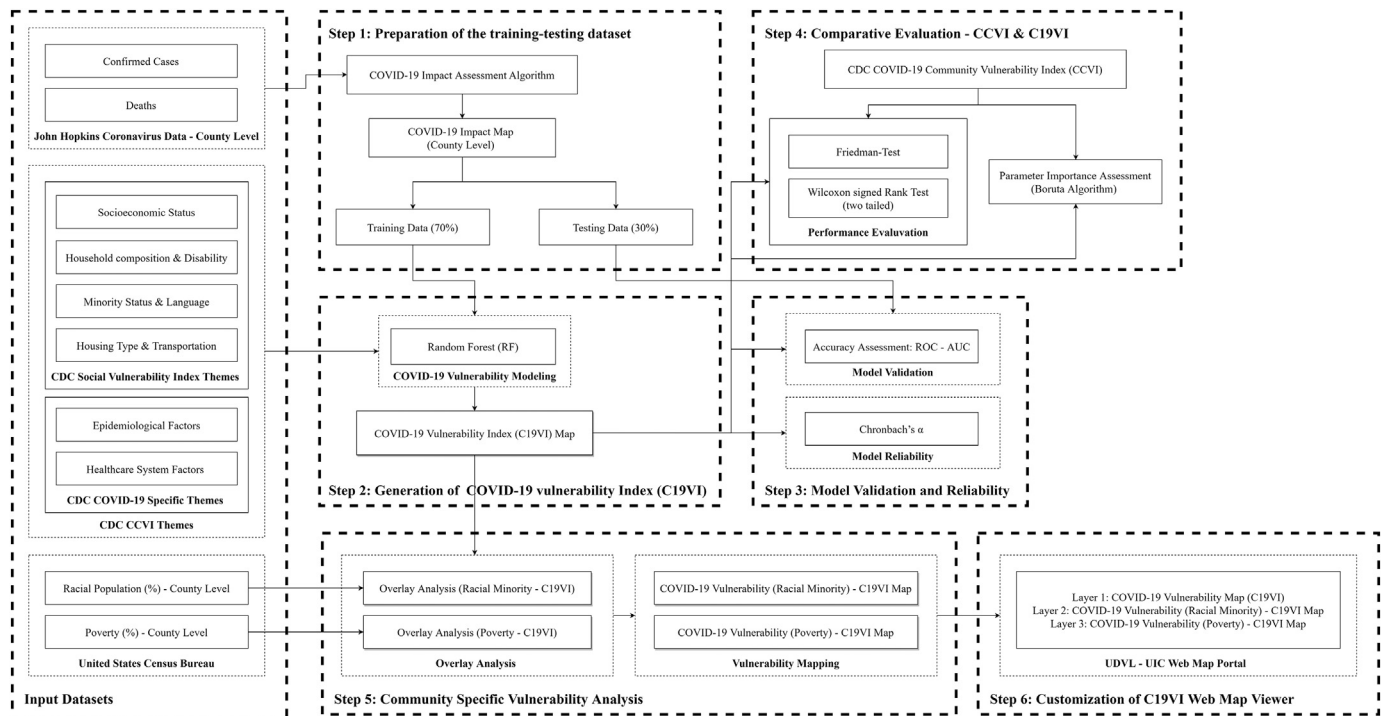


Fig. 3. Flow diagram of the C19VI methodology. The figure illustrates the C19VI methodology, from input datasets to customization of the C19VI web map viewer, in step-wise details.

2.3. Generation of COVID-19 Vulnerability Index (C19VI)

Our study methodology was built and tested in six steps (Fig. 3). First, the training-testing data was prepared using the “most affected” and the “non-significantly” affected counties using the proposed ‘COVID-19 Impact Assessment’ algorithm. Second, COVID-19 vulnerability map was generated using the RF machine learning technique (Breiman, 2001; Liaw and Wiener, 2002). Third, vulnerability modeling was validated using Receiver Operating Characteristic (ROC)-Area Under the ROC Curve (AUC) technique (Altman and Bland, 1994; Fan et al., 2006; DeLong et al., 1988) and Cronbach's α (Cronbach, 1951). Fourth, our C19VI modeling was comparatively assessed against the CDC's CCVI using Friedman (Friedman, 1937) and two-tailed Wilcoxon signed rank (Wilcoxon, 1945) test and later, the input themes contribution to the respective vulnerability index, the output, were ranked using, and Boruta technique (Kursa and Rudnicki, 2010). Fifth, C19VI was analyzed with racial minority population and poverty dataset to determine the disproportionate county-level impact of COVID-19 pandemic. Lastly, an interactive version of the C19VI map with other results was released to the public using the ESRI Web GIS customization toolkit (Builder EWA, 2018). Each step is further detailed below:

1. Preparation of the training-testing dataset: Proposed ‘COVID-19 Impact Assessment’ algorithm was used to map the impact of COVID-19 pandemic on all 3142 counties in the US using confirmed cases and deaths. Out of total 3142 counties, 200 very highly affected and 200 non-significantly affected counties were selected to prepare the COVID-19 vulnerability modeling training and testing dataset. 70% of the total counties (280) were randomly selected and implemented as a training dataset while rest 30% (120) were used for testing.
2. COVID-19 vulnerability modeling: COVID-19 vulnerability modeling was implemented using the RF machine learning technique (Breiman, 2001; Liaw and Wiener, 2002). This model predicts vulnerability of a given county on a continuous scale of 0 (least vulnerable) to 1 (most vulnerable). The map was graded according to the

COVID-19 Vulnerability Index into five vulnerability classes (Foundation S, 2020): very high (>80%), high (80%–60%), moderate (60%–40%), low (40%–20%) and very low (<20%) (Foundation S, 2020).

3. Validation of vulnerability modeling: The effectiveness of RF machine learning technique was specified by evaluating uncertainties in the resulting vulnerability map using the Receiver Operating Characteristic (ROC) - Area Under the Curve (AUC). The ROC-AUC is the standard technique most frequently employed in vulnerability modeling studies to evaluate the modeling accuracy. The ROC curve maps the true positive rate on the Y-axis and the false positive rate on the X-axis. It depicts the trade-off between the two rates. In the ROC technique, AUC (which varies from 0.5 to 1.0) are used for evaluating the model accuracy. The AUC for prediction curve was computed, based on the trapezium method. From the list of 200 very highly affected and 200 non-significantly affected counties, 30% (120) of the total counties were randomly selected for model validation. The ROC-AUC analyzed the conformity between the validation fold of the training-testing dataset and the products of the applied technique. We computed Cronbach's α (Cronbach, 1951) for the developed C19VI index to measure reliability by assessing the C19VI values, the output, with CDC's six theme variables, the input.
4. Comparison of the CCVI and C19VI: As both the CCVI and the C19VI models were developed using the same six thematic indicators, Friedman (Friedman, 1937) and two-tailed Wilcoxon signed rank (Wilcoxon, 1945) statistical tests were implemented to comparatively assess model vulnerability prediction ability. Next, Boruta feature importance assessment technique (Kursa and Rudnicki, 2010) was used to evaluate the relative importance of input indicators in CCVI and C19VI.
5. Community specific vulnerability analysis: Long-standing systemic, social and economic inequities across the counties have put many people from racial minority groups and living below the poverty line at increased risk of getting sick and dying from COVID-19 (Finch and Hernández Finch, 2020; Ahmed et al., 2020; van Dorn et al., 2020). By overlaying the C19VI map on racial minority

population percentage data, COVID-19 vulnerability specific to racial minority groups were identified. As recommended by CDC, a 13% of the racial minority threshold, i.e. a given county with more than 13% racial minorities residents, was used for computing the COVID-19 vulnerability for racial minority groups (Greener, 2019). Similarly, by overlaying the C19VI map on poverty percentage data, COVID-19 vulnerability specific to economically poor communities were identified. As defined by the Economic Research Service (ERS), United States Department of Agriculture (USDA) a 20% of the poverty threshold, i.e. a given county with more than 20% economically poor residents, was used to estimate the vulnerability for economically poor communities (Taylor, 2018; Mammen and Sano, 2018). ESRI ArcGIS overlay analysis tool (ArcGIS E, 2012) was used to conduct the community-specific vulnerability analysis.

- Customization of C19VI web map viewer: ESRI Web App Builder (Builder EWA, 2018) was used to develop an interactive 'C19VI web map' portal. This portal features three layers: 'C19VI' layer, 'COVID-19 Vulnerability (Racial Minority) - C19VI' layer, and 'COVID-19 Vulnerability (Poverty) - C19VI' layer. Every layer features its own attributes. The 'C19VI' layer displays the C19VI values, COVID-19 Impact Rank, total number of confirmed cases and deaths as of July 31st 2020 for each United States county. The 'COVID-19 Vulnerability (Racial Minority) - C19VI' layer displays the C19VI and minority population percentage for each United States county. The 'COVID-19 Vulnerability (Poverty) - C19VI' layer displays the C19VI and poverty percentage for each United States county. Web GIS portal is set to update every three months. Currently it features both the old web maps for 31st July and new web maps for 31st Oct 2020.

3. Results

3.1. COVID-19 impact assessment

Our 'COVID-19 Impact Assessment' algorithm performed a county-wise assessment of the pandemic using the confirmed cases, deaths and IFRs data from 22nd January 2020 to 31st July 2020. We generated a map of our assessment that groups the impact of the pandemic on all United States counties in one of the six categories (Fig. 4(A)). We found

88 counties with 'very high', 30 with 'high', 73 with 'moderate', 344 with 'low', 214 with 'very low,' and 2393 with 'non-significant' impact due to the COVID-19 pandemic (Fig. 4(B)). Top 200 counties with the most impact and the bottom 200 with non-significant impact were used as training and testing datasets for our COVID-19 vulnerability model.

3.2. COVID-19 vulnerability modeling

Using the impact assessment data of the selected United States counties, input themes and the RF technique, we developed COVID-19 Vulnerability Index (C19VI). Fig. 5(A) shows the C19VI map at the scale of 0 to 1. As presented in Fig. 5(B), we computed C19VI for all United States counties and classified them in one of the five vulnerability Classes, 'very high', 'high', 'moderate', 'low', and 'very low'. We found that 11.68% of the counties (367) fall into the 'very low' category, 22.34% (702) in the 'low,' 23.32% (733) in the 'moderate,' 24.34% (765) in the 'high,' and 18.30% (575) in the 'very high' category (Fig. 5(C)). Based on C19VI values, 20 most and least vulnerable counties and their corresponding input theme contribution to the vulnerability are displayed alongside CDC's CCVI in Fig. 6(A) and (B), respectively.

3.3. Model validation and reliability

We used the AUC-ROC technique to validate the prediction accuracy of our C19VI model. As shown in Fig. 7(A) and (B), we found 90% accuracy (AUC = 0.90) during the training phase and 84% accuracy (AUC = 0.84) during the testing phase, respectively. High internal consistency (Cronbach's $\alpha = 0.709$) of C19VI model was revealed using Cronbach's α test (Tavakol and Dennick, 2011; Glen, 2014). Overall, validation and reliability results indicate that the random forest machine learning modeling provides a high quality COVID-19 vulnerability map (C19VI) for the United States (DeLong et al., 1988; Tavakol and Dennick, 2011; Glen, 2014).

3.4. Comparative evaluation – CCVI & C19VI

Since we introduced a new COVID-19 vulnerability assessment index, we quantitatively evaluated its performance against the existing vulnerability model, CDC's CCVI, to assess C19VI's predictive power and applicability. We used Friedman test (Friedman, 1937), two-tailed

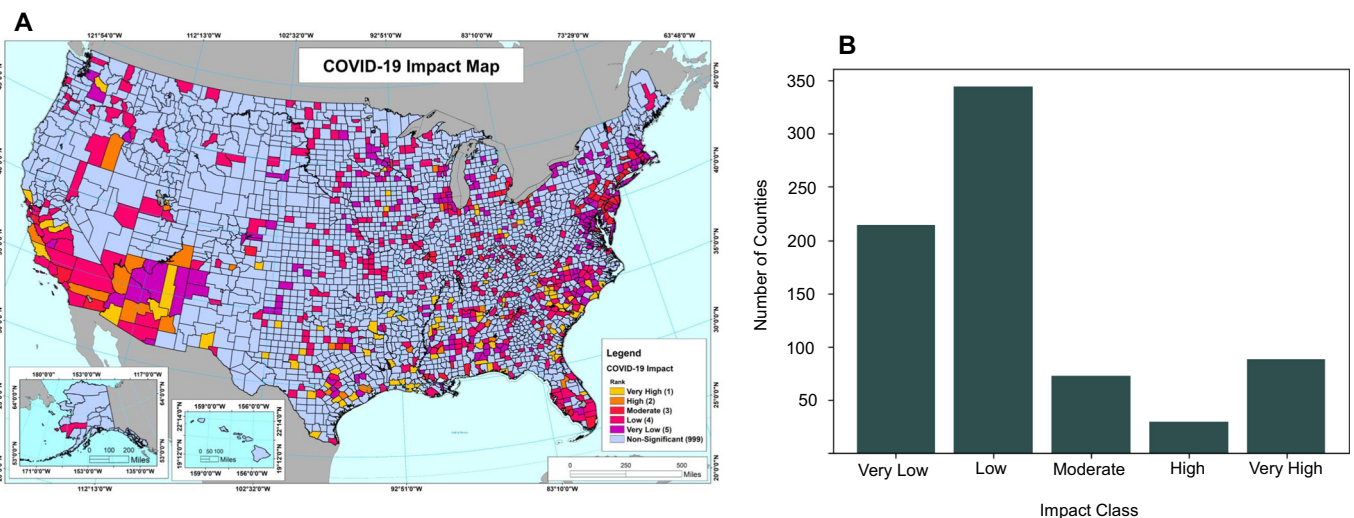


Fig. 4. 'COVID-19 Impact Assessment' algorithm output. (A) Map of COVID-19 impact showing classification of each US county in one of the six impact ranks generated using 'COVID-19 Impact Assessment algorithm'. The map shows counties with very high impact (Rank 1) in yellow, high (Rank 2) in orange, moderate (Rank 3) in red, low (Rank 4) in rose, very low (Rank 5) in purple, and insignificant (Rank 999) in tropical blue. (B) Bar graph depicting total number of counties in each COVID-19 Impact Rank.

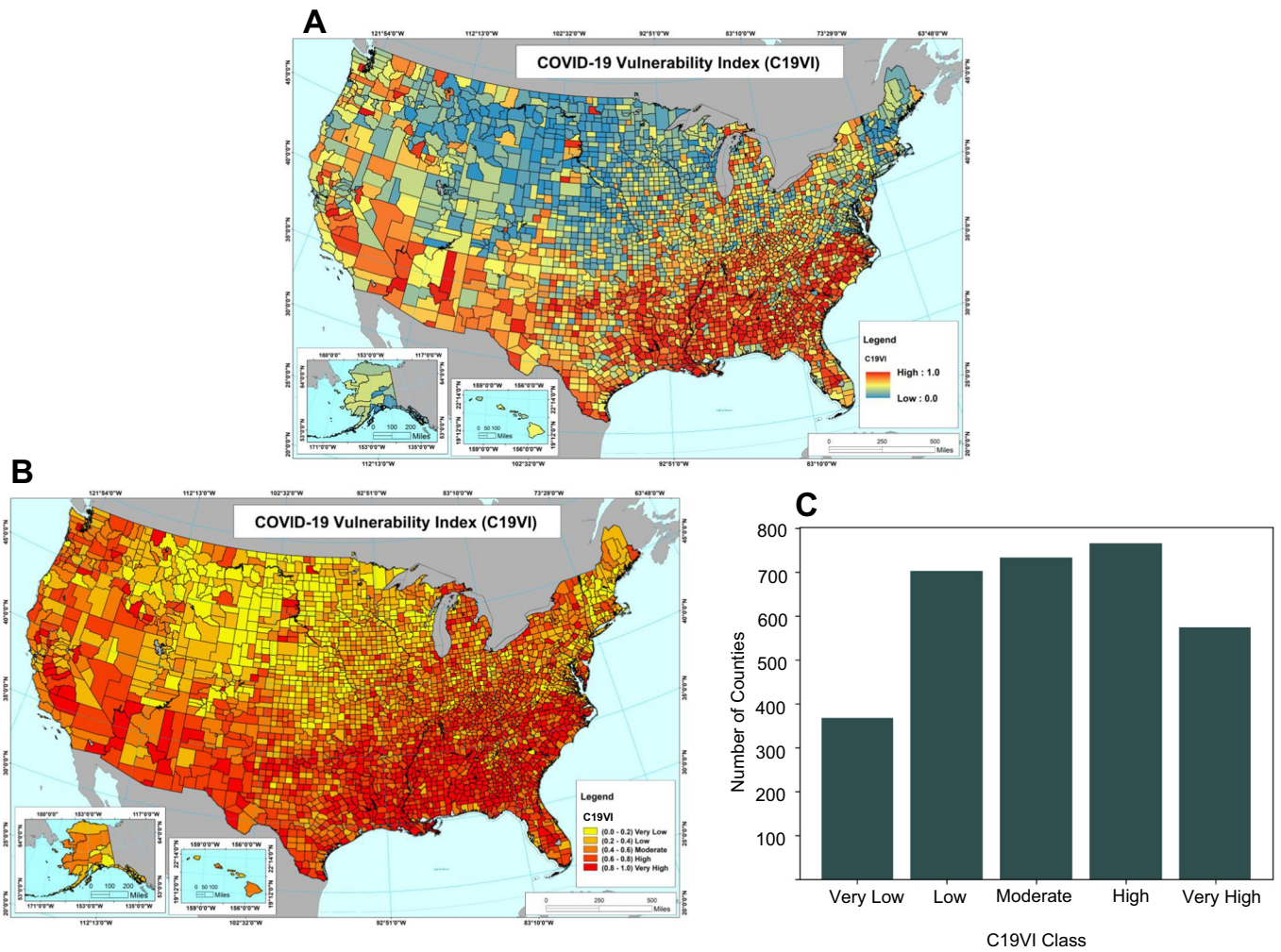


Fig. 5. COVID-19 vulnerability modeling output. (A) COVID-19 Vulnerability Index (C19VI) map on a continuous scale of 0 (least vulnerable) to 1 (most vulnerable). (B) Map of C19VI showing classification of each US county in of the six vulnerability classes generated using the RF machine learning-derived C19VI model. The map shows counties with very high vulnerability (C19VI 0.8–1.0) in red, high (0.6–0.8) in vermilion, moderate (0.4–0.6) in orange, low (0.2–0.4) in amber, and very low (0.2–0.0) in yellow. (C) Bar graph depicting the total number of counties in each C19VI class.

Wilcoxon signed rank test (Wilcoxon, 1945), and Boruta parameter importance assessment technique (Kursa and Rudnicki, 2010) to comparatively evaluate C19VI and CCVI.

1. Friedman and Wilcoxon tests: The Friedman and the two-tailed Wilcoxon signed rank tests detected significant differences ($p < 0.0001$) between the indices given by the two models, C19VI and CCVI. The mean rank of C19VI is 1.614 while mean rank of CCVI is 1.386. The full results of the Friedman and Wilcoxon tests can be found in Tables 2 and 3, respectively.
2. Boruta test: Individual importance of each CDC input themes in determining C19VI and CCVI were quantified using Boruta, a wrapper algorithm. The most important parameter in C19VI was theme 5 (135.17), while in CCVI, it was theme 6 (131.36) (Table 4). Moreover, theme 3 (Minority Status & Language) ranked second in parameter importance in the C19VI model as compared to CCVI, where it is ranked fourth (Table 4). The full results of the Boruta test on CCVI and C19VI are presented in Fig. 8(A) and (B), respectively.

3.5. Community specific vulnerability analysis

The racial minority populations of the United States reside more densely in the southern states and in urban areas (Bureau UC, 2018; Newkirk, 2020; Snyder and Parks, 2020). Our community-specific analysis reveals that the racial minorities disproportionately reside in

counties that are more vulnerable to COVID-19 (Fig. 9(A)). We found that 77.62% counties with racial minority populations >13%, have very high or high (CCVI >0.60) COVID19 vulnerability. Similar to racial minorities, economically poor communities are more likely to be affected by the virus and have higher mortality rates (Snyder and Parks, 2020). The C19VI derived COVID-19 vulnerability with reference to poverty is presented in Fig. 9(B). We find that 82.84% of economically poor counties, where poverty >20%, have very high or high (CCVI >0.60) COVID-19 vulnerability.

3.6. C19VI web map viewer

The Urban Data Visualization Lab (UDVL) at the University of Illinois at Chicago (UIC) have featured an interactive version of the C19VI map which can be accessed at the following URL <https://udv.lab.uic.edu/national-covid-19-vulnerability-index-c19vi>. This map portal is easily accessible on personal computers and mobile devices. See Supplementary material Fig. 1 (A), (B) and (C) for snapshots of the web map viewer.

4. Discussion

Ever since the United States declared a national emergency due to the COVID-9 pandemic in March 2020, the country is grappling against

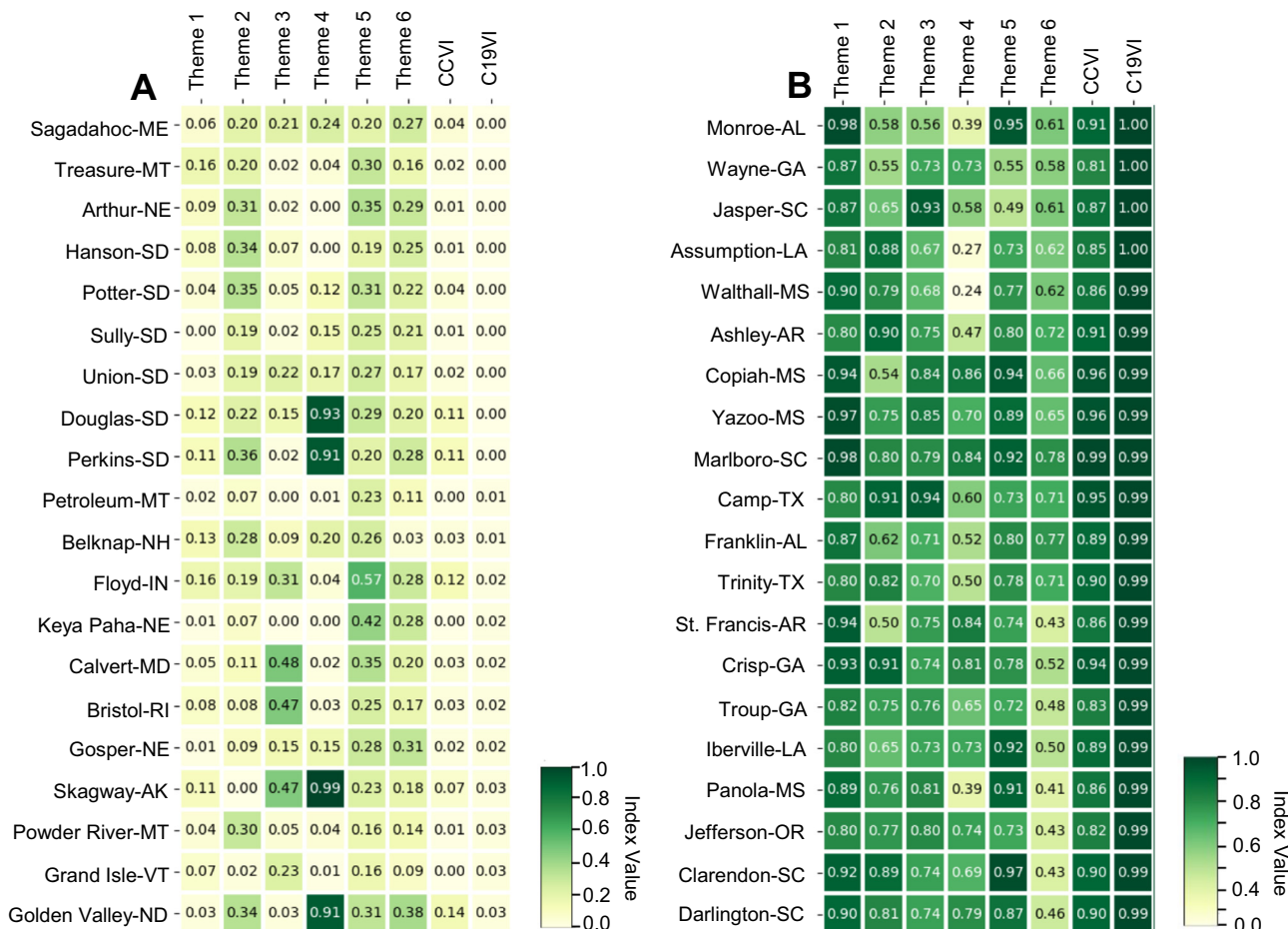


Fig. 6. Heat map of CCVI and C19VI comparative assessment. (A) Heat map of CCVI and C19VI alongside the six input theme indicators of the twenty most vulnerable counties shortlisted based on the C19VI. (B) Heat map of CCVI and C19VI alongside the six input theme indicators of the twenty least vulnerable counties shortlisted based on the C19VI.

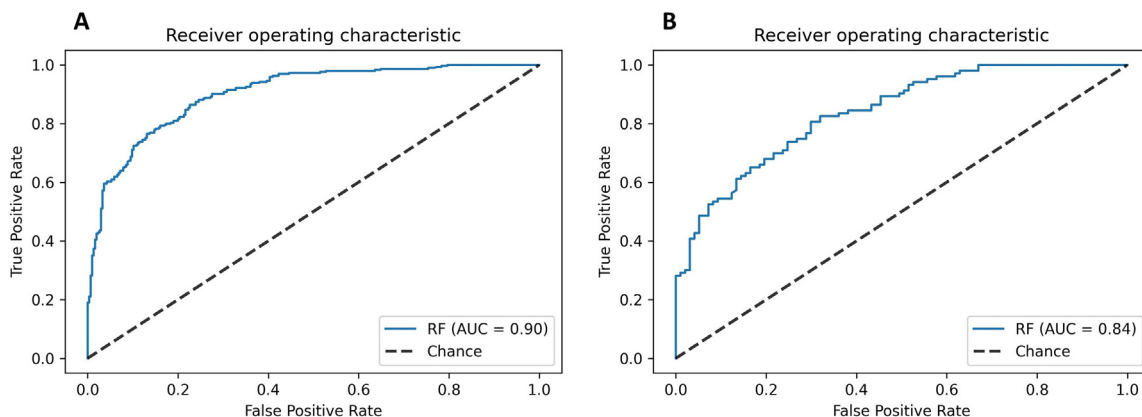


Fig. 7. C19VI model validation using ROC-AUC technique. (A) ROC-AUC curve with 90% accuracy during the training phase. (B) ROC-AUC curve with 84% accuracy during the testing phase.

a huge continuous surge in the incidence and mortality rates (COVID C, 2020; Liu et al., 2020). In the last month alone, June 30th, 2020 to July 31st, 2020, the total number of confirmed cases has risen from 2,729,764 to 4,713,014 while the total death count has increased from 130,313 to 156,826 (COVID C, 2020). It is expected that the United States' COVID-19 death toll will double, potentially reaching more than 0.4 million by the beginning of 2021 (COVID I and Murray, 2020). Thus, keeping in mind the uncontrollable spread, ineffective strategies to check the transmission, disproportionate impact based on systemic inequalities, and heterogeneous impact on different regions, we have developed a county-level COVID-19 Vulnerability Index

(C19VI) that assess the vulnerability of a region using an innovative methodology. This methodology considers the limitations of the existing vulnerability modeling techniques to assess COVID-19

Table 2
Results of Friedman test for CCVI and C19VI.

Index	Degrees of freedom	Chi-squared value	p-Value	Mean rank
C19VI	1	3.841	<0.0001	1.614
CCVI				1.386

Table 3
Comparison of CCVI and C19VI using two-tailed Wilcoxon signed-rank test.

Pairwise comparison	z-Statistic	p-Value
CCVI - C19VI	-12.461	<0.0001

Table 4
Importance assessment of the input themes in CCVI and C19VI using Boruta algorithm.

Theme	Indicator	CCVI - mean importance	C19VI - mean importance
1	Socioeconomic Status	74.56	105.60
2	Household Composition & Disability	43.20	34.45
3	Minority Status & Language	68.69	112.98
4	Housing Type & Transportation	47.81	67.86
5	Epidemiological Factors	102.84	135.17
6	Healthcare System Factors	131.36	72.25

vulnerability nationwide and performs a disproportionate analysis that points out the existing health disparities in the country. In the following sections, we discuss the unique characteristics of our C19VI model and the utility of C19VI in the nationwide and community-specific vulnerability assessment.

4.1. Novel approach of vulnerability modeling

Recently, many researchers (Boldog et al., 2020; Amiri et al., 2020; Acharya and Porwal, 2020; Kim and Bostwick, 2020; Sarkar and Chouhan, 2020; Mishra et al., 2020; Cahill et al., 2020; Marvel et al., 2020) have internationally conducted COVID-19 risk and vulnerability assessment using mathematical modeling (Amiri et al., 2020) or linear statistical techniques with sociodemographic, economic, and health indicators. Notorious examples of such analytical approaches are: i) equal weight assignment (Foundation S, 2020; Acharya and Porwal, 2020); ii) principal component analysis (PCA) (Kim and Bostwick, 2020; Sarkar and Chouhan, 2020); and iii) heuristic modeling (Mishra et al., 2020). In addition, a few studies were conducted by

performing numerical simulations of the total confirmed cases, deaths, and IFRs using statistical (Boldog et al., 2020; Cahill et al., 2020) and machine learning (Marvel et al., 2020) techniques to compute COVID-19 specific vulnerability. While these approaches have enhanced the domain of pandemic vulnerability modeling, they show at least one of the three underlying limitations recognized by the public health planners and policy makers that impair an optimal modeling process. Either they implement an equal weight assignment approach in vulnerability assessment, assume steady transmission rates in mathematical modeling, or treat confirmed cases, deaths, and IFR as constants for vulnerability assessment. However, it is known that 1) not all input themes variables are equally important in determining vulnerability (Acharya and Porwal, 2020), 2) confirmed cases, deaths, and IFR are not biological constants in a pandemic and thus, they do reflect the severity of the pandemic in a particular context, at a particular time (Ritchie and Roser, 2020), 3) selection of constant pandemic transmission rates for both mathematical analysis and data fitting is unrealistic in nature and does not encounter the implications of government implemented disease control actions and individuals' voluntary responses against COVID-19 (Ritchie and Roser, 2020; Wang, 2020).

Thus, we optimized these limiting factors by introducing RF machine learning, a non-linear, non-parametric predictive modeling technique which can efficiently compute large datasets and account for complex interactions between variables (Breiman, 2001; Liaw and Wiener, 2002). Additionally, in comparison to other machine learning techniques, RF is easier to tune and does not overfit the data making it suitable for the pandemic vulnerability modeling (Liaw and Wiener, 2002).

Furthermore, we optimized the dynamic characteristics of the pandemic by developing novel 'COVID-19 Impact Assessment' algorithm, which assesses the regional pandemic impact by performing trend and homogeneity analysis on daily datasets rather than static values for a defined period. Trend and homogeneity assessments help characterize the course of the pandemic and point out the COVID-19 response through changes in healthcare infrastructure or policies in a given region by identifying subtle changes in daily datasets (Mann, 1945; Kendall, 1955; Theil, 1950; Sen, 1968; Pettitt, 1979). Moreover, besides optimization, our impact assessment algorithm also serves to enhance

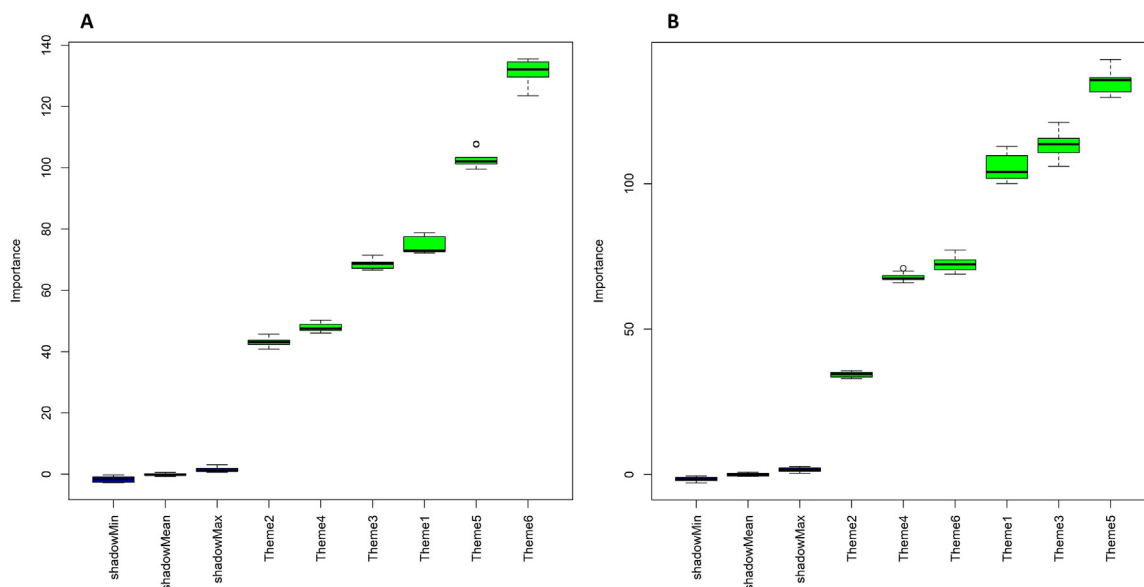


Fig. 8. Importance assessment of the six input theme indicators in CCVI and C19VI using Boruta algorithm. Panel A and B shows the box plot summary of the Boruta input parameter importance assessment of CCVI and C19VI, respectively. (A) Box plot (mean, median, minimum, and maximum Z) for each input theme in the increasing order of importance on the X-axis for the CCVI model. (B) Box plot (mean, median, minimum, and maximum Z) for each input theme in the increasing order of importance on the X-axis for the C19VI model.

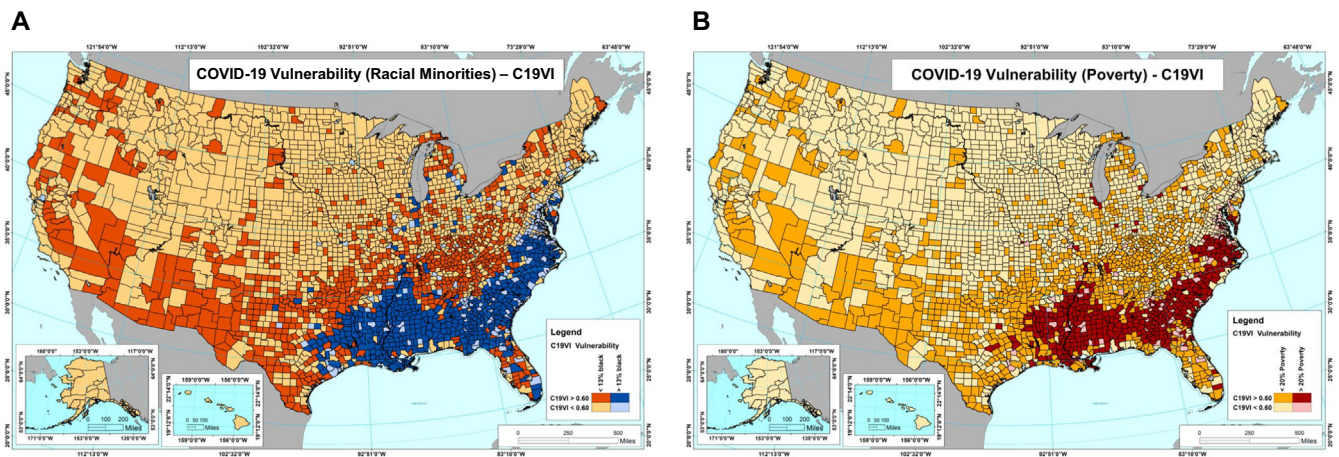


Fig. 9. C19VI facilitated community-specific vulnerability assessment. C19VI of all US counties was overlaid with racial minority population percentage data and poverty percentage data to generate panel A and B, respectively. (A) Map of the US counties showing COVID-19 vulnerability of the racial minorities. The map shows counties with high vulnerability (C19VI > 0.6) and higher than 13% racial minorities in cobalt, low vulnerability (C19VI < 0.6) and higher than 13% racial minorities in tropical blue, high vulnerability (C19VI > 0.6) and lower than 13% racial minorities in red, and low vulnerability (C19VI < 0.6) and lower than 13% racial minorities in chardonnay. (B) Map of the US counties showing COVID-19 vulnerability of the economically poor communities. The map shows counties with high vulnerability (C19VI > 0.6) and higher than 20% poverty in red, low vulnerability (C19VI < 0.6) and higher than 20% poverty in pink, high vulnerability (C19VI > 0.6) and lower than 20% racial minorities in orange, and low vulnerability (C19VI < 0.6) and lower than 20% poverty in chardonnay.

vulnerability modeling to be driven by the chronic disease burden, healthcare infrastructure, and policy impact such as lockdown phases.

In conjunction with the optimized impact assessment algorithm, high training (90%) and testing (84%) accuracy with favorable internal reliability score (Cronbach's $\alpha = 0.709$) of the RF machine learning-derived predictive modeling technique makes C19VI an accurate and reliable index. Besides, despite using the same input, our machine-learning derived C19VI produced significantly different and consistent results in contrast to the CDC's CCVI as elucidated through the Friedman and Wilcoxon signed-rank tests. Moreover, Boruta algorithm-based importance assessment of the variables for both methods shows that both methods handled the variables with major notable differences. The divergence between the two methods indicates that the C19VI was able to capture non-linear relationships in the variables which were not captured with the linear 'equal weight assignment approach' used in the CDC's CCVI model (Foundation S, 2020).

The ability of capturing non-linearity in the input variables alongside the unique characteristics of the C19VI methodology makes the C19VI an optimal index to be considered for vulnerability assessment.

4.2. Nationwide vulnerability analysis

Our nationwide vulnerability analysis reveals interesting patterns of vulnerability distributions around the country. We found that most of the vulnerable counties are concentrated in the southern states. As shown in the Fig. 5(B), nine of the top ten states with the most percentage of 'very highly' and 'highly' vulnerable counties—Alabama (94%), Mississippi (90%), Louisiana (89%), Georgia (76%), South Carolina (76%), Arkansas (76%), Tennessee (76%), North Carolina (74%), Florida (70%) and New Mexico (70%)—are southern states. Secondly, although the counties in the northeastern states had a significant number of confirmed cases, many states in this region showed low vulnerability. Four of the top ten states with the highest percentage of 'very low' and 'low' vulnerable counties—Connecticut (100%), New Hampshire (100%), Wyoming (96%), North Dakota (91%), Wisconsin (83%), Nebraska (83%), Montana (82%), Maine (81%), Vermont (79%), Minnesota (78%)—are northeastern states. Thus, we see that the counties with high and low vulnerability are clustered together, respectively, and the similar vulnerability classes are distributed discretely in the different geographical regions of the United States. This non-uniform, region-dependent

distribution of COVID-19 vulnerability in the United States can be associated with adoption of different public health strategies on a state-level and with the regional sociodemographic distribution in the United States.

This index can also be used alongside other epidemiological data, such as disease transmission, infection fatality rate, the proportion of cases needing hospitalization, intensive care unit admissions, or ventilator support to heighten the preparedness of a district or state, as well as planning and executing the response. We also recommend the use of our C19VI index alongside the CDC's Social Vulnerability Index (SVI) for developing disaster risk assessment and preparedness plans in COVID-19 affected regions. For example, in the times of COVID-19 pandemic, the C19VI should be used alongside the SVI for the disaster management in counties with frequent forest fires, tornadoes or hurricanes.

4.3. Racial/economic disproportionality using vulnerability

COVID-19 has brought previously unaddressed health disparities of racially marginalized and economically poor communities to the forefront of both disaster management officials and government concern. By overlaying the C19VI with the race and poverty data, we found that racial minorities and economically poor Americans disproportionately reside in communities that are more vulnerable to COVID-19. This finding is consistent with other evidences highlighting the disproportionate incidence of COVID-19 among minority groups and poor communities (Moore, 2020; Finch and Hernández Finch, 2020; van Dorn et al., 2020; Fortuna et al., 2020; Gaynor and Wilson, 2020; Sequist, 2020). The currently available county-level cases and deaths dataset, that is segregated by minority population and economic status, is not sufficient to generate reliable COVID-19 risk estimates. The analysis proposed here provides an excellent way to help the communities that disproportionately bear the burden of this crisis, by precisely identifying these areas.

Thus, the C19VI is intended to help policy makers, non-profit entities, private companies, local organizations, and the general public to improve the COVID-19 contingency planning. This index may also be useful for: i) a better management of distribution of resources; ii) addressing pandemic-associated healthcare disparities; iii) providing businesses with opportunities to grow where support is needed the most; and iv) raising public awareness of the COVID-19 pandemic.

Additionally, this methodology may be useful to develop advanced predictive modeling techniques by professionals in academia.

4.4. Limitations

Ideally, it would be possible to calculate the index at a census-tract level. However, several important variables used to define vulnerability were not available at this level. Hence, this analysis is restricted to the county-level. Secondly, being based on the ranking of counties for CDC six themes, our C19VI is a relative index of each county rather than being an absolute score. Thirdly, we were unable to test the external validity of C19VI since no accurate and stable measure of vulnerability was available. Fourthly, the 'COVID-19 Impact Assessment' algorithm requires to be evaluated for space and time complexity, and internal errors. Finally, more sophisticated techniques like Deep Learning, Heuristic and Statistical (weighted sum) with confirmed cases, deaths and IFR with hospitalized, ICU and asymptomatic patients can be used for accessing the detailed impact of pandemic in the future.

5. Conclusions

In this work, we proposed an innovative approach to conduct the vulnerability assessment of COVID-19 within the United States at the county level. This approach integrates the reliable and high-functioning domains of machine learning and predictive modeling. RF machine learning technique in conjunction with novel 'COVID-19 Impact Assessment' algorithm constituted the basis of our vulnerability model. In our model, the input data is processed under a non-linearly fashion, with high training-testing accuracy and favorable internal validity to generate a COVID-19 Vulnerability Index of the United States. Besides promising validation results, comparative assessment of our technique confirms that the C19VI predictive model is a reliable and pragmatic alternative to the CDC's CCVI. Of note, our innovative approach used to develop a vulnerability index has enhanced the nationwide capacity to predict the potential harms accurately which may help curtail the distressing course and consequences of the pandemic. When combined with the racial minority and poverty dataset, C19VI demonstrated its efficiency in disproportionate vulnerability analysis and helped validate the existing disparities. Thus, as a COVID-19 risk assessment tool, this index may assist the public health officials and other authorities in formulating and implementing policies to manage the pandemic. Lastly, the concept of our methodology could also be useful for vulnerability model in general, representing a potential progress in digital public health applications.

CRedit authorship contribution statement

AT conceived the study, designed methodology and programmed the model. AT and VA-BR performed statistical analysis and prepared the figures. AVD worked on the data acquisition and data interpretation. AT, AVD and ERAO performed data analysis. AT and AVD wrote the manuscript. AVD and ERAO critically revised the manuscript. All authors interpreted the results and approved the final version for submission.

Declaration of competing interest

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

Acknowledgements

We would like to thank Johns Hopkins University, Centers for Disease Control and Prevention (CDC), United States Census Bureau and Department of Homeland Security for sharing information necessary for the outbreak investigation, vulnerability modeling, analysis

and mapping. We sincerely acknowledge the support and collaboration of Dr. Moira Zellner, Director, Urban Data Visualization Lab (UDVL), CUPPA, UIC, Chicago, USA. We thank Anton Rozhkov, Ph.D. Candidate, Urban Data Visualization Lab (UDVL), CUPPA, UIC, Chicago, USA for customizing the web map portal and Abhilasha Dixit, Research Scholar, Centre of Excellence in Disaster Mitigation and Management (CoEDMM), Indian Institute of Technology-Roorkee, India for her expert technical assistance.

Data sharing

We used publicly available data and have referenced the sources in the paper. Developed data layers are also available in openly accessible GitHub repository. This repository updated regularly to include the latest datasets.

https://github.com/AnujTiwari/COVID19-Vulnerability-Index_C19VI

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.scitotenv.2021.145650>.

References

- Acharya, R., Porwal, A., 2020. A vulnerability index for the management of and response to the COVID-19 epidemic in India: an ecological study. *Lancet Glob. Health* 8 (9), e1142–e1151. [https://doi.org/10.1016/S2214-109X\(20\)30300-4](https://doi.org/10.1016/S2214-109X(20)30300-4).
- Agency. FEM. Bringing Resources to State, Local, Tribal & Territorial Governments. Washington, DC: US Department of the Homeland Security, Federal Emergency Management Agency. 2020.
- Ahmed F, Ahmed Ne, Pissarides C, Stiglitz J. Why inequality could spread COVID-19. *The Lancet Public Health* 2020; 5(5): e240.
- Altman, D.G., Bland, J.M., 1994. Diagnostic tests 3: receiver operating characteristic plots. *BMJ. British Medical Journal* 309 (6948), 188.
- Amiri S, Thorn EL, Mansfield JJ, Mellacheruvu P, Monsivais P. Data-driven Development of a Small-area COVID-19 Vulnerability Index for the United States. *medRxiv* 2020.
- Amram, O., Amiri, S., Lutz, R.B., Rajan, B., Monsivais, P., 2020. Development of a Vulnerability Index for Diagnosis With the Novel Coronavirus, COVID-19, in Washington State. *USA, Health & Place*.
- ArcGIS E, 2012. 10.1. Redlands, California, ESRI.
- Boldog, P., Tekeli, T., Vizi, Z., Dénes, A., Bartha, F.A., Röst, G., 2020. Risk assessment of novel coronavirus COVID-19 outbreaks outside China. *J. Clin. Med.* 9 (2), 571.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32.
- Builder EWA, 2018. What is Web App Builder for ArcGIS. Accessed September 24.
- Bureau UC, 2018. US Census Bureau QuickFacts: United States.
- Cahill G, Kutac C, Rider NL. Visualizing and Assessing US County-Level COVID19 Vulnerability. *medRxiv* 2020.
- COVID C. global cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). ArcGIS Johns Hopkins CSSE Retrieved August; 01: 2020.
- COVID, I., Murray, C.J., 2020. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *MedRxiv* <https://doi.org/10.1101/2020.03.27.20043752> In preparation. <https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1.full>.
- Cronbach LJ. Coefficient alpha and the internal structure of tests. *psychometrika* 1951; 16 (3): 297–334.
- Dang, H.-A., Huynh, T.L.D., Nguyen, M.-H., 2020. Does the Covid-19 pandemic disproportionately affect the poor? Evidence from a six-country survey.
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845.
- DHS. Homeland infrastructure foundation-level data. 2016.
- ESRI E, 1998. Shapefile Technical Description. An ESRI White Paper 4, 1.
- Fan, J., Upadhye, S., Worster, A., 2006. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine* 8 (1), 19–20.
- Finch, W.H., Hernández Finch, M.E., 2020. Poverty and Covid-19: rates of incidence and deaths in the United States during the first 10 weeks of the pandemic. *Front. Sociol.* 5, 47.
- Flanagan BE, Gregory EW, Hallisey EJ, Heitgerd JL, Lewis B. A social vulnerability index for disaster management. *Journal of homeland security and emergency management* 2011; 8(1).
- Fortuna, L.R., Tolou-Shams, M., Robles-Ramamurthy, B., Porche, M.V., 2020. Inequity and the disproportionate impact of COVID-19 on communities of color in the United States: the need for a trauma-informed social justice response. *Psychological Trauma, Theory, Research, Practice, and Policy*.
- Foundation S, 2020. The COVID-19 Community Vulnerability Index (CCVI).
- Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* 32 (200), 675–701.

- Gaynor, T.S., Wilson, M.E., 2020. Social vulnerability and equity: the disproportionate impact of COVID-19. *Public Adm. Rev.* 8 (5), 832–838. <https://doi.org/10.1111/puar.13264>.
- Glen, S., 2014. Cronbach's alpha: simple definition, use and interpretation. Retrieved February 18, 2019.
- Greener JR. Improving Health Equity for Black Communities in the Face of Coronavirus Disease-2019.
- Karaye, I.M., Horney, J.A., 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.
- Kendall MG. Rank Correlation Methods. 1955. Charles Griffin, London 1955.
- Kim SJ, Bostwick W. Social Vulnerability and Racial Inequality in COVID-19 Deaths in Chicago. *Health Educ. Behav.* 2020; 47(4).
- Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. *J. Stat. Softw.* 36 (11), 1–13.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R news* 2 (3), 18–22.
- Liu P, Beeler P, Chakrabarty RK. COVID-19 Progression Timeline and Effectiveness of Response-to-Spread Interventions across the United States. medRxiv 2020.
- Magnani, C., Azzolina, D., Gallo, E., Ferrante, D., Gregori, D., 2020. How large was the mortality increase directly and indirectly caused by the COVID-19 epidemic? An analysis on all-causes mortality data in Italy. *Int. J. Environ. Res. Public Health* 17 (10), 3452.
- Mammen, S., Sano, Y., 2018. Rural, low-income families and their well-being: findings from 20 years of research. *Family Science Review* 22 (1), 1–8.
- Mann, H.B., 1945. Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society* 245–259.
- Marvel S, House J, Wheeler M, et al. The COVID-19 Pandemic Vulnerability Index (PVI) Dashboard: monitoring county level vulnerability. medRxiv 2020.
- Meehan, M.T., Rojas, D.P., Adekunle, A.I., et al., 2020. Modelling insights into the COVID-19 pandemic. *Paediatr. Respir. Rev.* <https://doi.org/10.1016/j.prrv.2020.06.014> Submitted for publication.
- Mishra, S.V., Gayen, A., Haque, S.M., 2020. COVID-19 and urban vulnerability in India. *Habitat International* 103, 102230.
- Moore, J.T., 2020. Disparities in incidence of COVID-19 among underrepresented racial/ethnic groups in counties identified as hotspots during June 5–18, 2020–22 states, February–June 2020. *MMWR Morb. Mortal. Wkly Rep.* 69.
- Newkirk, V., 2020. The Coronavirus's unique threat to the south. *The Atlantic* 2, 2020.
- Organization WH, March 2020. Rolling updates on coronavirus disease (COVID-19). Updated 20, 2020.
- Oster, A.M., Kang, G.J., Cha, A.E., et al., 2020. Trends in number and distribution of COVID-19 hotspot counties—United States, March 8–July 15, 2020. *Morb. Mortal. Wkly Rep.* 69 (33), 1127.
- Pettitt, A., 1979. A non-parametric approach to the change-point problem. *J. R. Stat. Soc.: Ser. C: Appl. Stat.* 28 (2), 126–135.
- Ritchie, H., Roser, M., 2020. What do we know about the risk of dying from COVID-19. *Our World in Data*-March 25.
- Sambanis, A., Kim, S., Osiecki, K., 2019. Cailas MD. A New Approach to the Social Vulnerability Indices, Decision Tree-based Vulnerability Classification Model.
- Sarkar, A., Chouhan, P., 2020. COVID-19: district level vulnerability assessment in India. *Clinical Epidemiology and Global Health* 9, 204–215. <https://doi.org/10.1016/j.cegh.2020.08.017>.
- Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.* 63 (324), 1379–1389.
- Sequist TD. The disproportionate impact of Covid-19 on communities of color. *NEJM Catalyst Innovations in Care Delivery* 2020; 1(4).
- Snyder B, Parks V. Spatial Variation in Socio-ecological Vulnerability to COVID-19 in the Contiguous United States. Available at SSRN 3587713 2020.
- Tai, D.B.G., Shah, A., Doubeni, C.A., Sia, I.G., Wieland, M.L., 2020. The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States. *Clin. Infect. Dis.* <https://doi.org/10.1093/cid/ciaa815> In preparation.
- Tavakol, M., Dennick, R., 2011. Making sense of Cronbach's alpha. *Int. J. Med. Educ.* 2, 53.
- Taylor MM. Rural Health Disparities: The Economic Argument. Application of the Political Economy to Rural Health Disparities: Springer; 2018: 9–17.
- Theil H. A rank-invariant method of linear and polynomial regression analysis (Parts 1–3). *Ned Akad Wetensch Proc Ser A*; 1950; 1950. p. 1397–412.
- van Dorn A, Cooney RE, Sabin ML. COVID-19 exacerbating inequalities in the US. *Lancet (London, England)* 2020; 395(10232): 1243.
- Wang, J., 2020. Mathematical models for COVID-19: applications, limitations, and potentials. *Journal of public health and emergency* 4.
- Wilcoxon, F., 1945. Individual comparisons by ranking methods. *Biom. Bull.* 1, 80–83.
- Zhao J, Lee M, Ghader S, et al. Quarantine fatigue: first-ever decrease in social distancing measures after the COVID-19 pandemic outbreak before reopening United States. arXiv preprint arXiv:200603716 2020.