Evaluation of a Parsimonious COVID-19 Outbreak Prediction Model using publicly available datasets

Agrayan Kishan Gupta, Shaun Grannis, Suranga Kasthurirathne

Submitted to: Journal of Medical Internet Research
on: March 23, 2021

Table of Contents

Evaluation of a Parsimonious COVID-19 Outbreak Prediction Model using publicly available datasets

Agrayan Kishan Gupta[1]; Shaun Grannis[2, 3] MD, MS, FAAFP; Suranga Kasthurirathne[2, 3] PhD

[1]Indiana Univeristy Bloomington US
[2]Center for Biomedical Informatics Regenstrief Institute Indianapolis US
[3]Indiana University School of Medicine Indianapolis US

Corresponding Author:
Agrayan Kishan Gupta
Indiana Univeristy
107 S Indiana Ave
Bloomington
US

Abstract

Background: Coronavirus disease 2019 (COVID-19) pandemic has changed public health policies and personal lifestyles through lockdowns and mandates. Governments are rapidly evolving policies to increase hospital capacity and supply personal protective equipment to mitigate disease spread in distressed regions. Current models that predict COVID-19 case counts and spread, such as deep learning, offer limited explainability and generalizability. This creates a gap for highly accurate and robust outbreak prediction models which balance parsimony and fit.

Objective: We seek to leverage various readily accessible datasets extracted from multiple states to train and evaluate a parsimonious predictive model capable of identifying county-level risk of COVID-19 outbreaks on a day-to-day basis.

Methods: Our methods use the following data inputs: COVID-19 case counts per county per day and county populations. We developed an outbreak gold standard across California, Indiana, and Iowa. The model was trained on data between 3/1/20-8/31/20, then tested from 9/1/20 to 10/31/20 against the gold standard to derive confusion matrix statistics.

Results: The model reported sensitivities of 92%, 90%, and 81% for Indiana, Iowa, and California respectively. The precision in each state was above 85%, and the specificity and accuracy were generally greater than 95%.

Conclusions: The parsimonious model provide a generalizable and simple alternative approach to outbreak prediction. Our methodology could be tested on diverse regions to aid government officials and hospitals with resource allocation.

(JMIR Preprints 23/03/2021:28812)
DOI: https://doi.org/10.2196/preprints.28812

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title: A Parsimonious COVID-19 Outbreak Prediction Model

Abstract (250)

The coronavirus 19 (COVID-19) pandemic has changed public health policies and personal lifestyles through lockdowns and mandates. Governments are rapidly evolving policies to increase hospital capacity and supply personal protective equipment to mitigate disease spread in distressed regions. Current models that predict COVID-19 case counts and spread, such as deep learning, offer limited explainability and generalizability. This creates a gap for highly accurate and robust outbreak prediction models which balance parsimony and fit. This paper presents a novel rule-based COVID-19 outbreak prediction model that utilizes minimal data input. Using case counts per county per day and county populations, the model was trained and tested across Indiana, Iowa, and California. The model's precision in each state was above 85%, and the specificity and accuracy were both generally greater than 95%. Based on these results along with COVID-19's unique epidemiological impact, we propose a rule-based model as a simple and generalizable approach to identify outbreaks.

Introduction

COVID-19 cases continue to rise even as governments and local authorities make efforts to contain the pandemic. Second and third waves of viral outbreaks place health systems under extended strain, especially healthcare workers who have worked at maximum capacity for months. COVID-19 has led to hospital bed shortages, personal protective equipment shortages, and other healthcare disruptions[1].

To combat this, scientists have quickly expanded scientific literature to develop a deeper understanding of COVID-19 virus. The rapidly evolving information landscape has led governments to adapt policies in response to the changing circumstances. In the last decade, public health has expanded data sources to track influenza activity, providing a complementary public health tool along with real-time testing [2]. Since the beginning of 2020, data scientists have collaborated with governmental organizations to create public-facing dashboards to provide easy access information on COVID-19 statistics and other incidence metrics. Despite advances in information exchange, gaps exist in current research to understand and predict disease pathology and transmission. For example, there is a lack of uniformity in data reporting techniques across states; there is limited epidemiological information on the disease; and published public health data are being retrospectively adjusted.

To cope, predicting the amount of cases or whether areas are experiencing COVID-19 outbreaks allows for the proper allocation of limited medical resources. Any outbreak is difficult for a government entity as they adapt hospital management and capacity to meet emergent demands. Predicting outbreaks enables decision makers across governmental and healthcare sectors to devote constrained resources, allowing patient needs to be met. Analytical methods can be applied to predict potential outbreaks. Some of these methods such as deep learning and complex may yield superior results, but offer limited explainability and generalizability especially with data source variability[3]. In contrast, a less complex modelling approach uses selective and sufficient data to promote brevity and offer partial explanations to fit across a variety of settings[4].

Parsimonious models tend to underfit tested data sets as they tend to better discriminate the signal from noise, allowing for greater generalizability and functionality[5]. Other studies during COVID-19 have also explored predicting case counts through parsimonious approaches, for example exponential growth and compartmental models[6]. Our data-driven approach incorporates inputs from a state's COVID-19 cases per day per county and a state's population. The novelty of our approach from a methodological perspective is its usage of minimal parameters to effectively predict outbreaks.

We leveraged a variety of readily accessible county level data elements to train and evaluate heuristic, rule-based models capable of predicting risk of outbreaks. We utilized a minimum number of predictor variables to output a duplex indicator tested across a curated gold standard.

Methods

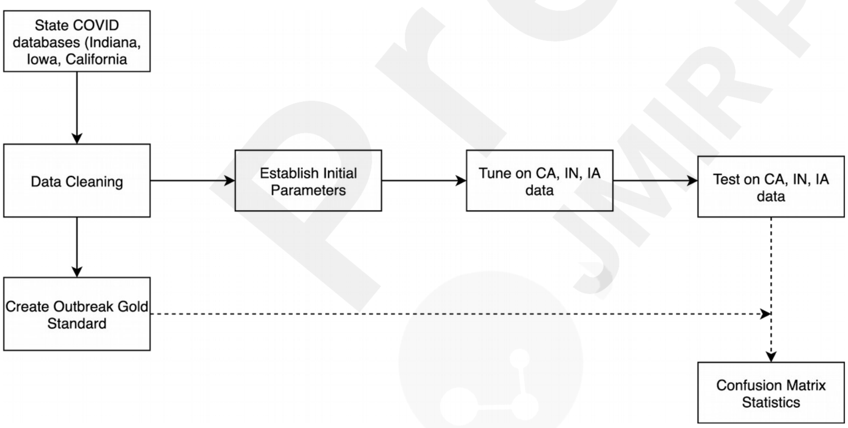


Figure 1: Depicted is the flow diagram for an overview of the study methodology. Data from Indiana, Iowa, and California was cleaned to establish an outbreak gold standard. Concurrently, model rules were developed and tuned on each state's data from March 1st 2020 to August 31st 2020. Confusion matrix statistics were derived by comparing the model's results on the test data set from September 1st 2020 to October 31st 2020 to the outbreak gold standard.

Data Cleaning:

We developed models for three states: California, Indiana, and Iowa. These states were selected based on geographical factors, governmental regulations, and data availability. We also considered the quality of datasets through state tracking systems, basic data sources, and general completeness of reporting[7]. Indiana and Iowa are similar in county sizes and population, whereas California allowed us to evaluate model generalizability in a coastal and more populous state[8]. We extracted county-level datasets obtained from the Indiana State Department of Health and New York Times between March 1st, 2020 and October 31st, 2020. Datasets were organized by county, state, and day using R statistical software[9 10]. Days with negative case counts were changed to 0 as they most likely were due to errors or omissions in data collection. Additionally, we removed an “unknown” county category from the Iowa and California data sets.

Gold Standard Preparation:

A gold standard was created to identify whether each county in Indiana, Iowa, and California was in an outbreak on any given day.

To do so a, human expert reviewer considered factors listed below to label outbreak ranges:

1. How case counts trended in each county, and if there was a general baseline of cases over time.
2. A county’s population size as counties with more people experienced more cases.
3. Duration of outbreak to assign a binary indicator of ‘outbreak detected’ or ‘outbreak not detected’ to each day and county.

Based on our approach, a county could have multiple outbreaks separated by various time windows. Outbreaks lasted a minimum of three days to account for testing lags as data was not always reported on the same day, especially during the initial phases of the pandemic. Furthermore, lower case counts at the tails of an outbreak and on weekends due to closed testing centers were considered.

Model Building:

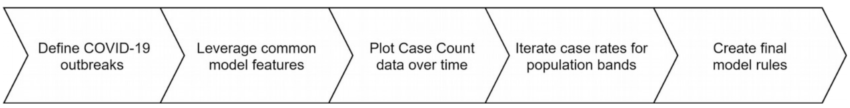


Figure 2: We underwent an iterative process to develop the model rules. We plotted case count data by population to establish a generalized trendline. By using features from other models and defining outbreak parameters, we constructed a rough model. The rules were trained and finalized on the train data for the test data set.

We created a rule-based outbreak prediction model using the training datasets obtained from all three states, and evaluated its performance across the holdout test datasets and gold standard. To begin building the model, we evaluated features from other common and published models.

The State of Wisconsin's COVID-19 dashboard used a Case Rate metric, defined as a per capita running 7 day sum of the case counts per county per day[11]. Case Rate is a metric that standardizes COVID-19 severity across counties of differing populations while also accounting for data lags and providing insight on transmission. Furthermore, predictive models for infectious diseases such as SEIRs provide guidance on disease transmission and outbreak causation. We plotted Case Rate vs. Indiana county populations in R to generate a general trendline that could differentiate between 'outbreak detected' or 'outbreak not detected' days. Our logarithmic graph semi-accurately depicted a horizontal line that separated outbreak days. The next steps were to leverage and apply the trendline results on states and counties with various populations.

We started building the model by dividing counties based on population size, initially at 100,000 intervals. Since Case Rate is more sensitive to less populated counties, we added more intervals for counties under 100,000 people. Each population interval was given a Case Rate baseline value that served as a binary indicator for outbreak determination. Additionally, to account for data lag, we implemented a criteria where counties were under outbreak if they were 4 standard deviations above the cumulative mean. As depicted in the system flow diagram in figure 1, we established these parameters values and trained the model rules on the state data sets between March 1st and August 31st 2020. The train to test partition was roughly 71% to 29%, respectively, which is described as close to optimal for large datasets[12]. Then, the model was tested against the gold standard on each state between September 1st and October 31st 2020. The confusion matrix statistics between the test data set results and gold standard are discussed in the following section.

Results

	Indiana	Iowa	California
Number of Counties	92	99	58
Counties below 100k	75	93	23
Counties between 100k and 500k	16	6	19
Counties between 500k and 1mil	1	0	7
Counties above 1 mil	0	0	9
Smallest County	5875	3602	1129
Largest County	964582	490161	10039107
% Urban Population	72.44%	64.02%	94.95%
Household Median Income	59892	68718	70489
Case Count Average	7.98	6.18	70.33
Case count St. Dev	21.02	15.75	231.857

	Metropolitan				Small Town				Rural		
RUCA Codes	1	2	3	4	5	6	7	8	9	10	Average
CA	7139	325	53	254	31	8	58	16	2	151	1.397163
IA	319	96	5	80	36	5	94	34	3	152	4.287621
IN	929	185	28	129	72	14	67	10	17	57	2.397215

Table 1a and 1b: Table 1 shows state county population sizes and population statistics based on census counts. Table 2 breaks down each state's RUCA Code classification, which are counted by census tracts. The codes are grouped at a higher level into Metropolitan, Small Town, and Rural. (Source: US Census Bureau, USDA)

Table 1a presents descriptive statistics, including number of counties, population sizes, and urbanization, to highlight each state's fundamental differences. Indiana and Iowa have similar county population distributions, with both having a majority of counties below 100,000 citizens. However, Indiana has more midsize counties with its largest close to 1 million people, while California has several counties with populations greater than 1 million. Moreover, California has the largest percentage of urban population (94.95%), with Indiana (and Iowa far behind. California's urbanization is further explored in Table 1b as a majority of its census tracts have metropolitan RUCA codes. RUCA codes classify a state's census tracts into Metropolitan, Small Town, and Rural areas based on population density and daily commuting[13]. Indiana is slightly more urbanized than Iowa.

	Indiana			Iowa			California		
Date Range	3-1 to 8-31	9-1 to 10-31	3-1 to 10-31	3-1 to 8-31	9-1 to 10-31	3-1 to 10-31	3-1 to 8-31	9-1 to 10-31	3-1 to 10-31
	Train	Test	Total	Train	Test	Total	Train	Test	Total
Number of Outbreaks	26	65	83	43	85	114	35	26	40
Average Duration	25	19.18	22.86	18.175	14.622	15.787	47.29	19.31	53.92
Total Days in Outbreak	650	1247	1897	727	1199	1926	1655	502	2157
Outbreak Days as %	3.86%	22.59%	8.45%	4.01%	20.15%	7.97%	15.59%	14.43%	5.24%

Table 2: COVID-19 outbreak prevalence descriptors from the gold standard. Indiana, Iowa, and California datasets were divided by train, test, and total date ranges to characterize the outbreak periods.

Table 2 showcases the prevalence of the number of outbreaks and their durations in each state over the train, test, and total time periods. In Indiana and Iowa, the number of outbreaks doubled from the train date range to the test range, despite the train data set being almost three times as large as the test data set. Furthermore, the percentage of days in an outbreak quadrupled to 22.6% and 20.1% for Indiana and Iowa, respectively, between the train and test ranges. The percentage of outbreak days in California stayed relatively the same and the average outbreak duration decreased from 47 to 19 days. Fewer counties in California accounted for the lower number of outbreaks during the overall time period.

Model Rules:

Using the above data, we generated model parameters to predict COVID-19 outbreaks. Rules and population bands are further outlined in Figure 3 and Appendix 1.

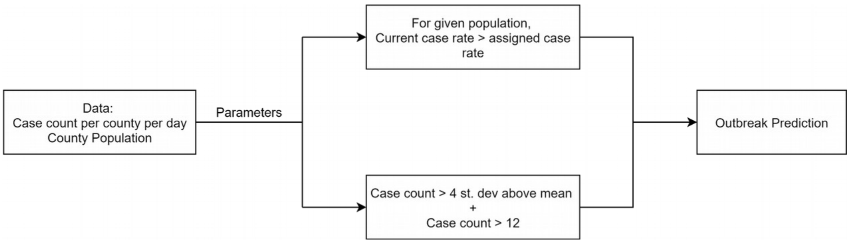


Figure 3: COVID-19 case counts per county per day and county population data was pulled. Using the model rules, counties were defined to be in outbreak through two ways. First, depending on a county's population, it was assigned a Case Rate, defined as a per capita running 7 day sum of the case counts per county per day, for its population band. If the current case rate was greater than the assigned case rate, the county was predicted as under an outbreak. The other method is if the case count on a given day was greater than 12 and four standard deviations above the cumulative mean.

Using the aforementioned population groupings, counties experienced an outbreak through the following methods:

1. For the specified population band, a county's Case Rate on a given day was greater than the minimum case rate assigned to that population
2. The county's case count on a specific day was greater than 12 and 4 standard deviations above the rolling COVID-19 county case count average

By combining these rules with the gold standard previously developed, a confusion matrix was utilized to provide an analysis on the model's performance.

As shown in Table 3, the rule-based prediction model was tested on the test date range.

	Indiana	Iowa	California
Test Date Range	9-1 to 10-31	9-1 to 10-31	9-1 to 10-31
Sensitivity	92.33%	90.05%	80.86%
Specificity	95.56%	97.40%	99.57%
Precision	85.04%	89.83%	96.96%
Accuracy	94.86%	95.91%	96.85%

Table 3: Confusion matrix results when the prediction model was applied to the curated gold standard during the test data date range from September 1st to October 31st 2020. Indiana and Iowa had similar results, while California maintained high accuracy and precision with lower sensitivity.

Almost all four key confusion matrix statistics -- sensitivity, specificity, precision, and accuracy-- were above 80% in each state during the test range. The specificity and accuracy consistently stayed very high in each state, with both above 94%. This was largely attributed to most days being classified as true negatives, which are fundamentally easier to detect than true positives. Indiana and

Iowa both had sensitivity metrics 10% greater than California, however their precision results were respectively 11% and 7% lower. For Indiana and Iowa, this means the model computed fewer false negative readings, which could be attributed to having more and longer outbreaks, as shown with Table 2. California's higher precision but lower specificity means the model was more precise in predicting when outbreaks happened, but was less successful in capturing all outbreaks.

Discussion

Our efforts resulted in a heuristic model capable of detecting COVID-19 outbreaks with strong predictive performance. The model reported sensitivities of 92%, 90%, and 81% for Indiana, Iowa, and California respectively. Specificity and accuracy stayed extremely high in all states indicating that the model excelled in predicting when there wasn't an outbreak.

As the pandemic progressed, states enhanced their data reporting systems. As described by Khakharia et al., some regions had sudden significant drops and rises in case counts, making it difficult for models to forecast future cases[14]. Though outbreaks are not fundamentally different, the train and test data sets can be characterized separately. Despite the test range being shorter, Indiana and Iowa both had twice as many outbreaks. During this period, states experienced a second wave of COVID-19 cases as schools started, governors relaxed state lockdown laws, and citizens returned to work[15]. For example, as a liberal state, California was one of the last states to begin lifting restrictions in midsize and large counties, contributing to why they didn't have a similar rise in outbreaks as Indiana and Iowa[16 17]. In result, counties reentered or for the first time realized outbreak periods. Additionally, cases counts were typically higher during the test period than the train period.

California remains a state of interest due to its outbreak characterization and confusion matrix results. Unlike Indiana and Iowa, California has several counties with populations over 1,000,000 people, and furthermore, it was the only state with less outbreaks and percentage of days in outbreak between the train and test periods. In the results, California had a significantly lower sensitivity but a higher precision value. Respective to other states, the model was worse in capturing all of California's outbreaks but more precise in predicting them, similar to a quality over quantity argument.

This parsimonious rule-based prediction model is easily replicable in other states, as it only utilizes county population and COVID-19 cases per day per county data. Using those figures, along with following the model's rules, states would be able to detect and predict outbreaks with high accuracy. Current outbreak prediction approaches center around machine learning algorithms.

Though they generally have very high accuracies, these models incorporate a variety of data points and can overfit models. The rule-based model's data simplicity enables it to be easily implemented in other regions, especially those with limited reported systems. It is also an understandable and accurate method to relay a county's current state of COVID-19 to the general public, who are not as informed in health metrics.

Limitations

Our work was impacted by limitations with the current data collecting systems as well as the rule-based approach. The inconsistency of data reporting presents a significant systematic challenge for model building. For instance, states closed most COVID-19 testing centers on weekends, leading to lower case count values for roughly 28% of the week. Further, many states did not publish most of their raw COVID-19 data, meaning we pulled cases per day per county data from the New York Times instead of a state's Department of Health which is more accurate. NYT would retroactively change case data, making it more unreliable since there were days with negative values. We noticed data from Indiana, downloaded from the Indiana State Department of Health, was easier to access and more complete in all regards.

The lack of prior research on gold standard curation also presents limitations. With no industry standard on defining an outbreak, we created the gold standard based on intuition and the specified criteria outlined above. Therefore, this process may be subject to biases and confounding that may have influenced our model's results. However, the gold standard was generated prior to any model development, limiting any causational effects. Furthermore, the rule-based model approach is subject to several limitations. Since the model incorporated a 7-day moving Case Rate, there was a lag at the tails of outbreaks as the increased case counts were not initially detected. Even with a parsimonious approach, the parameters derived from our results can greatly differ when applied to another region. This uncertainty, formed through parameters, social mandates, and vaccination, is a feature of any prediction model. We help lessen this uncertainty through our generalizable approach demonstrated on diverse states.

Future Work

The ongoing global pandemic has led to most major institutions allocating tremendous resources for its resolution. The model would benefit from a larger sample size of US states, and possibly international regions, to test generalizability on a more expansive scale. Additionally, the model's data range could be expanded to the third wave of cases and as the COVID-19 vaccine is distributed to a majority of the populace to determine its functionality past the study's scope. Study results could also be translated to provide a clearer outlook of epidemiological diseases. Since the model can predict outbreaks with high accuracy, it could be tested on historical COVID-19 data to easily determine when most outbreaks occurred in a region. Moreover, trends and patterns could be found across outbreaks between various factors such as lockdown policies, population density, and civilian obedience. Understanding outbreak causation presents interesting research on public policy adaptation in current and future situations.

Conclusion

This paper presents an accurate, generalizable, and explainable COVID-19 outbreak prediction model. During the test date range, the model's sensitivity was above 90% in Indiana and Iowa, and 80% in California. Furthermore, the specificity and accuracy were greater than 94% in every state. These results coupled with the minimal data inputs create an understandable and easy to implement model that governments and policy-makers can utilize to assess COVID-19 severity in diverse regions. There should be future work on testing the model in more states and countries using more recent data. Moreover, the model should be used to identify outbreaks to investigate correlations between external factors and outbreak development.

Appendix

Population Band Upper Limit	Min Case Rate
7500	450
10000	400
15000	340
20000	275
50000	225
100000	125
350000	115
500000	95
750000	85
1000000	75
2000000	65
5000000	55
10000000	45

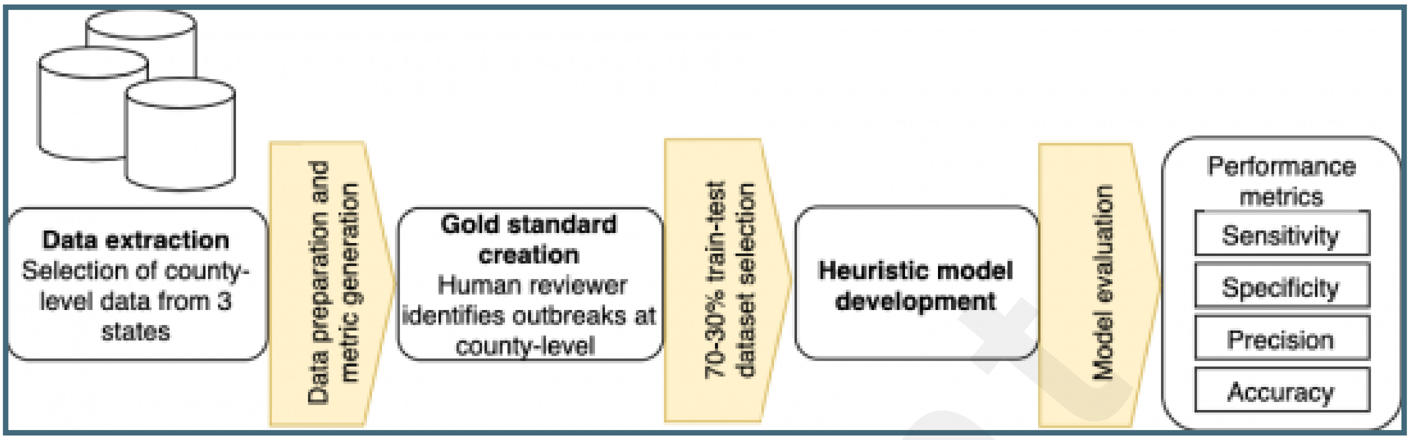
1. Liu Q, Luo D, Haase JE, et al. The experiences of health-care providers during the COVID-19 crisis in China: a qualitative study. Lancet Glob Health 2020;**8**(6):e790-e98 doi: 10.1016/S2214-109X(20)30204-7[published Online First: Epub Date]|.

2. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc Natl Acad Sci U S A 2015;**112**(47):14473-8 doi: 10.1073/pnas.1515373112[published Online First: Epub Date]|.
3. Saez C, Romero N, Conejero JA, Garcia-Gomez JM. Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset. J Am Med Inform Assoc 2021;**28**(2):360-64 doi: 10.1093/jamia/ocaa258[published Online First: Epub Date]|.
4. Sokol K, Flach P. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv191205100S (accessed December 01, 2019).
5. Vandekerckhove J, Matzke D, Wagenmakers E-J. Model comparison and the principle of parsimony. The Oxford handbook of computational and mathematical psychology. New York, NY, US: Oxford University Press, 2015:300-19.
6. Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proc Natl Acad Sci U S A 2020;**117**(29):16732-38 doi: 10.1073/pnas.2006520117[published Online First: Epub Date]|.
7. State Reporting Assessments. Secondary State Reporting Assessments. https://covidtracking.com/about-data/state-reporting-assessments.
8. County Population Totals: 2010-2019. Secondary County Population Totals: 2010-2019. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html.
9. COVID-19 County Statistics
. Secondary COVID-19 County Statistics
. https://www.in.gov/mph/1194.htm.
10. Coronavirus (Covid-19) Data. Secondary Coronavirus (Covid-19) Data. https://developer.nytimes.com/covid.
11. COVID-19: Disease Activity by Region and County. Secondary COVID-19: Disease Activity by Region and County 2021. https://www.dhs.wisconsin.gov/covid-19/disease.htm.
12. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics 2011;**4**:31 doi: 10.1186/1755-8794-4-31[published Online First: Epub Date]|.
13. Rural-Urban Commuting Area Codes. Secondary Rural-Urban Commuting Area Codes August 17, 2020. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx.
14. Khakharia A, Shah V, Jain S, et al. Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning. Annals of Data Science 2021;**8**(1):1-19 doi: 10.1007/s40745-020-00314-9[published Online First: Epub Date]|.
15. Maragakis LL. Coronavirus Second Wave? Why Cases Increase. Secondary Coronavirus Second Wave? Why Cases Increase 2020. https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus.
16. Reclosing Plans and Mask Requirements. Secondary Each State's COVID-19 Reopening and Reclosing Plans and Mask Requirements February 22, 2021. https://www.nashp.org/governors-prioritize-health-for-all/.
17. Li Y, Zhang R, Zhao J, Molina MJ. Understanding transmission and intervention for the COVID-19 pandemic in the United States. Sci Total Environ 2020;**748**:141560 doi: 10.1016/j.scitotenv.2020.141560[published Online First: Epub Date]|.

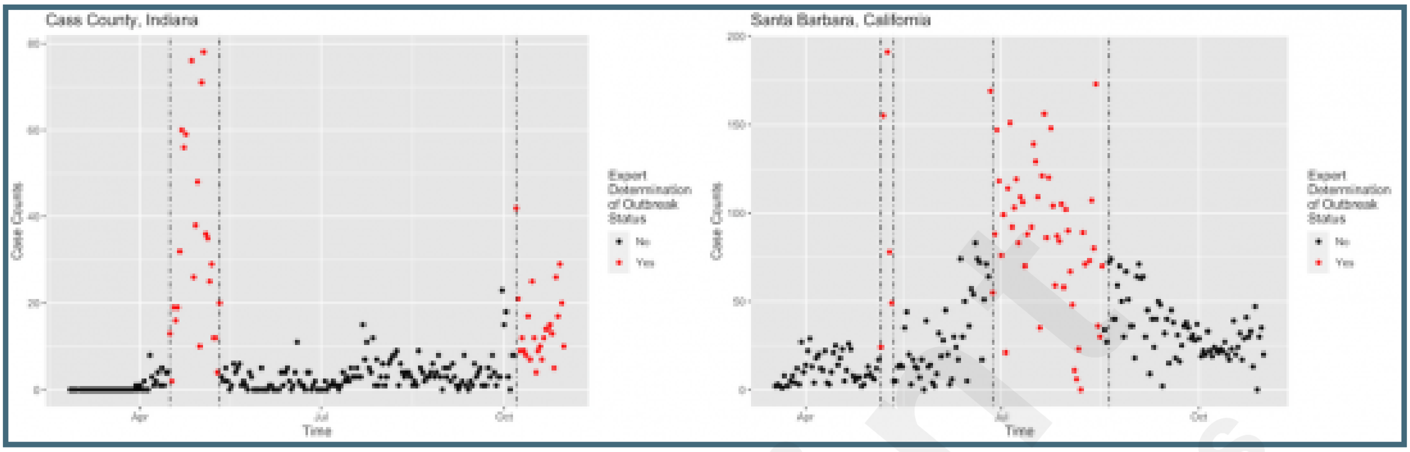
Supplementary Files

Figures

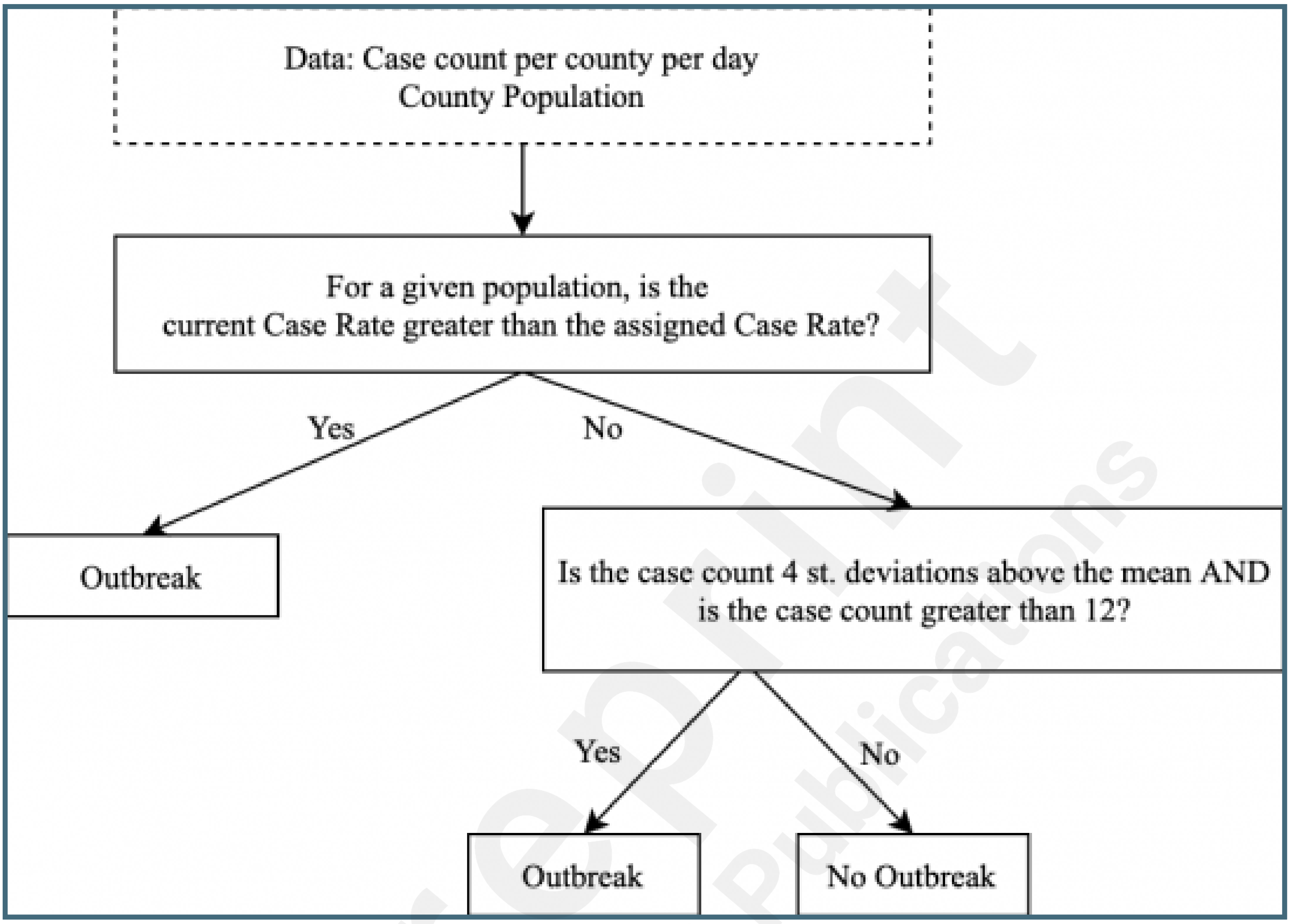
Flow diagram for an overview of the study methodology.



Visualization of the COVID-19 case counts in Cass County, Indiana, and Santa Barbara County, California, between March 1st, 2020 and October 31st, 2020. Days determined to be outbreak are colored red while normal days are black.



Heuristic model decision making process.



Multimedia Appendixes

Population bands with their respective minimum case rates.
URL: http://asset.jmir.pub/assets/20b5f12d2979baf181588e15e07840ec.xls

Outbreak gold standard for each county in California, Indiana, and Iowa per day.
URL: http://asset.jmir.pub/assets/8d7a8130cf6166b763a303152b6388b3.xlsx