Early Stage Machine Learning Based Prediction of US County Vulnerability to the COVID-19 Pandemic

Mihir Mehta, Juxihong Julaiti, Paul Griffin, Soundar Kumara

Submitted to: JMIR Public Health and Surveillance
on: April 17, 2020

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Early Stage Machine Learning Based Prediction of US County Vulnerability to the COVID-19 Pandemic

Mihir MehtaMSc, MS, ; Juxihong JulaitiMSc, ; Paul GriffinDPhil, ; Soundar KumaraDPhil,

Corresponding Author:
Paul GriffinDPhil,
Phone: +1765-496-7395
Email: paulgriffin@purdue.edu

Abstract

Background: The rapid spread of COVID-19 means that government and health services providers have little time to plan and design effective response policies. It is therefore important to rapidly provide accurate predictions of how vulnerable geographic regions such as counties are to the spread.

Objective: To develop county level prediction around near future disease movement for COVID-19 occurrences using publicly available data.

Methods: We estimate county level COVID-19 occurrences using data from March 14-31, 2020 based on data fused from multiple publicly available sources inclusive of health statistics, demographics, and geographical features. We developed a 3-stage model to quantify, firstly the probability of COVID-19 occurrence for unaffected counties using XGBoost classifier and secondly, the number of potential occurrences of a county via XGBoost regression. Thirdly, these results are combined to compute the county level risk. This risk is then used as an estimated after-five-day-vulnerability of the county.

Results: Using data from March 14-31, 2020, the model shows a sensitivity over 71.5% and specificity over 94%. We found that population, population density, percentage of people aged 70 or greater and prevalence of comorbidities play an important role in predicting COVID-19 occurrences. We found a positive association between affected and urban counties as well as less vulnerable and rural counties.

Conclusions: The developed model can be used for identification of vulnerable counties and potential data discrepancies. Limited testing facilities and delayed results introduces significant variation in reported cases and produces a bias in the model.
Clinical Trial: Not Applicable

(JMIR Preprints 17/04/2020:19446)
DOI: https://doi.org/10.2196/preprints.19446

Preprint Settings

Original Manuscript

Early Stage Machine Learning Based Prediction of US County Vulnerability to the COVID-19 Pandemic

Abstract

Background: The rapid spread of COVID-19 means that government and health services providers have little time to plan and design effective response policies. It is therefore important to quickly provide accurate predictions of how vulnerable geographic regions such as counties are to the spread.

Objective: To develop county level prediction around near future disease movement for COVID-19 occurrences using publicly available data.

Methods: We estimate county level COVID-19 occurrences for the period March 14-31, 2020 based on data fused from multiple publicly available sources inclusive of health statistics, demographics, and geographical features. We developed a three-stage model using XGBoost- a machine learning algorithm to firstly quantify the probability of COVID-19 occurrence and secondly, estimate the number of potential occurrences for unaffected counties. Finally, these results are combined to predict the county level risk. This risk is then used as an estimated after-five-day-vulnerability of the county.

Results: The model predictions have shown a sensitivity over 71% and specificity over 94% for models built using data from 14th March to 31st March 2020. We found that population, population density, percentage of people aged 70 or greater and prevalence of comorbidities play an important role in predicting COVID-19 occurrences. We observed a positive association at county level between urbanicity and vulnerability to COVID-19.

Conclusions: The developed model can be used for identification of vulnerable counties and potential data discrepancies. Limited testing facilities and delayed results introduces significant variation in reported cases and produces a bias in the model.

Trial Registration: Not Applicable

Key Words: COVID-19, prediction model, county-level vulnerability, machine learning, XGBoost

Introduction

The continued spread of confirmed cases of COVID-19, absence of a vaccine, limited resources for testing and assisting people with confirmed cases have presented a great challenge for our public health and healthcare provider systems. To this point, nonpharmaceutical interventions such as social distancing are the only effective mitigation measures. The rapid spread of the disease means that government and health services have very little time to plan and design effective response policies such as resource and workforce planning. Accurately predicting the near future COVID-19 spread at sufficient granularity would provide these organization with better information and time to appropriately plan and respond.

We have developed a three-stage machine learning model to estimate COVID-19 spread outcomes at the US county level. In the first stage, we estimate the probability that a county has at least one confirmed COVID-19 case. In the second stage, we estimate the number of COVID-19 occurrences given that county has at least one case. Finally, we combine the results from the two stages to estimate those counties that have the greatest and least vulnerability for changes in disease prevalence for the next five-day period.

There has been significant epidemiological work for previous coronavirus pandemics such as MERS and SARS [1]. For example, Badawi et al. [2] performed systematic analysis of prevalence of comorbidities in MERS using data from 12 studies and found that diabetes and hypertension were present in 50% of the cases. Matsuyama et al. [3] systematically reviewed studies involving laboratory confirmed MERS cases to measure both the risk of admission to the Intensive Care Unit (ICU) and death. They compared risks by age, gender, and underlying comorbidities. Park et al. [4] reviewed characteristics and associated risks factors of MERS. Bauch et al. [5] surveyed SARS modeling literature focused on understanding the basic epidemiology of the disease and evaluating control strategies. Surveyed SARS models varied in the terms of population studied and geographical characteristics [6,7]. Different designs were used for SARS modeling comprising of deterministic compartmental models [7], stochastic compartmental models [6], a combination of stochastic and deterministic compartmental models [8], discrete-time models [9], logistics curve fitting models [10], contact network models [11] and likelihood-based models [12]. Studies associated with risk factors for SARS [13] and MERS [3,14–20] have found an association between comorbidities and infected cases.

MERS and SARS epidemiological modeling has been done at different granularities such as the country [21,22], specific region [23], and case clusters [6]. Given the much broader reach of COVID-19 compared to MERS and SARS, it is very important to make predictions at a sufficiently high level of granularity. This is particularly important since previous studies have shown that there is considerable heterogeneity in space, transmissibility, and susceptibility [5]. Our approach is developed at county level with inclusion of a variety of health statistics, demographics, and geographical features of counties. Further, we use publicly available data so that any organization could leverage the model. To the best of our knowledge, no work has been done to predict near future infection risk at the county level using the combination of health statistics, demographics, and geographical features of counties.

Methods

Recruitment

We performed an epidemiological study at the US county level using publicly available data to develop a machine learning predictive model. Data analysis was performed from February 15, 2020, to April 3, 2020. The study was reviewed by the Penn State Integrated Research Ethics Board and deemed exempt because it was a deidentified, secondary data analysis. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline [24].

We used US Census data to obtain county level population statistics for age, gender, and density [25,26]. We obtained county level data for diagnosed adult diabetics percentage and cancer crude rate statistics from the Center for Disease Control and Prevention (CDC) [27,28]. We used county level hypertension estimates and chronic respiratory disease mortality rates obtained from the Global Heath Data Exchange (GHDx) [29,30] website, provided by the Institute for Health Metrics and Evaluation. We obtained the centroids for each county from ArcGIS [31]. Finally, we obtained US Census Cartographic Boundary files for each county in JSON format [32] and county level COVID-19 daily occurrences data (confirmed cases) from NYTimes GitHub page [33,34].

Statistical Analysis

Outcomes

There are three primary outcomes for our predictive model: i) the probability that a county has at least one confirmed case of COVID-19, which we define as a positive instance, ii) the number of confirmed COVID-19 cases within a county, which we define as occurrences, and iii) vulnerability of the county.

Previous studies have shown angiotensin-converting enzyme 2 (ACE2) facilitates the infection of COVID-19 [35–37], and that patients with diabetes, hypertension and cardiovascular diseases have an increased expression of ACE2 [35]. County population factors such as density, age, and sex have a significant impact on the spread of an epidemic [38]. Cancer and chronic respiratory diseases have also been shown to increase mortality risk for COVID-19 [39].

The dataset used for our three-stage model contains correlated variables. For example, diabetes and hypertension prevalence, cancer crude rate and old population. Additionally, the underlying relationship between variables was assumed to be non-linear.

Precursor to the Prediction Model

Machine learning techniques help us to derive insights and predict trends using data without the explicit need of programming. They are mainly divided into two types based on the explicit availability of outcomes for a given set of observations: supervised and unsupervised techniques. In supervised techniques, the outcome or dependent variable is available for a given set of observations. Supervised techniques are further divided into regression or classification techniques depending upon data type of the outcome variable: continuous or categorical [40]. In the literature, artificial neural network based deep learning and tree-based gradient tree boosting techniques have demonstrated better prediction capabilities in exploring non-linear relationship among correlated predictors [41–48].

XGBoost (Extreme Gradient Boosting) [49,50] is a gradient tree-based supervised machine learning technique capable of performing both regression and classification tasks. The underlying algorithm combines the results from multiple individual trees with weak predictions (weak learners) to yield accurate final predictions. During the combining process, the algorithm prevents the over-fitting by regularizing objective function. The performance of this technique depends upon effective tuning of multiple hyper-parameters such as learning rate, maximum depth with respect to underlying data distribution. These hyper-parameters can be tuned with the help of random or exhaustive search as well as by using Bayesian optimization. Bayesian optimization method has shown efficiency in terms of accuracy and time [51].

Developing the Prediction Model

To predict COVID-19 outcomes, we divided the problem into three stages. In the first stage, we classified each county either as a positive or negative instance and used the same as a dependent variable. Hence, we built an XGBoost classifier model to learn from the data.

In the second stage, to predict number of occurences- a continuous variable, we leveraged an XGBoost regression model that included data only for positive instances with number of occurrences as the response.

In the last stage, we combined results from the first two stages and calculated the expected occurrences for counties as a measure of county vulnerability. For the calculation of expected occurrences, we multiplied the probability of county belonging to the positive instances derived using the classification model, with potential occurrences the same county will have if it becomes a positive instance derived using the regression model.

Evaluating the Prediction Model

The evaluation process is illustrated with an example for the date 14 March 2020. For this date, modeling data comprised of COVID-19 cases reported at a county level at the end of 14th March along with all other variables obtained from fusion process.

In the first stage - classification problem, this data was divided into an 80:20 ratio for training and testing simultaneously ensuring equivalent representation of both classes (positive and negative instance). With this setup and leveraging HyperOpt package, multiple hyper-parameters of the model were tuned using area under the receiver operating characteristic curve (AUC) and accuracy values as the evaluation criteria. The resultant model was used to compute county level probability score.

In the second stage, regression problem, the dataset was filtered to include only positive instance counties as of 14th March with number of occurences being a dependent variable. Like the first stage, this data was divided into an 80:20 proportion for testing and training and hyper parameters were optimized by leveraging HyperOpt package. Regression problem used root mean squared error (RMSE) value as an evaluation criterion. The best model was used to calculate the number of occurrences associated with counties.

In the final stage, the vulnerability of a county was determined by multiplying the stage one probability score with stage two number of occurrences. This calculated value was utilized to identify riskiest and safest counties. The model is serving as a proxy for estimating after-five-day-vulnerability, the third stage outcome was evaluated using actual COVID-19 numbers observed at the end (14+5) 19th March 2020. To measure sensitivity, among the top 5% riskiest counties estimated at the end of the third stage of the model, the number of counties which were observed to be positive instance as of 19th March were identified. The corresponding fraction was defined as sensitivity. Similarly, the specificity, among the top 10% least vulnerable counties was estimated by the third stage of the model. The number of counties which were continued to be observed as a negative instance were identified and corresponding fraction was reported as specificity. The third stage model was accessed by both sensitivity and specificity.

Finally, the consistency of the three-stage modeling process was verified by repeating this process daily from 14th March till 26th March and assessing the same from 19th March to 31st March.

Results

The variable importance for the overlapping predictors between the final classification and regression models for March 16th is shown in Figure 1. Total population (TOT_POP) was the most important variable for both the classification and regression models. Other important variables included population density, longitude, hypertension prevalence, chronic respiratory mortality rate, cancer crude rate, and diabetes prevalence. Latitude (we use this to identify neighboring counties and the presence or absence of positive class in the neighborhood) and percentage of populations older than 70 years were found to be the least important features of those considered, though still played a role.

Figure 2 shows a map of the USA with the predicted probability of being a positive instance for each county in the USA as a color gradient. County level statistics can be viewed by moving the cursor of the county of interest. The example of New York County as of March 14th is shown in the Figure 2.

Accuracy and AUC for the first stage model is shown in Table 1. Predictions of the model for all US counties are consistent over the 18 days with little variation in AUC and accuracy values. Similarly, RMSE for the second stage model for all US counties is presented in Table e1. The results for first two stages of the model were evaluated till 31st March.

The sensitivities and specificities for the vulnerability predictions for the three-stage model trained on data from March 14th to March 26th are shown in Tables 2 and 3. The values are given for each day. The sensitivity (Table 2) is given by percentage of counties that had no confirmed cases but were identified as being among the 5% most vulnerable had at least one confirmed COVID-19 case five days later. The specificity (Table 3) is given by the percentage of counties identified as being among the 10% least vulnerable with no confirmed cases that still had no confirmed cases five days later.

The dataset is comprised of 37% urban and 63% rural counties based on the urban and rural county definition for year 2013 [52]. To determine if there is an association between urbanicity and vulnerability, we performed a set of one-sided t-tests. The null hypothesis - the 10% least vulnerable counties would have the same proportion of rural counties as the actual proportion of rural counties in the dataset - was rejected for every day from March 14th to March 26th. Additionally, the null hypothesis - the actual positive instances counties would the same proportion of urban counties as the actual proportion of urban counties in the dataset - was also rejected for every day over the analysis period. It can therefore be concluded that there is a positive association between urban and most vulnerable counties as well as rural and least vulnerable counties. The continuous decreasing trend in the confidence interval of the urban counties proportion estimate within actual positive instance counties can be used to infer that COVID-19 is propagating from urban counties to rural counties.

Discussion

We developed a three-stage machine learning model using publicly available data to predict the five-day vulnerability of a US county. The model estimates the likelihood and impact that a county with no documented COVID-19 cases will have within a five-day period and using them, vulnerability prediction for a county is made. Using data from March 14th to Marth 31st, 2020, the model showed a sensitivity over 71.5% and specificity over 94%. We found a positive association between affected counties and urban counties as well as top 10% least vulnerable counties and rural counties. Further, counties with higher population density, a greater percentage of 70 years of above age people, higher diabetes, cardiac illness, and respiratory diseases prevalence are more vulnerable to COVID-19 than their counterparts.

Our model serves multiple purposes. First, it can help in identifying potentially vulnerable counties. This prediction would be a vital component in managing COVID-19 spread by providing vulnerability information based on the likelihood and magnitude of change within five days. That can help health organizations to plan effectively for management of hospital resources and workforce, rapid response teams, and COVID testing kits and testing locations. In addition, there are multiple counties with limited testing facilities, and with current swab-based testing, it takes multiple days to get the results. Thus, occurrences associated with each county fluctuate rapidly daily.

Limitations

There are multiple limitations to our work. First, there are several predictors that we did not include in the model that have known associations with COVID-19. However, one of our goals was to make sure that any organization could use our model by only including data that is publicly available. Second, our analysis (Table e2) found that there is an increasing trend for the coefficient of variation

(CV) for occurrences associated with positive instances counties. Note that CV is a proxy for economic inequality [53–56]. Hence, there is a bias in the response variable, which can reduce the accuracy of the prediction. As testing facilities improve in terms of numbers and efficiency, this bias would be minimized and would be reflected in the model. Given this point, it would useful to look at top riskiest and top safest counties predicted by the three-stage model and examine for potential data discrepancies. Finally, additional feature engineering and stacking methods can be utilized to enhance the prediction capabilities of existing models.

Our work uses open source programming and publicly available data. The full dataset, sample modeling and result outputs available with instructions for use on: https://github.com/mihirpsu/covid_19

Commentary on present models

Presently multiple research groups are providing COVID-19 projections on death and hospitalization cases numbers. Particularly for USA, the CDC website maintains a list of projection providing research groups. These projections are available along with an ensemble projection. As COVID-19 is approached a flattened curve stage, states deployed varied level of easing of restrictions. Thus, these restrictions are expected to alter presently observed dynamics of the disease spread. Hence, they play an important factor in projections. To account for the same, some of these models assume stationary parameters during the projection period while others assume some form of dynamic nature [57]. These projections are provided at different levels: US level [58], states level [59], metropolitan area level [60] and at the county level [61,62]. These projections are developed using - variants of SEIR models [62], deep learning models [63], agent-based models [64], variants of mechanistic disease transmission models [65], renewal equations-based models [66] and statistical models [61]. In all these models, Columbia University’s Meta-Population SEIR Model [62] and University of Iowa's

[61] non-parametric spatial-temporal model provide projections at a county level. Columbia University’s initial model leveraged US Census county level daily commute data during daytime and nighttime to account for the movement of the disease. However, this model does not account for county level population heterogeneity. The University of Iowa's approach is developed using a combination of statistical and mathematical modeling techniques with an assumption of parameter agnostic exponential family based conditional distribution of COVID-19 cases and deaths. This model leverages county level data on intervention policies, demographic characteristics, health-care infrastructure, socioeconomic factors, urban rate, and geographical information. However, they do not account of county level prevalence of co-morbidities. Finally, The University of Texas at Austin [60] model provides projections at the metropolitan area level using mobile-based data. With the better availability of data and information about COVID-19, current models can forecast projections for a longer period with better accuracy than our model. Yet, our model still presents a unique assumption free county level modeling approach accounting for heterogeneity using demographic, health, and geographical features.

Conflicts of Interest: None declared.

References:

1. Baldwin I, Mauro BW di. Economics in the time of COVID-19: A new eBook. VOX CEPR Policy Portal. 2020. ISBN:978-1-912179-28-2
2. Badawi A, Ryoo SG. Prevalence of comorbidities in the Middle East respiratory syndrome coronavirus (MERS-CoV): a systematic review and meta-analysis. International Journal of Infectious Diseases. 2016. [doi: 10.1016/j.ijid.2016.06.015]
3. Matsuyama R, Nishiura H, Kutsuna S, Hayakawa K, Ohmagari N. Clinical determinants of the severity of Middle East respiratory syndrome (MERS): A systematic review and meta-analysis. BMC Public Health 2016; [doi: 10.1186/s12889-016-3881-4]
4. Park JE, Jung S, Kim A. MERS transmission and risk factors: A systematic review. BMC Public Health 2018; [doi: 10.1186/s12889-018-5484-8]
5. Bauch CT, Lloyd-Smith JO, Coffee MP, Galvani AP. Dynamically modeling SARS and other newly emerging respiratory illnesses: Past, present, and future. Epidemiology. 2005. [doi: 10.1097/01.ede.0000181633.80269.4c]
6. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, Leung GM, Ho LM, Lam TH, Thach TQ, Chau P, Chan KP, Lo SV, Leung PY, Tsang T, Ho W, Lee KH, Lau EMC, Ferguson NM, Anderson RM. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. Science 2003; [doi: 10.1126/science.1086478]
7. Hsieh YH, Chen CWS, Hsu SB. SARS Outbreak, Taiwan, 2003. Emerging Infectious Diseases 2004;10(2):201–206. [doi: 10.3201/eid1002.030515]
8. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, Fisman D, Murray M. Transmission dynamics and control of severe acute respiratory syndrome. Science 2003; [doi: 10.1126/science.1086616]
9. Choi BCK, Pak AWP. A simple approximate mathematical model to predict the number of severe acute respiratory syndrome cases and deaths. Journal of Epidemiology and Community Health 2003; PMID:14573591
10. Zhou G, Yan G. Severe Acute Respiratory Syndrome Epidemic in Asia. Emerging Infectious Diseases. 2003. [doi: 10.3201/eid0912.030382]
11. Masuda N, Konno N, Aihara K. Transmission of severe acute respiratory syndrome in dynamical small-world networks. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 2004; [doi: 10.1103/PhysRevE.69.031917]
12. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of Epidemiology 2004; [doi: 10.1093/aje/kwh255]
13. World Health Organization. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS). Who/Cds/Csr/Gar/200311 2003;
14. Omrani AS, Matin MA, Haddad Q, Al-Nakhli D, Memish ZA, Albarrak AM. A family cluster of middle east respiratory syndrome coronavirus infections related to a likely unrecognized asymptomatic or mild case. International Journal of Infectious Diseases 2013; [doi: 10.1016/j.ijid.2013.07.001]
15. Memish ZA, Cotten M, Watson SJ, Kellam P, Zumla A, Alhakeem RF, Assiri A, Rabeeah AAA, Al-Tawfiq JA. Community Case Clusters of Middle East Respiratory Syndrome Coronavirus in Hafr Al-Batin, Kingdom of Saudi Arabia: A Descriptive Genomic study. International Journal of Infectious Diseases 2014; PMID:24699184
16. Almekhlafi GA, Albarrak MM, Mandourah Y, Hassan S, Alwan A, Abudayah A, Altayyar S, Mustafa M, Aldaghestani T, Alghamedi A, Talag A, Malik MK, Omrani AS, Sakr Y. Presentation and outcome of Middle East respiratory syndrome in Saudi intensive care unit patients. Critical Care 2016; [doi: 10.1186/s13054-016-1303-8]
17. Alraddadi BM, Watson JT, Almarashi A, Abedi GR, Turkistani A, Sadran M, Housa A, Almazroa

MA, Alraihan N, Banjar A, Albalawi E, Alhindi H, Choudhry AJ, Meiman JG, Paczkowski M, Curns A, Mounts A, Feikin DR, Marano N, Swerdlow DL, Gerber SI, Hajjeh R, Madani TA. Risk factors for primary middle east respiratory syndrome coronavirus illness in humans, Saudi Arabia, 2014. Emerging Infectious Diseases 2016; [doi: 10.3201/eid2201.151340]

18. Kang CK, Song KH, Choe PG, Park WB, Bang JH, Kim ES, Park SW, Kim H bin, Kim NJ, Cho S il, Lee JK, Oh MD. Clinical and epidemiologic characteristics of spreaders of middle east respiratory syndrome coronavirus during the 2015 outbreak in Korea. Journal of Korean Medical Science 2017; [doi: 10.3346/jkms.2017.32.5.744]

19. Zhao J, Alshukairi AN, Baharoon SA, Ahmed WA, Bokhari AA, Nehdi AM, Layqah LA, Alghamdi MG, al Gethamy MM, Dada AM, Khalid I, Boujelal M, al Johani SM, Vogel L, Subbarao K, Mangalam A, Wu C, Eyck P ten, Perlman S, Zhao J. Recovery from the Middle East respiratory syndrome is associated with antibody and T cell responses. Science Immunology 2017; [doi: 10.1126/sciimmunol.aan5393]

20. Saad M, Omrani AS, Baig K, Bahloul A, Elzein F, Matin MA, Selim MAA, Mutairi M al, Nakhli D al, Aidaroos AYA, Sherbeeni N al, Al-Khashan HI, Memish ZA, Albarrak AM. Clinical aspects and outcomes of 70 patients with Middle East respiratory syndrome coronavirus infection: A single-center experience in Saudi Arabia. International Journal of Infectious Diseases 2014; [doi: 10.1016/j.ijid.2014.09.003]

21. Park HY, Lee EJ, Ryu YW, Kim Y, Kim H, Lee H, Yi SJ. Epidemiological investigation of MERS-CoV spread in a single hospital in South Korea, may to june 2015. Eurosurveillance 2015; [doi: 10.2807/1560-7917.ES2015.20.25.21169]

22. Sha J, Li Y, Chen X, Hu Y, Ren Y, Geng X, Zhang Z, Liu S. Fatality risks for nosocomial outbreaks of Middle East respiratory syndrome coronavirus in the Middle East and South Korea. Archives of Virology 2017; [doi: 10.1007/s00705-016-3062-x]

23. Chowell G, Fenimore PW, Castillo-Garsow MA, Castillo-Chavez C. SARS outbreaks in Ontario, Hong Kong and Singapore: The role of diagnosis and isolation as a control mechanism. Journal of Theoretical Biology 2003; [doi: 10.1016/S0022-5193(03)00228-5]

24. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Annals of Internal Medicine. 2007. PMID:17938396

25. U.S. Census Bureau PD. Annual Resident Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for States and Counties: April 1, 2010 to July 1, 2018 [Internet]. 2019 [cited 2020 Mar 19]. Available from: https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/counties/asrh/cc-est2018alldata.csv

26. Website AFF. 2010 County Level Population Density [Internet]. 2010. [cited 2020 Mar 19]. Available from: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk

27. Centers for Disease Control and Prevention UD of H and HS. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA [Internet]. CdcGov. 2020 [cited 2020 Mar 19]. p. 664–671. [doi: 10.1111/j.1465-3362.2012.00432.x]

28. United States Department of Health and Human Services C for DC and P and NCI. 2005–2016 Database: National Program of Cancer Registries and Surveillance, Epidemiology, and End Results SEER*Stat Database: NPCR and SEER Incidence – U.S. Cancer Statistics Public Use Research Database with Puerto Rico, November 2018 submission (2005–20 [Internet]. 2019 [cited 2020 Mar 19]. Available from: https://www.cdc.gov/cancer/uscs/public-use/

29. United States Hypertension Estimates by County 2001-2009 | GHDx [Internet]. [cited 2020 Apr 3]. Available from: http://ghdx.healthdata.org/record/ihme-data/united-states-hypertension-estimates-county-2001-2009

30. United States Chronic Respiratory Disease Mortality Rates by County 1980-2014 | GHDx

[Internet]. [cited 2020 Apr 3]. Available from: http://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014

31. Minn 2010-2014 County Cancer Profiles [Internet]. [cited 2020 Apr 3]. Available from: https://pennstate.maps.arcgis.com/home/item.html?id=ab5ab6a44f124ecc876a9d7c9eaf859c

32. GeoJSON and KML data for the United States - Eric Celeste [Internet]. [cited 2020 Apr 3]. Available from: https://eric.clst.org/tech/usgeojson/

33. COVID-19/Coronavirus Live Updates With Credible Sources in US and Canada | 1Point3Acres [Internet]. [cited 2020 Apr 3]. Available from: https://coronavirus.1point3acres.com/

34. NYTimes. NYtimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. [Internet]. 2020 [cited 2020 Apr 1]. Available from: https://github.com/nytimes/covid-19-data

35. Fang L, Karakiulakis G, Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? The Lancet Respiratory Medicine 2020; [doi: 10.1016/s2213-2600(20)30116-8]

36. Jia X, Yin C, Lu S, Chen Y, Liu Q, Bai J, Lu Y. Two Things About COVID-19 Might Need Attention. Preprints 2020; [doi: 10.20944/preprints202002.0315.v1]

37. del Rio C, Malani PN. COVID-19-New Insights on a Rapidly Changing Epidemic. JAMA 2020; PMID:32108857

38. Bin S, Sun G, Chen CC. Spread of infectious disease modeling and analysis of different factors on spread of infectious disease based on cellular automata. International Journal of Environmental Research and Public Health 2019; [doi: 10.3390/ijerph16234683]

39. Chow N, Fleming-Dutra K, Gierke R, Hall A, Hughes M, Pilishvili T, Ritchey M, Roguski K, Skoff T, Ussery E. Preliminary Estimates of the Prevalence of Selected Underlying Health Conditions Among Patients with Coronavirus Disease 2019 — United States, February 12–March 28, 2020. MMWR Morbidity and Mortality Weekly Report [Internet] 2020 Apr 3 [cited 2020 Apr 4];69(13):382–386. [doi: 10.15585/mmwr.mm6913e2]

40. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. American Journal of Epidemiology 2019; PMID:31509183

41. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics 2001; [doi: 10.2307/2699986]

42. Richardson M, Dominowska E, Ragno R. Predicting clicks: Estimating the click-through rate for new ads. 16th International World Wide Web Conference, WWW2007 2007. [doi: 10.1145/1242572.1242643]

43. Pan B. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conference Series: Earth and Environmental Science 2018. [doi: 10.1088/1755-1315/113/1/012127]

44. Chang W, Liu Y, Xiao Y, Xu X, Zhou S, Lu X, Cheng Y. Probability Analysis of Hypertension-Related Symptoms Based on XGBoost and Clustering Algorithm. Applied Sciences [Internet] 2019 Mar 22;9(6):1215. [doi: 10.3390/app9061215]

45. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics 2017; [doi: 10.1093/bib/bbx044]

46. Zhu J, Pande A, Mohapatra P, Han JJ. Using Deep Learning for Energy Expenditure Estimation with wearable sensors. 2015 17th International Conference on E-Health Networking, Application and Services, HealthCom 2015 2015. [doi: 10.1109/HealthCom.2015.7454554]

47. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B. Deep Neural Networks for Acoustic Modeling in Speech Recognition. Ieee Signal Processing Magazine 2012; [doi: 10.1109/MSP.2012.2205597]

48. Alanazi HO, Abdullah AH, Qureshi KN. A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care. Journal of Medical Systems 2017; [doi: 10.1007/s10916-017-0715-6]

49. Guo J, Yang L, Bie R, Yu J, Gao Y, Shen Y, Kos A. An XGBoost-based physical fitness evaluation

model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring. Computer Networks Elsevier B.V.; 2019 Mar 14;151:166–180. [doi: 10.1016/j.comnet.2019.01.026]

50. Chen T. XGBoost : A Scalable Tree Boosting System.

51. Putatunda S, Rama K. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. ACM International Conference Proceeding Series 2018. [doi: 10.1145/3297067.3297080]

52. Ingram DD, Franco SJ. 2013 NCHS urban-rural classification scheme for counties. Vital and Health Statistics, Series 2: Data Evaluation and Methods Research 2014; PMID:24776070

53. Champernowne DG, Cowell FA. Economic Inequality and Income Distribution [Internet]. Cambridge University Press; 1998 [cited 2016 Feb 28]. Available from: https://books.google.com/books?hl=en&lr=&id=lk5cccSd-v4C&pgis=1ISBN:0521589592

54. Campano F, Salvatore D. Income Distribution: Includes CD. Income Distribution: Includes CD. 2006. [doi: 10.1093/0195300912.001.0001]ISBN:9780195300918

55. Bellù LG, Liberati P. Policy Impacts on Inequality. Welfare Based Measures of Inequality. The Atkinson Index. EASYPol 2006;

56. Coefficient of variation - Wikipedia [Internet]. [cited 2020 Apr 4]. Available from: https://en.wikipedia.org/wiki/Coefficient_of_variation#cite_note-Bellu2006-20

57. Forecasts of Total Deaths | CDC [Internet]. [cited 2020 Jun 22]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

58. auquan COVID-19 Dashboard [Internet]. [cited 2020 Jun 22]. Available from: https://covid19-infection-model.auquan.com/

59. Covid Act Now [Internet]. [cited 2020 Jun 22]. Available from: https://covidactnow.org/?s=54069

60. The University of Texas COVID-19 Modeling [Internet]. [cited 2020 Jun 22]. Available from: https://covid-19.tacc.utexas.edu/projections/

61. Wang L, Wang G, Gao L, Li X, Yu S, Kim M, Wang Y, Gu Z. Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States. 2020 Apr 29 [cited 2020 Jun 22]; Available from: http://arxiv.org/abs/2004.14103

62. Pei S, Shaman J. Initial Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US. medRxiv [Internet] Cold Spring Harbor Laboratory Press; 2020 Mar 27 [cited 2020 Jun 22];2020.03.21.20040303. [doi: 10.1101/2020.03.21.20040303]

63. COVID-19 Response, AdityaLab, Georgia Tech [Internet]. [cited 2020 Jun 22]. Available from: https://www.cc.gatech.edu/~badityap/covid.html#forecasting

64. Keskinocak P, Oruc Aglar BE, Baxter A, Asplund J, Serban N. The Impact of Social Distancing on COVID19 Spread: State of Georgia Case Study. medRxiv. 2020. [doi: 10.1101/2020.04.29.20084764]

65. team IC-19 health service utilization forecasting, Murray CJ. Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries [Internet]. medRxiv. Cold Spring Harbor Laboratory Press; 2020 Apr. [doi: 10.1101/2020.04.21.20074732]

66. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, Munday JD, Meakin S, Doughty EL, Chun JY, Chan Y-WD, Finger F, Campbell P, Endo A, Pearson CAB, Gimma A, Russell T, Flasche S, Kucharski AJ, Eggo RM, Funk S. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research [Internet] F1000 Research Ltd; 2020 Jun 1 [cited 2020 Jun 22];5:112. [doi: 10.12688/wellcomeopenres.16006.1]

Table 1: XGBoost Classification Training and Testing Details

Dataset	Evaluation Metrics	Mean Value	Minimum Value	Maximum Value	Standard Deviation	Number of Days
Test	Accuracy	83%	77%	92%	5%	18
	AUC	78%	71%	83%	3%	18
Train	Accuracy	94%	82%	100%	5%	18
	AUC	91%	80%	100%	6%	18

Table 2: Sensitivity of the three-stage Model

Date	Number of 5% Most Vulnerable Counties Identified on a Given Date (with 0 confirmed case)	Number of Counties that reported cases after 5 Days	Sensitivity
3/14/2020	92	61	66.30%
3/15/2020	119	90	75.63%
3/16/2020	151	99	65.56%
3/17/2020	199	144	72.36%
3/18/2020	144	110	76.39%
3/19/2020	176	115	65.34%
3/20/2020	198	146	73.74%
3/21/2020	166	125	75.30%
3/22/2020	158	120	75.95%
3/23/2020	84	66	78.57%
3/24/2020	89	65	73.03%
3/25/2020	336	208	61.90%
3/26/2020	104	72	69.23%

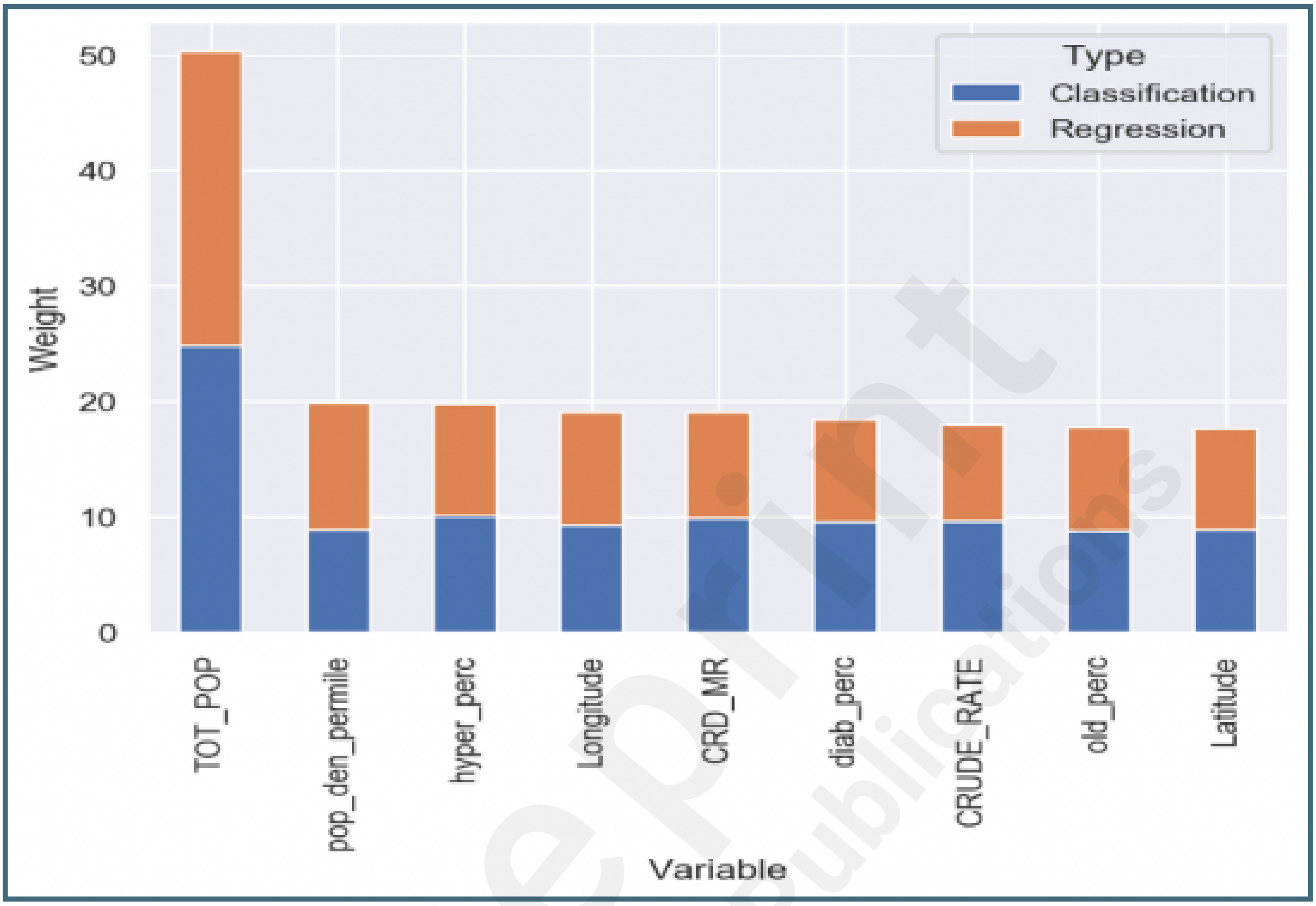
Table 3: Specificity of the three-stage Model

Date	Number of Top 10% Least Vulnerable Counties Identified on a Given Date (0 confirmed case)	Number of Counties with 0 case after 5 Days	Specificity
3/14/2020	276	274	99.28%
3/15/2020	282	276	97.87%
3/16/2020	46	44	95.65%
3/17/2020	313	304	97.12%
3/18/2020	297	281	94.61%
3/19/2020	214	198	92.52%
3/20/2020	295	266	90.17%
3/21/2020	312	291	93.27%
3/22/2020	15	14	93.33%
3/23/2020	310	289	93.23%
3/24/2020	303	270	89.11%
3/25/2020	214	197	92.06%
3/26/2020	231	218	94.37%

Supplementary Files

Figures

Variable Importance for the Classification and Regression Models.

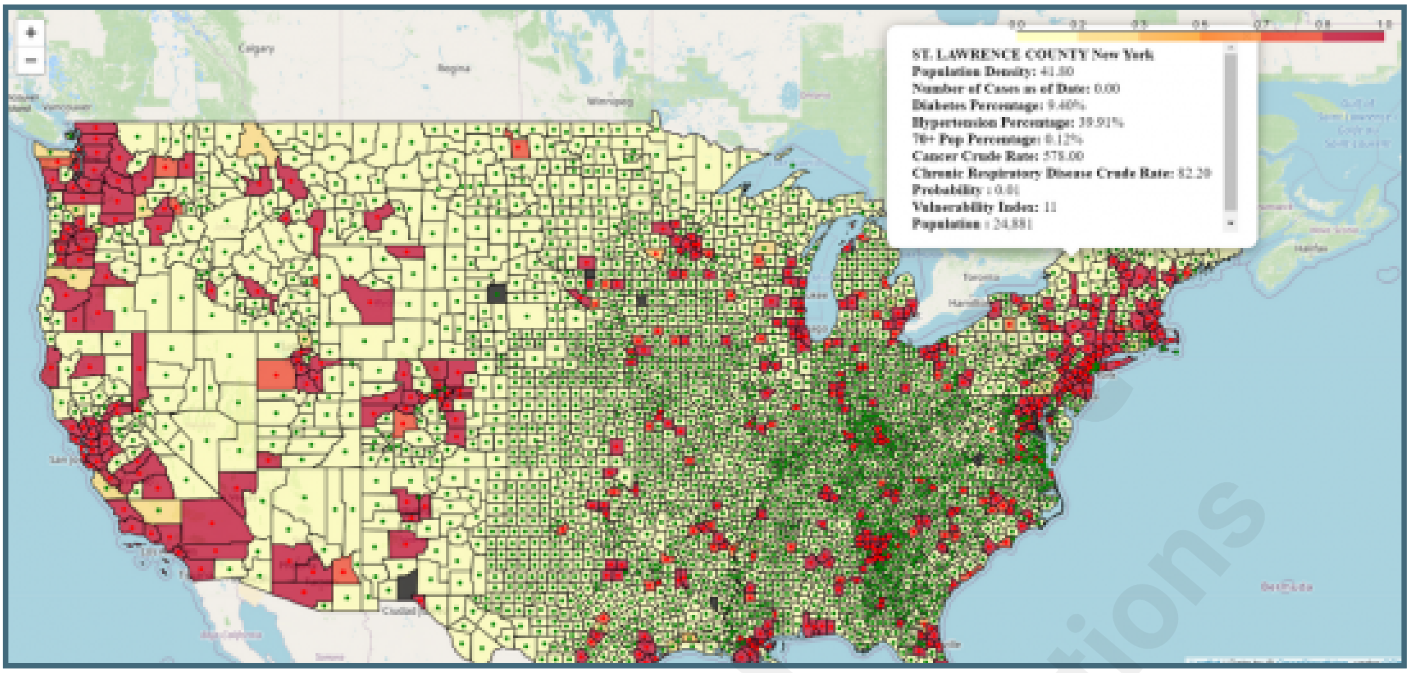


Multimedia Appendixes

Supplementary Materials.
URL: https://asset.jmir.pub/assets/64ccb589218b4dabacaebec2463683cc.docx

Figures

Predicted probability of being a positive instance for each county in the US.



ST. LAWRENCE COUNTY New York
Population Density: 41.80
Number of Cases as of Date: 0.00
Diabetes Percentage: 9.40%
Hypertension Percentage: 39.91%
70+ Pop Percentage: 0.12%
Cancer Crude Rate: 578.00
Chronic Respiratory Disease Crude Rate: 82.20
Probability : 0.01
Vulnerability Index: 11
Population : 24,881