Evaluation of a Parsimonious COVID-19 Outbreak Prediction Model: Heuristic Modeling Approach using publicly available datasets

Agrayan Kishan Gupta, Shaun Grannis, Suranga Kasthurirathne

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

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

Evaluation of a Parsimonious COVID-19 Outbreak Prediction Model: Heuristic Modeling Approach 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

Original Manuscript

Original Paper

Agrayan K Gupta[1]; Shaun J Grannis[2,3], MD, MS; Suranga N Kasthurirathne[2,3], PhD

[1]Indiana University, Indianapolis, Indiana, USA
[2]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, Indiana, USA
[3]Indiana University School of Medicine, Indianapolis, Indiana, USA

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

Abstract (450)

Background:
Coronavirus disease 2019 (COVID-19) pandemic has changed public health policies and human and community behaviors through lockdowns and mandates. Governments are rapidly evolving policies to increase hospital capacity and supply personal protective equipment and other equipment to mitigate disease spread in affected regions. Current models that predict COVID-19 case counts and spread are complex by nature and offer limited explainability and generalizability. This has highlighted the need for accurate and robust outbreak prediction models which balance model parsimony and performance.

Objective:
We sought to leverage 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. We sought to leverage 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 modeling approach leveraged 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 utilized a per capita running seven-day sum of the case counts per county per day and cumulative case count mean to develop baseline values. The model was trained on data recorded between 3/1/20 to 8/31/20, and tested on data recorded between 9/1/20 to 10/31/20.

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

Conclusion:
Our parsimonious model provides a generalizable and simple alternative approach to outbreak prediction. This methodology could be applied to diverse regions to aid state officials and hospitals with resource allocation, and to guide risk management, community education and mitigation strategies.

Keywords: COVID-19; predictive modeling; coronavirus; modeling disease outbreak; precision public health; emerging outbreak

Introduction

The Coronavirus disease 2019 (COVID-19) pandemic has impacted the health and wellbeing of individuals, communities, and economies across the globe at a hitherto unprecedented scale[1-3]. On March 11th, 2020, the World Health Organization (WHO) declared COVID-19 as a worldwide pandemic with over 118,000 confirmed cases and 4,291 deaths in over 114 countries[4]. To date, the pandemic has resulted in over 170 million confirmed cases, with over 3.5 million deaths globally[5]. In the United States, COVID-19 has infected 33 million people and claimed more than 600,000 lives[5].

At the height of the pandemic, waves of viral outbreaks placed health systems across the globe under extended strain leading to shortages in hospital beds, personal protective equipment, and healthcare personnel, causing significant disruptions to healthcare delivery and loss of life[2 6]. Experts estimate the cumulative financial costs of the COVID-19 pandemic related to lost output and health reduction at $16 trillion, or approximately 90% of the annual gross domestic product of the US[7].

In contrast to historical pandemics, the availability of public and population health information systems has enabled researchers to collaborate on many research activities in response to COVID-19[8 9]. Since the onset of the pandemic, data scientists have partnered with governmental organizations to create various public-facing COVID-19 dashboards that provide easy access to descriptive statistics and other metrics[10 11]. Information on COVID-19 related mortality, utilization of healthcare resources, and recovery has been crucial in increasing situational awareness to inform ongoing pandemic response efforts across communities[12 13].

Most recently, COVID infection rates have started to fall in response to increased vaccinations and public education efforts[14 15]. To date, 40% of the US population is fully vaccinated[16]. These improvements have led to an interest in relaxing or revoking various restrictions enforced at state and county levels. While important to the well-being of both communities and economies, such decisions may be dangerous if undertaken without adequate pre-planning and awareness of potential risks. As such, effective identification of potential outbreaks of COVID-19 offers the ability to inform decision-makers across governmental and public health sectors on how to open up their communities to normal day-to-day life activities and deploy limited human and treatment resources to where they are most needed[17].

Prior research has demonstrated the potential to apply analytical models to identify potential outbreaks in response to other diseases[18]. However, these methods rely on large, complex datasets extracted from a specific health system or region[19-21]. Such datasets may be challenging and time-consuming to collect, leading to delays in generating timely predictions. Further, models trained using locale-specific datasets may not be generalizable across other locations[22], hindering the potential of re-using such models across other patient populations and regions. A variety of models are trained using complex algorithmic approaches such as neural networks and deep learning models. Such machine learning approaches may yield superior results but fail to achieve widespread acceptance[23] due to challenges in explainability and interpretation[24].

In contrast, a less complex modeling approach that uses a subset of easily obtainable key elements widely captured across broad geographic regions may be less challenging to develop. Further, such models may also deliver adequate predictive performance without sacrificing explainability and interpretability. Such parsimonious models may also present less risk of overfitting on training datasets, allowing for greater generalizability[25].

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

We selected three states for our COVID-19 outbreak prediction modeling efforts: California, Indiana, and Iowa. These states were selected based on geographical factors, governmental regulations, and availability of datasets for public use. For example, Indiana and Iowa are similar in number of counties and total populations[26]. In contrast, California represented a more populous, urban state[26]. We also considered the general completeness of reporting, the quality of basic COVID-19 data sources, and the accuracy of state tracking systems[27].

Data extraction and cleaning:

For each state, we extracted a variety of county-level data elements captured daily between March 1st, 2020 and October 31st, 2020. Data for the state of Indiana were obtained from the Indiana State Department of Health, while data for the states of Iowa and California were obtained from the New York Times online repository[5 28]. We selected March 1st as a start date as most states began collecting COVID-19 data at this time. October 31st marked the end of our analysis time period. Each dataset was organized by county, state, and date reported using R programming software[29]. Several errors or omissions in the datasets were addressed as follows; days with negative case counts were changed to 0 and a county labeled as ‘unknown’ reported by Iowa and California were removed from further evaluation.

Preparation of a Gold Standard:

We created a gold standard indicating if each county under study was at an outbreak on any given day. A human expert reviewer created the gold standard by assigning an outbreak label (yes/no) to each county/date combination considering the following criteria:

- How do case counts trend in each county? Is there a general baseline of cases over time?
- How large is the county’s population size (counties with more people report more cases)
- 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 over time. Outbreaks lasted a minimum of three days to account for testing lags as data were not always reported on the same day, especially during the initial phases of the pandemic[30]. Furthermore, lower case counts at the end of an outbreak and on weekends due to closed testing centers were also taken into consideration using seven-day average metrics.

Model Building:

We created a heuristic outbreak prediction model using the training datasets obtained from all three states and evaluated its performance across the holdout test datasets and the gold standard. For each county, data collected between March 1st, 2020 and August 31st, 2020 were labelled as the train dataset, while data collected between September 1st, 2020 and October 31st,2020 were labeled as the test dataset. As a preliminary step towards model development, we considered features used in other common models, including susceptible-infected-recovered (SIR) epidemic[31] and time delay[32] models, severity of lockdown measures[33], cumulative cases (both reported and not reported)[34], and daily test reports[35]. Furthermore, predictive models for infectious diseases such as SIRs provide guidance on disease transmission and outbreak causation. The State of Wisconsin's COVID-19 dashboard used a Case Rate metric defined as a per capita running seven-day sum of the case counts per county per day[36]. Case Rate standardizes COVID-19 severity across counties of differing populations while also accounting for data lags and providing insight on transmission. We plotted Case Rate vs. Indiana county populations 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 following 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 population intervals. Since Case Rate is more sensitive to less populated counties, we added intervals for counties under 100,000 people. Each population interval was given an assigned Case Rate baseline value that served as a binary indicator for outbreak determination. We implemented a criterion where counties were under outbreak if they were four standard deviations above the cumulative mean to account for data lag. As depicted in the system flow diagram (Figure 1), we established these parameter values and trained the model rules on the train datasets (data sets reported between March 1st, 2020 and August 31st, 2020). The train to test partition was roughly 71% to 29%, respectively, which is close to optimal for large datasets[37]. Then, the model was tested against the gold standard with the test datasets (datasets reported between September 1st, 2020 and October 31st, 2020).

Figure 1 presents a flow diagram depicting our study approach.

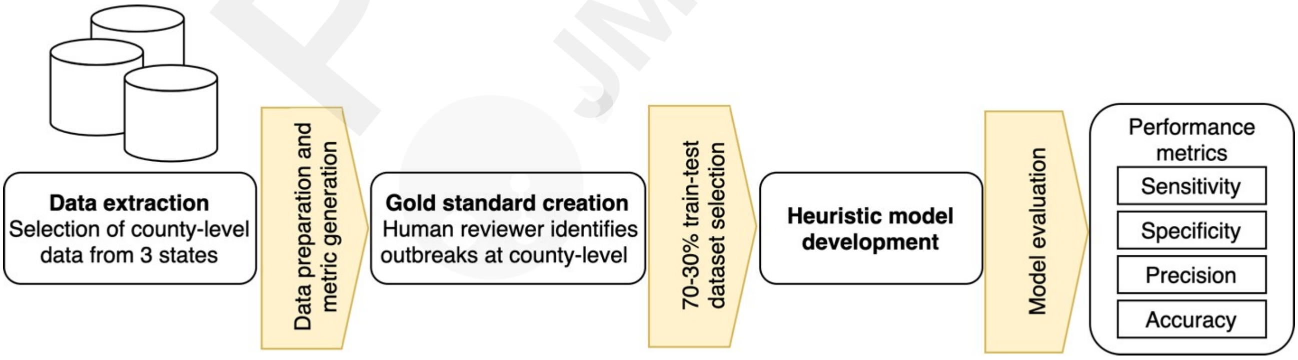

Data extraction
Selection of county-level data from 3 states
Data preparation and metric generation
Gold standard creation
Human reviewer identifies outbreaks at county-level
70-30% train-test dataset selection
Heuristic model development
Model evaluation
Performance metrics
Sensitivity
Specificity
Precision
Accuracy

Figure 1: Flow diagram for an overview of the study methodology.

Results

We collected data on a total of 249 counties from across all three states. Table 1 presents descriptive statistics on each state including number of counties, population sizes, and urbanization to highlight each state's fundamental differences[26]. Previous research has identified multiple factors in determining urban vs. rural areas, including total population, population density, and commuting flow[38].

Indiana and Iowa have similar county population distributions, with both having a majority of counties less than 100,000 people. 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 (72.44%) and Iowa (64.02%) far behind.

Table 1: Table 1 shows state and county population sizes and population statistics based on census counts.

	Indiana	Iowa	California
County-level Statistics			
Number of Counties	92	99	58
Counties where population < 100,000	75	93	23
Counties where population >= 100,000 and < 500,000	16	6	19
Counties where population >= 500,000 and < 1,000,000	1	0	7
Counties where population > 1,000,000	0	0	9
Population of Smallest County	5,875	3,602	1,129
Population of Largest County	964,582	490,161	10,039,107
Percentage of Urban Population	72.44%	64.02%	94.95%
Household Median Income ($)	59,892	68,718	70,489
Case Counts Per Day			
Average (St. Dev)	7.98 (21.02)	6.18 (15.75)	70.33 (231.86)

Figure 2 presents an example visualization of outbreak determination in Cass County, Indiana, and Santa Barbara County, California for gold standard preparation. Cass and Santa Barbara counties reported populations of 37,689 and 446,499 respectively[26].

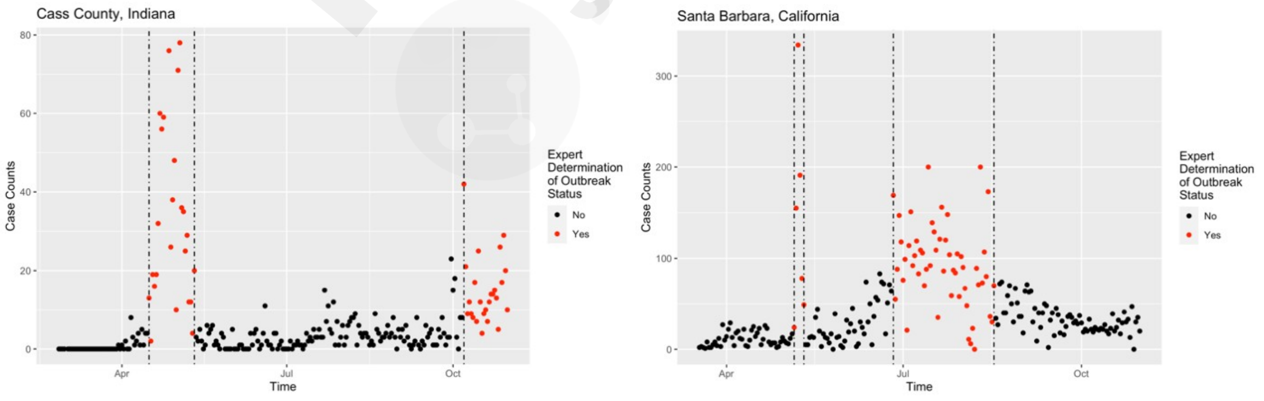


Figure 2: Visualization of the COVID-19 case counts in Cass County, Indiana, and Santa Barbara County, California, between March 1st, 2020 and October 31st, 2020. Days determined to be outbreak are colored red while normal days are black.

Table 2 shows 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 to the test date 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 between the train and test ranges quadrupled to 22.6% and 20.1% for Indiana and Iowa, respectively. The percentage of outbreak days in California remained relatively stable while the average outbreak duration decreased from 47 to 19 days. Because counties were our unit of analysis and since California had fewer and more populated counties than Indiana or Iowa, we believe these factors contributed to the reduced number of outbreaks in California.

Table 2: COVID-19 outbreak prevalence descriptors from the gold standard. Indiana, Iowa, and California datasets were divided by train (March 1st, 2020 and August 31st, 2020), test (September 1st, 2020 and October 31st, 2020), and total (March 1st, 2020 and October 31st, 2020) date ranges to characterize the outbreak periods.

	Indiana			Iowa			California		
	Train	Test	Total	Train	Test	Total	Train	Test	Total
Number of Outbreaks	26	65	83	43	85	114	35	26	40
Average Duration of outbreak	25	19.18	22.86	18.18	14.62	15.79	47.29	19.31	53.92
Total Days in Outbreak	650	1247	1897	727	1199	1926	1655	502	2157
Outbreak Days (%)	3.86 %	22.59 %	8.45%	4.01%	20.15%	7.97%	15.59%	14.43%	5.24%

Model Rules:

Using the above datasets, we developed model parameters to predict COVID-19 outbreaks. Figure 3 presents a top-down decision tree our model behavior. Rules and the assigned case rate associated with each population band used in the decision making process are further outlined in Appendix 1.

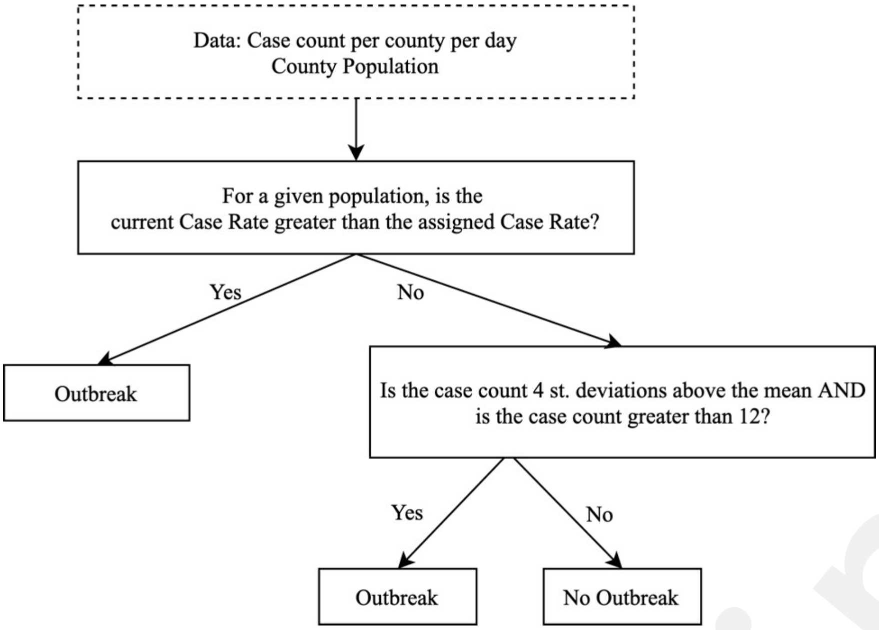

Data: Case count per county per day
County Population
For a given population, is the
current Case Rate greater than the assigned Case Rate?
Yes
No
Outbreak
Is the case count 4 st. deviations above the mean AND
is the case count greater than 12?
Yes
No
Outbreak
No Outbreak

Figure 3: Heuristic model decision making process. The current Case Rate is defined as the per capita running seven-day sum of the case counts per county per day.

As described in Figure 3, the heuristic model determined that counties experienced an outbreak through the following methods:

- 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.
- The county's case count on a specific day was greater than 12 and was 4 standard deviations above the rolling COVID-19 county case count mean.

If a county on a specific day met either requirement, then the county was labeled as "in outbreak". By combining these rules with the gold standard previously developed, a confusion matrix was utilized to provide an analysis on the model's performance.

Table 3 shows the confusion matrix results when the prediction model was applied to the curated gold standard during the test date range from September 1st, 2020 to October 31st, 2020. Sensitivity is the proportion of correctly identified positives, while specificity is the proportion of correctly identified negatives. All four key confusion matrix statistics -- sensitivity, specificity, precision, and accuracy-- were above 80% in each state during the test range. The model specificity and accuracy were both above 94% for each state. This was attributed to most days being classified as true negatives, which are fundamentally more straightforward to detect than true positives. Model sensitivity for Indiana and Iowa was 10% greater than California. However their precision was 11% and 7% lower, respectively. For Indiana and Iowa, this means the model computed fewer false negative readings, which could be attributed to having more and more prolonged outbreaks (Table 1). California's higher precision but lower sensitivity means the model was more precise in predicting when outbreaks happened but was less successful in capturing all outbreaks.

Table 3: Test data set model results against gold standard.

	Indiana	Iowa	California
Test Date Range	9/1/20-10/31/20	9/1/20-10/31/20	9/1/20-10/31/20

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%

Discussion

Our efforts resulted in the development of a heuristic model capable of detecting COVID-19 outbreaks with predictive measures between 80% and 99%. The model reported sensitivities of 92%, 90%, and 81% for Indiana, Iowa, and California respectively. This demonstrated that the model was capable of identifying a clear majority of outbreaks across each state. The model also reported precision scores of 85%, 89%, and 96% for Indiana, Iowa, and California, demonstrating that a majority of positive predictions made by the model were accurate. These performance metrics indicate that the model is fit for use in real-life settings. Additionally, the train and test periods displayed distinct outbreak characteristics due to the increased spread of COVID-19.

These performance metrics are also notable given that prevalence of outbreaks in train datasets was considerably low, and could have resulted in weak predictive models had we used more traditional classification based modelling approaches which significantly underperform when trained using unbalanced datasets[39 40]. As the pandemic progressed, each state worked to enhance their data reporting systems. As described by Khakharia et al., some regions reported sudden and significant changes in case counts, making it difficult for models to forecast future cases[41]. 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 reported twice as many outbreaks during the test period. This can be attributed to a second wave of COVID-19 cases that occurred during the test period as schools started, governors relaxed state lockdown laws, and citizens returned to work[42]. For example, California was one of the last states to begin lifting restrictions in midsize and large counties, which may have contributed to relatively fewer outbreaks than Indiana and Iowa[43 44]. Thus, counties reentered or for the first time realized outbreak periods during the test period.

California remains a state of interest due to the characterization of its outbreaks as well as predictive performance results on the holdout test dataset. Unlike Indiana and Iowa, California has several counties with populations over 1,000,000 people, and furthermore, it was the only state with fewer outbreaks and percentage of days in outbreak between the train and test periods. The California model revealed a significantly lower sensitivity but higher precision. Thus, to Indiana and Iowa, the California model captured proportionally fewer outbreaks but predicted the subset with greater precision.

This parsimonious prediction model is easily replicable in other states, as it only utilizes county population and COVID-19 cases per day per county data. States can detect and predict outbreaks with high accuracy following the model's rules. 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[22]. The heuristic 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. In addition to public and internal communication, forecasting models can be applied to aid in outbreak preparation and community mitigation methods[45].

In addition to a high performing heuristic model, our efforts also led to the development of a well curated gold standard dataset consisting of outbreak status for each county on a day to day basis. we share this dataset as an appendices (Appendix 2) to enable additional research around this significant line of research.

Limitations:

Our work was impacted by limitations in data collection systems currently deployed by each state. The inconsistency of data reporting presented a significant systematic challenge for model building activities. For instance, states closed most COVID-19 testing centers on weekends, leading to lower case count values on Saturdays and Sundays. Further, many states did not publish most of their own 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. New York Times would retroactively change case data, making it more unreliable since there were days with negative values.

The lack of prior research on curating gold standards on disease outbreaks 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 could have been subject to confounding that may have influenced our model's results. Furthermore, the rule-based model approach is subject to several limitations. Since the model incorporated a seven-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 other regions. This uncertainty, formed through parameters, social mandates, and vaccination, is a feature of any prediction model. We helped lessen this uncertainty through our generalizable approach demonstrated in 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, we could expand the model's data range for 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 determine when most outbreaks occurred in a region easily. Moreover, trends and patterns could be found across outbreaks between various factors such as lockdown policies, air pollution levels, 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. The model reported sensitivity scores > 90% in Indiana and Iowa, and > 80% in California. Furthermore, model specificity and accuracy scores were > 94% in every state. These results, coupled with the minimal data inputs required, creates an explainable and easy to implement model that governments and policy-makers can utilize to assess COVID-19 severity across diverse geographic regions. Future work includes testing the model in other states and countries using more recent data. Moreover, the model should be used to identify outbreaks to investigate correlations between external factors such as socioeconomic risks, air pollution, county level laws and outbreak development.

Acknowledgments:

None.

Conflicts of Interest:

None declared.

Appendix 1

Appendix 1: Population bands with their respective minimum case rates.

Appendix 2

Appendix 2: Outbreak gold standard for each county in California, Indiana, and Iowa per day

References

1. Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SN. Predicting COVID-19 mortality with electronic medical records. NPJ digital medicine 2021;**4**(1):1-10
2. Hartley DM, Perencevich EN. Public health interventions for COVID-19: emerging evidence and implications for an evolving public health crisis. Jama 2020;**323**(19):1908-09
3. Baker SR, Bloom N, Davis SJ, Terry SJ. Covid-induced economic uncertainty: National Bureau of Economic Research, 2020.
4. McNeil DG. Coronavirus Has Become a Pandemic, W.H.O. Says. Secondary Coronavirus Has Become a Pandemic, W.H.O. Says 2020. https://www.nytimes.com/2020/03/11/health/coronavirus-pandemic-who.html.
5. Coronavirus (Covid-19) Data. Secondary Coronavirus (Covid-19) Data. https://developer.nytimes.com/covid.
6. 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]|.
7. Cutler DM, Summers LH. The COVID-19 Pandemic and the $16 Trillion Virus. JAMA 2020;**324**(15):1495-96 doi: 10.1001/jama.2020.19759[published Online First: Epub Date]|.
8. Rubin R. NIH Launches Platform to Serve as Depository for COVID-19 Medical Data. Jama 2020;**324**(4):326-26
9. Bennett TD, Moffitt RA, Hajagos JG, et al. The National COVID Cohort Collaborative: Clinical Characterization and Early Severity Prediction. medRxiv:2021.01. 12.21249511
10. Dixon BE, Grannis SJ, McAndrews C, et al. Leveraging Data Visualization and a Statewide Health Information Exchange to Support COVID-19 Surveillance and Response: Application of Public Health Informatics. Journal of the American Medical Informatics Association 2021
11. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases 2020;**20**(5):533-34
12. Jewell NP, Lewnard JA, Jewell BL. Caution warranted: using the Institute for Health Metrics and Evaluation model for predicting the course of the COVID-19 pandemic: American College of Physicians, 2020.
13. Budd J, Miller BS, Manning EM, et al. Digital technologies in the public-health response to

COVID-19. Nature medicine 2020;**26**(8):1183-92

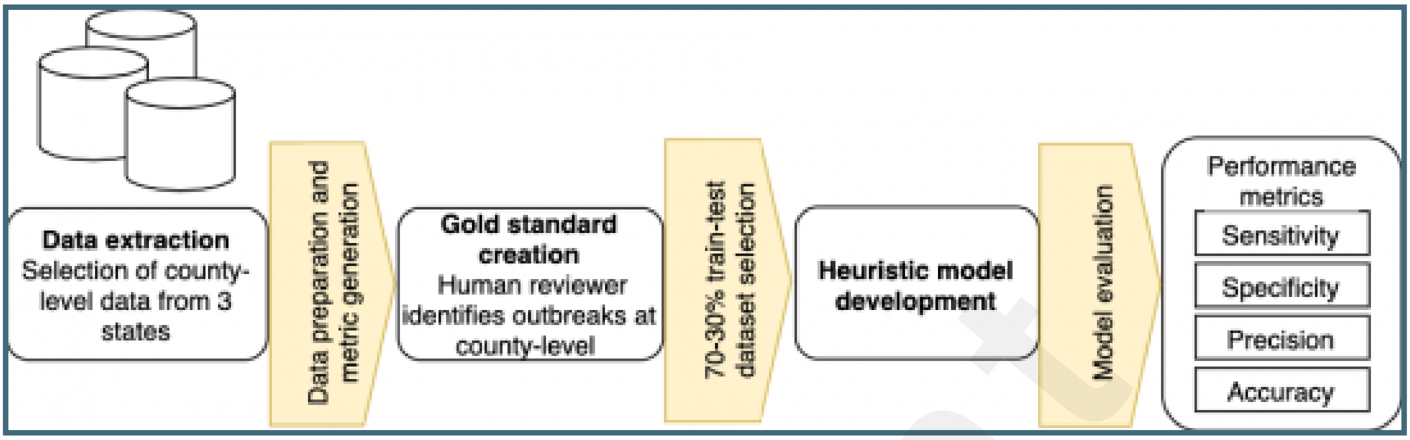
14. Dagan N, Barda N, Kepten E, et al. BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. N Engl J Med 2021 doi: 10.1056/NEJMoa2101765[published Online First: Epub Date]|.
15. Del Rio C, Malani P. COVID-19 in 2021-Continuing Uncertainty. JAMA 2021 doi: 10.1001/jama.2021.3760[published Online First: Epub Date]|.
16. COVID Data Tracker. Secondary COVID Data Tracker 2021. https://covid.cdc.gov/covid-data-tracker/#datatracker-home.
17. Weissman GE, Crane-Droesch A, Chivers C, et al. Locally Informed Simulation to Predict Hospital Capacity Needs During the COVID-19 Pandemic. Ann Intern Med 2020;**173**(1):21-28 doi: 10.7326/M20-1260[published Online First: Epub Date]|.
18. Dai Y, Wang J. Identifying the outbreak signal of COVID-19 before the response of the traditional disease monitoring system. PLoS Negl Trop Dis 2020;**14**(10):e0008758 doi: 10.1371/journal.pntd.0008758[published Online First: Epub Date]|.
19. Adnan M, Gao X, Bai X, et al. Potential Early Identification of a Large Campylobacter Outbreak Using Alternative Surveillance Data Sources: Autoregressive Modelling and Spatiotemporal Clustering. JMIR public health and surveillance 2020;**6**(3):e18281
20. Tseng Y-J, Shih Y-L. Developing epidemic forecasting models to assist disease surveillance for influenza with electronic health records. International Journal of Computers and Applications 2020;**42**(6):616-21
21. Verma M, Kishore K, Kumar M, Sondh AR, Aggarwal G, Kathirvel S. Google search trends predicting disease outbreaks: an analysis from India. Healthcare informatics research 2018;**24**(4):300
22. Dexter GP, Grannis SJ, Dixon BE, Kasthurirathne SN. Generalization of machine learning approaches to identify notifiable conditions from a statewide health information exchange. AMIA Summits on Translational Science Proceedings 2020;**2020**:152
23. 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]|.
24. Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence; 2019.
25. 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.
26. 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.
27. State Reporting Assessments. Secondary State Reporting Assessments. https://covidtracking.com/about-data/state-reporting-assessments.
28. COVID-19 County Statistics
. Secondary COVID-19 County Statistics
. https://www.in.gov/mph/1194.htm.
29. Ripley BD. The R project in statistical computing. MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network 2001;**1**(1):23-25
30. Maxmen A. Why the United States is having a coronavirus data crisis. Secondary Why the United States is having a coronavirus data crisis 2020. https://www.nature.com/articles/d41586-020-02478-z.
31. Yoshida N, Hara T. Global stability of a delayed SIR epidemic model with density dependent birth and death rates. Journal of Computational and Applied Mathematics 2007;**201**(2):339-47 doi: https://doi.org/10.1016/j.cam.2005.12.034[published Online First: Epub Date]|.

32. Zaman G, Kang YH, Jung IH. Optimal treatment of an SIR epidemic model with time delay. Biosystems 2009;**98**(1):43-50 doi: https://doi.org/10.1016/j.biosystems.2009.05.006[published Online First: Epub Date]|.
33. Giordano G, Blanchini F, Bruno R, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nat Med 2020;**26**(6):855-60 doi: 10.1038/s41591-020-0883-7[published Online First: Epub Date]|.
34. Lin Q, Zhao S, Gao D, et al. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. Int J Infect Dis 2020;**93**:211-16 doi: 10.1016/j.ijid.2020.02.058[published Online First: Epub Date]|.
35. Casella F. Can the COVID-19 epidemic be controlled on the basis of daily test reports? 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200306967C (accessed March 01, 2020).
36. 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.
37. 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]|.
38. Hall SA, Kaufman JS, Ricketts TC. Defining urban and rural areas in U.S. epidemiologic studies. J Urban Health 2006;**83**(2):162-75 doi: 10.1007/s11524-005-9016-3[published Online First: Epub Date]|.
39. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 2016;**5**(4):221-32
40. Kasthurirathne SN, Grannis SJ. Machine Learning Approaches to Identify Nicknames from A Statewide Health Information Exchange. AMIA Summits on Translational Science Proceedings 2019;**2019**:639
41. 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]|.
42. 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.
43. Each State's COVID-19 Reopening and 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/.
44. 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]|.
45. Lutz CS, Huynh MP, Schroeder M, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health 2019;**19**(1):1659 doi: 10.1186/s12889-019-7966-8[published Online First: Epub Date]|.

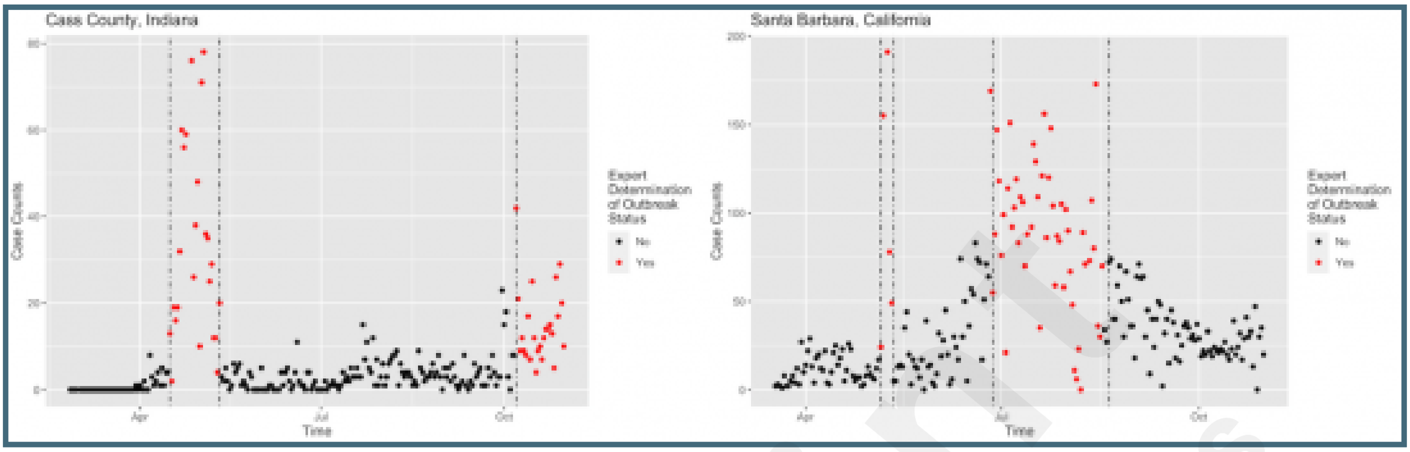
Supplementary Files

Figures

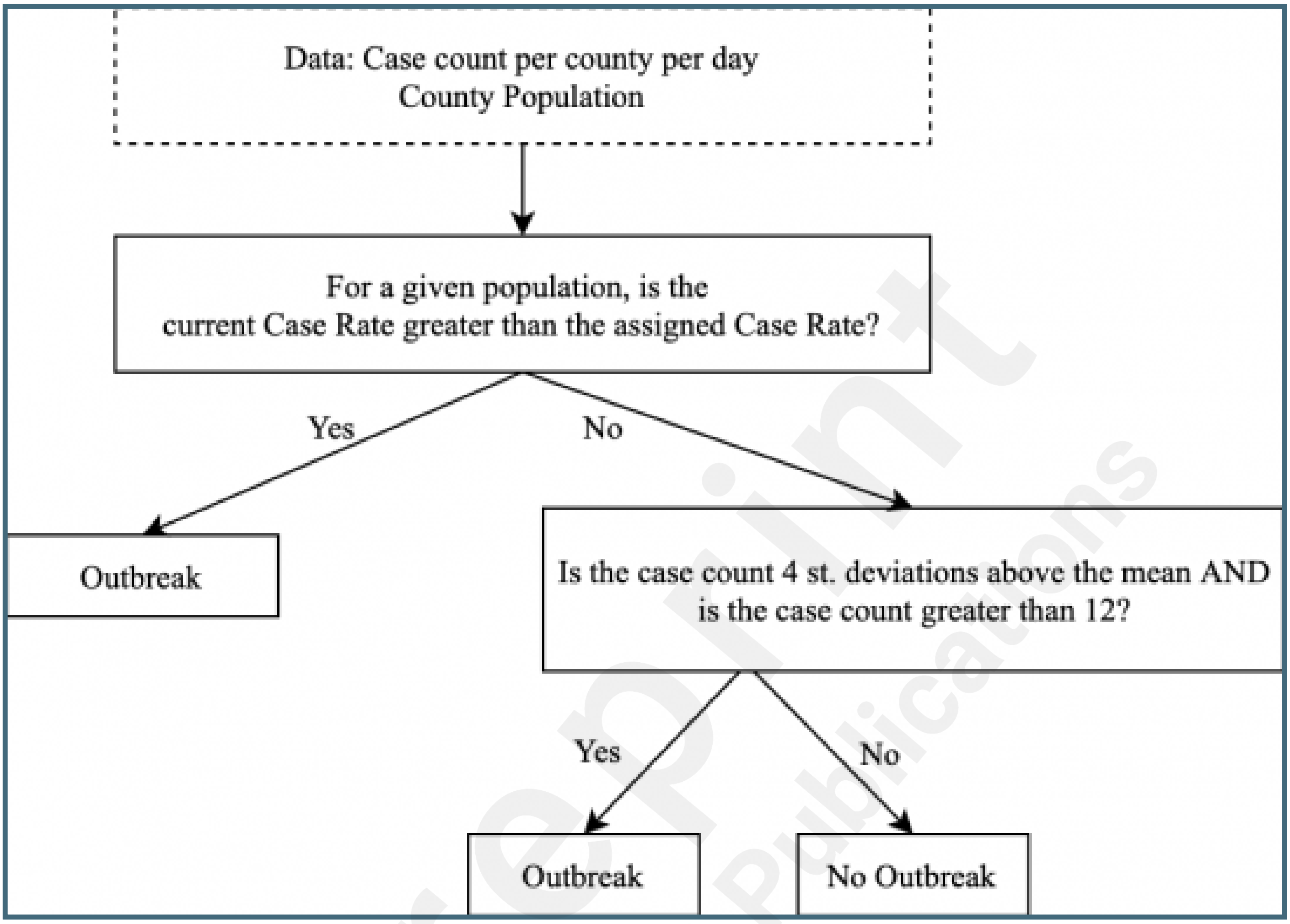
Flow diagram for an overview of the study methodology.


Data extraction
Selection of county-level data from 3 states
Data preparation and metric generation
Gold standard creation
Human reviewer identifies outbreaks at county-level
70-30% train-test dataset selection
Heuristic model development
Model evaluation
Performance metrics
Sensitivity
Specificity
Precision
Accuracy

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.


Data: Case count per county per day
County Population
For a given population, is the current Case Rate greater than the assigned Case Rate?
Yes
No
Outbreak
Is the case count 4 st. deviations above the mean AND is the case count greater than 12?
Yes
No
Outbreak
No Outbreak

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