PLOS ONE

RESEARCH ARTICLE

Predicting increases in COVID-19 incidence to identify locations for targeted testing in West Virginia: A machine learning enhanced approach

Bradley S. Price[1,2]*, Maryam Khodaverdi[1], Adam Halasz[3], Brian Hendricks[4], Wesley Kimble[1], Gordon S. Smith[4], Sally L. Hodder[1,5]

1 West Virginia Clinical and Translational Science Institute, Morgantown, West Virginia, United States of America, **2** Management Information Systems Department, West Virginia University, Morgantown, West Virginia, United States of America, **3** School of Mathematics and Data Science, West Virginia University, Morgantown, West Virginia, United States of America, **4** Department of Epidemiology and Biostatistics, West Virginia University, Morgantown, West Virginia, United States of America, **5** West Virginia University School of Medicine, Morgantown, West Virginia, United States of America

* brad.price@mail.wvu.edu

Citation: Price BS, Khodaverdi M, Halasz A, Hendricks B, Kimble W, Smith GS, et al. (2021) Predicting increases in COVID-19 incidence to identify locations for targeted testing in West Virginia: A machine learning enhanced approach. PLoS ONE 16(11): e0259538. https://doi.org/10.1371/journal.pone.0259538

Editor: Sanjay Kumar Singh Patel, Konkuk University, REPUBLIC OF KOREA

Received: August 24, 2021

Accepted: October 20, 2021

Published: November 3, 2021

Data Availability Statement: The data used in this manuscript is owned by a third party. To request access contact the West Virginia Department of Health and Human Resources (https://dhhr.wv.gov/Pages/contact.aspx). Relevant code and data similar to what is used in this manuscript can be found in a Github Repository: https://github.com/MKhodaverdi/Covid19_County_Prediction.

Abstract

During the COVID-19 pandemic, West Virginia developed an aggressive SARS-CoV-2 testing strategy which included utilizing pop-up mobile testing in locations anticipated to have near-term increases in SARS-CoV-2 infections. This study describes and compares two methods for predicting near-term SARS-CoV-2 incidence in West Virginia counties. The first method, R_t Only, is solely based on producing forecasts for each county using the daily instantaneous reproductive numbers, R_t. The second method, ML+R_t, is a machine learning approach that uses a Long Short-Term Memory network to predict the near-term number of cases for each county using epidemiological statistics such as R_t, county population information, and time series trends including information on major holidays, as well as leveraging statewide COVID-19 trends across counties and county population size. Both approaches used daily county-level SARS-CoV-2 incidence data provided by the West Virginia Department Health and Human Resources beginning April 2020. The methods are compared on the accuracy of near-term SARS-CoV-2 increases predictions by county over 17 weeks from January 1, 2021- April 30, 2021. Both methods performed well (correlation between forecasted number of cases and the actual number of cases week over week is 0.872 for the ML+R_t method and 0.867 for the R_t Only method) but differ in performance at various time points. Over the 17-week assessment period, the ML+R_t method outperforms the R_t Only method in identifying larger spikes. Results show that both methods perform adequately in both rural and non-rural predictions. Finally, a detailed discussion on practical issues regarding implementing forecasting models for public health action based on R_t is provided, and the potential for further development of machine learning methods that are enhanced by R_t.

Funding: The project described was supported by the National Institute Of General Medical Sciences, 5U54GM104942-04 and 5U54GM104942-05S3.

Competing interests: No authors have competing interests.

Introduction

The novel coronavirus (SARS-CoV-2) pandemic has had a large impact on health systems and prior to vaccines being available public health measures such as testing, contact tracing and social distancing were the prevention methods available [1, 2]. Rural communities in the United States (US) have also been heavily impacted by the pandemic. SARS-CoV-2 related deaths have occurred disproportionately among rural areas of the US, and negative impacts on health and economic well-being have been described to be more severe among rural populations [3, 4]. Persons living in rural communities often have multiple barriers to health care and laboratory diagnostic testing due to geographic, transportation, and cost [5]. Early in the COVID-19 pandemic, the state of West Virginia (WV) provided county-specific data on SARS-CoV-2 testing results so that daily instantaneous reproductive numbers (R_t) could be calculated for each WV county to indicate viral transmission dynamics. An aggressive SARS-CoV-2 testing strategy was implemented that included static as well as mobile testing units. The Rapid Acceleration of Diagnostics in Underserved Populations (RADx-UP), funded by the National Institutes of Health, provided the opportunity to deliver pop-up mobile testing in those areas predicted to have the greatest increases in SARS-CoV-2 incidence. The objective of RADx-Up was to increase testing in those communities most likely to have a near-term (within 7–10 days) increase in COVID-19 cases, thereby potentially providing early identification of SARS-CoV-2 infected persons who may then quarantine more rapidly in an effort to blunt the anticipated increase in new cases.

Two strategies to predict near-term increases in SARS-CoV-2 cases were developed using recent county-specific incidence of infections and R_t–one method is a dynamical algorithm-based prediction using R_t and the serial interval while the second method uses a Long Short-Term Memory (LSTM) machine learning strategy. The objective of this study, in support of RADx-Up, was to recommend counties of outbreak for targeted testing. The accuracy of the two methods to predict short-term increases in county-specific SARS-CoV-2 incidence is compared and a discussion on conditions favoring one method or the other is also provided. This study was conducted prior to the emergence of the new Delta SARS-CoV-2 variant which has proven to be more transmissible and with increased mortality than the original strain that prevailed over the study period [6, 7].

Data and methods

Data

To obtain estimates of near-term increases in SARS-CoV-2 cases, a likelihood-based model underlying the EpiEstim package in R and developed in Cori et al. [8] and Thompson et al. [9] was deployed. using software provided by Imperial College London [10] Two methods were employed: 1) the R_t Only method, a forecast based on the reproduction number and associated serial interval that predicts the future R_t that is then extrapolated to estimate the number of future cases; 2) a LSTM machine learning model (ML+R_t) that utilizes the reproduction number from the R_t Only method as an input, but also utilizes total cases and population, among other inputs, to predict the total number of cases for a given period of time.

The data in this study is based on daily reports of all daily COVID-19 polymerase chain reaction (PCR) and antigen testing results conducted in WV since March 2020 directly from the WV Department of Health and Human Resources (WVDHHR). Noteworthy is that all SARS-COV-2 testing data are required to be reported to WVDHHR. Information for each unique patient is collected and contains test procurement date, test result date, patient zip code, patient county of residence, testing site name, county where the test is obtained, and test

result. As patients who test positive may be tested multiple times, only the first positive tests on a patient is considered. When applying this filter, data obtained from all testing sites is used (i.e., hospital, clinic, pharmacy, drive-through, mobile van). The number of daily cases for each county is calculated by adding the lab confirmed cases and clinical confirmed cases after filtering out repeated tests or COVID-19 diagnoses. This daily incidence data on first diagnosed infection is the basis for calculation of R_t.

R_t Only method: Producing short term predictions

Our R_t Only method relies on the methodology used in the EpiEstim package and the underlying modeling approach of Cori et al. [8] and Thompson et al. [9]. This approach relates the daily incidence (number of new cases) to past cases through an instantaneous reproduction number R_t which characterizes the daily dynamics of transmission reflects a multitude of factors relating to individual and group behavior in the community of interest.

As a brief review, daily infections within a community occur as independent random events drawn from a Poisson distribution. The probability that exactly k cases occur is $p_k = \frac{\Gamma^k}{k!}e^{-\Gamma}$, and the rate parameter Γ coincides with the average daily incidence, $\langle k \rangle = \Gamma$. In the instantaneous R_t framework, the expected incidence on day t is a product of two quantities, the infection potential and the reproduction number, $\Gamma_t = \Lambda_t R_t$. The infection potential Λ_t summarizes the record of past cases in the community and the typical variation of the infectiousness of an individual over time.

The infection potential Λ_t is determined by the incidence I_{t-s} on prior days $s = 1,2,\ldots$ and the serial interval distribution w_s.

$$\Lambda_t = \sum_{s=1}^{S_{\max}} I_{t-s} w_s$$

The serial interval distribution w_s reflects the time course of infectiousness of one infected individual at $s = 1,2,\ldots$ days from the primary infection. It encapsulates the relative increase and decrease of infectiousness of an individual, assuming all other conditions in the community remain unchanged. In practice, the serial interval is typically obtained as the normalized ($\sum_{s=1}^{S_{\max}} w(s) = 1$) distribution of time intervals between known infector-infected pairs. Based on studies done by Gostic et al. [11] and Challen et al. [12], a serial distribution extending over 100 days was used. The infection potential can be understood as the sum of the expected number of infections on day t, due to past cases in the community, under ideal "steady state" conditions, such that over time, each primary case causes exactly one secondary case.

The time varying reproduction number, R_t, captures conditions of transmission that are external to the infected individuals and reflect community behavior. In this framework, R_t is a random variable with a Gamma distribution $f(R) = \frac{1}{b^a \Gamma(a)} R^{a-1} e^{-R/b}$. The parameters a_t, b_t are determined for each day through Bayesian (maximum a posteriori probability) estimation. The parameters of interest are estimated using incidence data up to and including the current day, $I_1, I_2, \cdots I_t$ as follows:

$$a_t = \sum_{s=0}^{\tau-1} I_{t-s} + a_{\text{prior}},\ \frac{1}{b_t} = \sum_{s=0}^{\tau-1} \Lambda_{t-s} + \frac{1}{b_{\text{prior}}},\ \Lambda_t = \sum_{s=1}^{S_{\max}} I_{t-s} w_s$$

This estimated R_t distribution applies to the most recent τ days, but it requires the values of $I_{t'}$ for $t' \leq t$ going back to $t' = t - s_{\max}$ where $s_{\max}$ is the length of the serial interval distribution. For the serial interval w_s a discretized gamma distribution was used with mean and standard deviation of t_s = 7.0 ± 4 days, provided in the software similar to Cori [8].

For the serial interval w_s a gamma distribution was used with mean and standard deviation of $\tau_s = 6.99 \pm 4.02$ days, as given by Flaxman et al. [13]. Following Cori and Thompson's method, a prior distribution was used with mean and standard deviation equal to 5 ($a_{\text{prior}} = 1$, $b_{\text{prior}} = 5$) [8, 9].

The semi-deterministic model for future incidence, based on Cori's method regards the daily distributions of R_t (values of a_t, b_t) as inputs that summarize the current conditions for disease transmission within the community of interest. The serial interval distribution w_s, which is fixed with regard to time, is also an input. Thus, on day t the distribution of R_t is known and applies to this day (assessed using the most recent τ days, similar to a trailing moving average).

Next day prediction. Assuming time series of past daily incidences $\{I_u\}_{u=0,1,..t}$ ending on day t is observed, the number of infections on the next day t+1 follows a Poisson distribution, with parameter $\Gamma_{t+1} = \Lambda_{t+1}R_{t+1}$, where R_{t+1} is also a random variable. Assuming the parameters a,b of $f(R_{t+1}|a,b)$ are known, the probability of exactly k new infections on day t+1 is:

$$P(k|R_{t+1},\Lambda_t) = \frac{(\Lambda_{t+1}R_{t+1})^k}{k!}e^{-\Lambda_{t+1}R_{t+1}} \rightarrow P(k|\Lambda_{t+1},a,b) = \int_0^\infty P(k|R,\Lambda_{t+1})f(R|a,b)\,dR$$
$$= \frac{(b\Lambda_{t+1})^k}{(b\Lambda_{t+1}+1)^{a+k}}\prod_{j=1}^{k}\frac{(a+j)}{j}$$

The expected number of new infections coincides with the infection potential multiplied by the expected R.

$$\langle I_{t+1}\rangle_{R_{t+1}} = \Lambda_{t+1}R_{t+1} \rightarrow \langle\langle I_{t+1}\rangle_{R_{t+1}}\rangle_{a,b} = \Lambda_{t+1}\langle R_{t+1}\rangle_{a,b} = \Lambda_{t+1}ab$$

For the purpose of predicting a likelihood range for the daily incidence, CDF of R_{t+1} is defined as:

$$P(\bar{I}_t \in [I_1,I_2]|a,b,\Lambda) = gamcdf\left(\frac{I_2}{\Lambda}|a,b\right) - gamcdf\left(\frac{I_1}{\Lambda}|a,b\right) = \frac{1}{b^a\Gamma(a)}\int_{I_1/\Lambda}^{I_2/\Lambda}R^{a-1}e^{-\frac{R}{b}}dR$$

A [5% - 95%] credibility interval is obtained for the daily incidence using the inverse CDF for $R$and multiplying by the corresponding infection potential. This provides a smaller variance than the discrete distribution $P(k)$ but is a more practical indication of the incidence rate.

Extrapolation over multiple days. To go beyond the "next" day, the one-day prediction is iterated, using predicted values to expand the incidence data. One can reasonably extrapolate the current distribution of R_t to t+1 and any number of days in the future. For the short term (7 day) predictions discussed here, the value of the most recent available R_t remains the same over the prediction interval, $\bar{R}_{t+k} = R_t$.

The estimated incidence for day t+1 requires the infection potential on that day Λ_{t+1}, which is computed based on incidence up to the preceding day t.

$$\bar{I}_{t+1|t} = \Lambda_{t+1}\bar{R}_{t+1} = \Lambda_{t+1}R_t,\ \Lambda_{t+1} = \sum_{s=1}^{s_{max}} I_{(t+1)-s}w_s = I_t w_1 + I_{t-1}w_2 + \cdots I_{(t+1)-s_{max}}w_{s_{max}}$$

Predictions for day t+2 and beyond can be obtained using the predictions for preceding days for the incidence and iteratively applying the approach for any number of k days into the

future.

$$\bar{I}_{t+1|t} = \Lambda_{t+1} R_t \qquad \bar{\Lambda}_{t+2|t} = \sum_{s=2}^{s_{\max}} I_{t+1-s} w_s + \bar{I}_{t+1|t} w_1$$

$$\bar{I}_{t+2|t} = \bar{\Lambda}_{t+2} R_t \quad \bar{\Lambda}_{t+3|t} = \sum_{s=3}^{s_{\max}} I_{t+2-s} w_s + \sum_{s=1}^{2} \bar{I}_{t+2-s} w_s$$

$$\bar{I}_{t+k|t} = \bar{\Lambda}_{t+k} R_t \quad \bar{\Lambda}_{t+k|t} = \sum_{s=k}^{s_{\max}} I_{t+k-s} w_s + \sum_{s=1}^{k} \bar{I}_{t+k-s} w_s$$

The estimate of the credibility intervals are similar to the one-day case, using only the corresponding range for the reproduction number R_t, and not compounding with uncertainty for each estimated incidence $\bar{I}_{t+k}$ or with the additional uncertainty due to the Poisson distribution of the daily (integer) incidence. While this provides a narrower range, the credible interval serves as a relative measure of the uncertainty affecting the prediction.

Correction for imported cases. Not accounting for imported SARS-CoV-2 cases into a county will lead to over estimation of R_t. In practice, imported cases are not able to be directly identified, so an adjustment must be made to identify them. Assuming the daily incidence I_t can be separated into imported and community-spread parts:

$$I_t = I_t^{(\text{local})} + I_t^{(\text{imported})}$$

Then, imported cases are an additional input to the model. Imported cases are included in the infection potential because they contribute to new local infections, but are not included in the number of new cases when estimating the reproduction number:

$$a_t = \sum_{s=0}^{\sigma-1} I_{t-s}^{(\text{local})} + a_{\text{prior}}, \frac{1}{b_t} = \sum_{s=0}^{\sigma-1} \Lambda_{t-s} + \frac{1}{b_{\text{prior}}}, \Lambda_t = \sum_{s=1}^{s_{\max}} I_{t-s} w_s$$

Turning to predictions, the reproduction number and infection potential computed in the standard framework can only predict the local cases:

$$R_t \sim gampdf(a_t, b_t), I_t^{(local)} \sim poisspdf(R_t \Lambda_t) \rightarrow \langle I_t^{(local)} \rangle = \Lambda_t \langle R_t \rangle = \Lambda_t a_t b_t$$

By definition, imported cases cannot be predicted in the R_t model; however, events when the observed number of new cases vastly exceeds the expectation from local transmission can be identified. This hindsight is used to improve the estimate of the reproduction number as follows.

The likely number of imported cases on a given day is estimated by comparing the actual incidence to the Bayesian credible interval for new local cases estimated from the previous days. This estimated past incidence is then incorporated in a corrected estimate for R_t.

In an initial pass the a_t, b_t parameters are computed for time point t based on the incidence time series $\{I_\tau\}_{\tau = 0,1,\cdots t-1}$. The one-day predicted incidence on day t is computed as described above, using the infection potential Λ_t and the distribution of $\bar{R}_t \equiv R_{t-1}$. The value corresponding to the upper $\theta = 95\%$ credible interval is used as a cutoff and identify the part of the incidence that exceeds the cutoff with imported cases.

$$I_t^{(\text{local,high})} = \Lambda_t \, \text{gaminv}(\theta, a_{t-1}, b_{t-1}), I_t^{(\text{imported,est})} = \max(I_t - I_t^{(\text{local,high})}, 0)$$

The estimated local incidence $I_t^{(\text{local,est})}$ is used to provide revised estimates for the reproduction number as described above (also consistent with Cori and Thompson's approach). Finally,

the resulting R_t parameters are used for the most recent day and the full incidence to provide revised estimates for days t+1, t+2,. . .t+k.

ML+R_t method: Using long short-term memory network to forecast outbreaks

As previously mentioned, the LSTM method implemented in this project is meant to build on the widely used R_t Only approach described in the previous section. The novelty of this LSTM approach is that it provides for the input of epidemiological modeling while taking advantage of cutting-edge machine learning techniques. The combination of the two allows the LSTM model to incorporate the epidemiological principles used to produce the R_t estimate while adding additional information such as temporal and demographic information that can be leveraged with traditional machine learning models. Further, the calculation of R_t using the Rt Only method uses independent data sets for each county in turn creating a unique model for each county that does not consider the impact of possible relationships between counties. By contrast, the ML+R_t approach uses global trends across counties. By training on all the data, the model is not only able to take advantage of global trends, but by including spatial information, the model is also able to show how these trends impact specific counties.

Daily county-specific R_t, summary statistic information on the estimated R_t such as standard deviation, confidence intervals, and the probability of $R_t > 1$ are also provided. Rt values computed using both 7 and 14 day intervals are included. All these factors along with temporal information such as daily information on whether it is a weekend or not, holiday or not, days passed from last major holiday, days to the next major holiday were utilized as inputs for our model.

As mentioned previously, due to the length of time it takes to receive a test result (lag time), the deflated effect on the positive cases when considering test procurement date must be considered. An average lag of 3 days for results to achieve close to actual levels was observed. To mitigate the effect of the testing lag day *t*, *t-1*, *t-2* are imputed with the actual SARS-CoV-2 cases for days *t-3*, *t-4*, *t-5* respectively.

A LSTM recurrent neural network [14] was implemented in Python with an Adam optimizer, as our model of interest for this analysis, permitting consideration of all available county-specific input information for the past 7 days with a prediction of the number of positive cases for the county as an output. Other advantages of the LSTM approach are the ability to exploit autocorrelation between time points and the utilization of dropout layers to remove redundant information.

In general, the LSTM models are more complex versions of recursive neural networks (RNNs). The multi-layer LSTM method deployed here follows the framework described in Fig 1 where the input layer is defined by a matrix combining the number of positive cases for county c at time point t, $Y_{c,t}$, and all inputs for county c at time point t. The inputs then move their way through the network (i.e., through the LSTM layer and dense layer) to obtain an output. The output is defined as, $\hat{Y}_{c,t+7}$, the predicted daily number of cases for county c at time point t+7. LSTM can be viewed as a network where information between time points is shared. Each LSTM cell, diagramed in Fig 1, shares two pieces of information with other LSTM cells; the current state of the cell, C_t, and output of the cell, h_t, is calculated with the following formulas given input data, x_t:

$$\tilde{C}_t = \sigma'(W_c.[h_{t-1}, x_t] + b_c)$$

$$g_t^i = i_t \times \tilde{C}_t$$

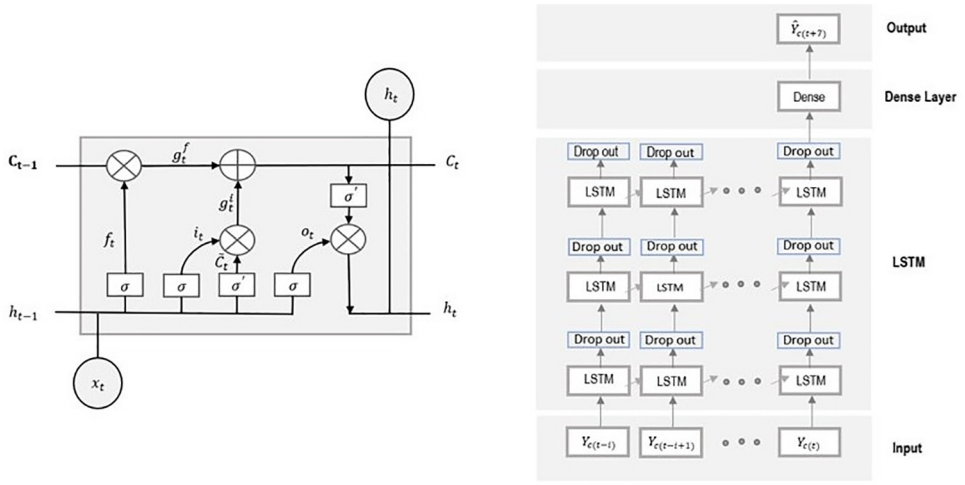

Fig 1. The LSTM framework deployed for the proposed ML+R_t method on right, and structure of each LSTM cell on left.

https://doi.org/10.1371/journal.pone.0259538.g001

$$g_t^f = f_t \times C_{t-1}$$

$$C_t = g_t^f + g_t^i$$

$$h_t = o_t \times \sigma' C_t$$

Where, w are the weight variables (traditionally thought of like regression coefficients), and b are the bias variables (traditionally thought of as intercept terms). Activation functions, σ and σ' are non-linear transformation functions such as, sigmoid and hyperbolic tangent. A feature of each cell is input, output, and forget gates. These gates are what give the LSTM, the memory property which allows it to account and adjust for auto correlation. Define:

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o)$$

The above are gates that define the memory of the LSTM cell and are distinct linear combinations of inputs and outputs from the previous LSTM cell with specific activations functions.

In addition, the importance of the inputs is not guaranteed (including R_t and associated summary statistics), thus dropout layers were added to allow for the identification of important inputs. The drop out layers filtered out inputs that would be considered insignificant in order to detect the important signals coming from the input data and also protect against overfitting.

Once predictions for a given week were determined, the summary statistics of the results were produced. Summary statistics included: 1) predicted number of positive cases by county, 2) predicted percent change in cases per 100,000 persons by county compared to the previous week, 3) predicted increase in number of cases compared to the previous week, and 4) predicted number of cases relative to the population size.

Evaluations of models in deployment

Metrics and evaluation

To evaluate performance of the two methods, the predicted values for new SARS-CoV-2 cases were benchmarked against the actual number of positive cases recorded for each week from January 1, 2021 through April 30, 2021. As a main goal of these new case forecasts was to target areas for diagnostic testing, each week's prediction is used as a recommendation. These recommendations were ranked on many several metrics but most predominately on the percentage increase in cases over the previous week. To evaluate the recommendations, the total discounted cumulative gain (DCG) of each method is measured [15]. DCG is a commonly used metric in page ranking calculations and is suitable here as the information shared was used similar to page ranking calculations. As a reminder the goal of this analysis is to recommend counties of increased incidences for intervention (i.e., increased SARS-CoV-2 testing), not to predict the actual number of incidence. DCG provides a metric for comparison of differing recommendation methods, which is how both the ML+R_t and Rt Only are being used. Unlike most metrics used in machine learning such as squared error or absolute error, larger DCG values indicate better performance.

To better study performance of the ML+R_t and R_t Only methods, two separate DCG metrics are used to consider the cost of poor recommendations. The first is on the ability to identify the top counties of increase regardless of the level of increase, while the second metric considers the size of the increase (percentage) in the comparison.

To define the first metric let $\hat{y}_{c,t}$ and $y_{c,t}$ represent the number of predicted cases and actual cases over a 7-day period for the cth county at time point t respectively. To keep from biasing the evaluation towards rural areas with a low incidence, only consider those with $y_{c,t+1}>10$ are considered. Define S_t to be the set of indices, the largest 10 values of $\frac{y_{c,t+1}}{y_{c,t}}$ for a given time point. The Binary Discounted Cumulative Gain (Binary DCG) of a set of rankings at time point t is defined as:

$$\sum_{i=1}^{q} \frac{I(i \in S_t)}{\ln(i+1)}$$

where $I(i \in S_t)$ is an indicator of a correct identification of a top 10 ranking in the actual percentage increases, and q is the number of rankings used in the calculation. For example, if $q = 10$, then $BDCG_t$ would only evaluate the top 10 rankings, in our setting this would be the top 10 counties, returned by a method. One may view B-DCG as a weighted identifier to measure the quality of the rankings for purposes of identifying case increases (or spikes) of the top q recommendations.

As the closeness of the predicted number of cases to the actual case number, i.e., the "quality" of the prediction, a second metric was used to consider the quality of the prediction rather than just considering a binary outcome. To accomplish this, Spike DCG is defined as:

$$\sum_{i=1}^{q} \frac{y_{i,t+1}}{y_{i,t}\ln(i+1)}.$$

Spike DCG considers the relative size of the spike for the top q recommendations. While Binary DCG investigates the ability of a method to correctly identify the top 10 counties, Spike DCG places value on the recommendations that are produced by identification of larger spikes. This comparison is of great importance as targeted interventions may only have finite

resources to deploy so understanding the level of trust and impact expected by the two methods is of importance.

As both the R_t Only and ML+R_t methods are used to recommend county level locations for testing, it is important to investigate the quality of the top recommendations, disregarding the order and quality of the ranked predictions. This evaluation gives a sense of the quality of the recommendations produced by the methods, relative to others.

Finally, as this study is being deployed in a state with many rural areas, differences in methods between rural and non-rural areas were also analyzed. This study uses the 2013 Rural-Urban Continuum Codes (RUCC) to define rurality [16], which define a rural area as a non-metro area with population under 20,000 and is not adjacent to an urban metro area. To asses quality of the predictions provided by each method, we examined correlations between predicted and actual 7-day positive case totals. The quality of Binary DCG and Spike DCG in both rural and non-rural areas is assessed by investigating the performance of ML+R_t and R_t Only methods among lower population communities with less access to large healthcare systems. Both R_t Only and ML+R_t methods were deployed each week from January 1, 2021 through April 30,2021 using all available training data beginning in April 2002 for each of the 55 counties in the state of WV, and resulting county recommendations were retained for comparison against the actual number of cases. The code for fitting and deploying the models is publicly available [17].

Results

The daily number of tests from April 2020-April 2021 were highly variable (Fig 2 with some weeks having very low testing rates as illustrated by Fig 3). Each of the two prediction methods utilized all available data and was updated weekly to obtain county level predictions. Note that this study specifically focuses on evaluating predictions in the latter part of this time frame, and coincided with vaccinations becoming available to different demographics of residents of West Virginia residents, though data from the entire study was used to train each of the methods.

The correlation between forecasted number of cases and the actual number of cases week over week is 0.872 for the ML+R_t method and 0.867 for the R_t Only method. Fig 4 shows a scatter plot of the relationship between forecasted cases and the actual corresponding cases.

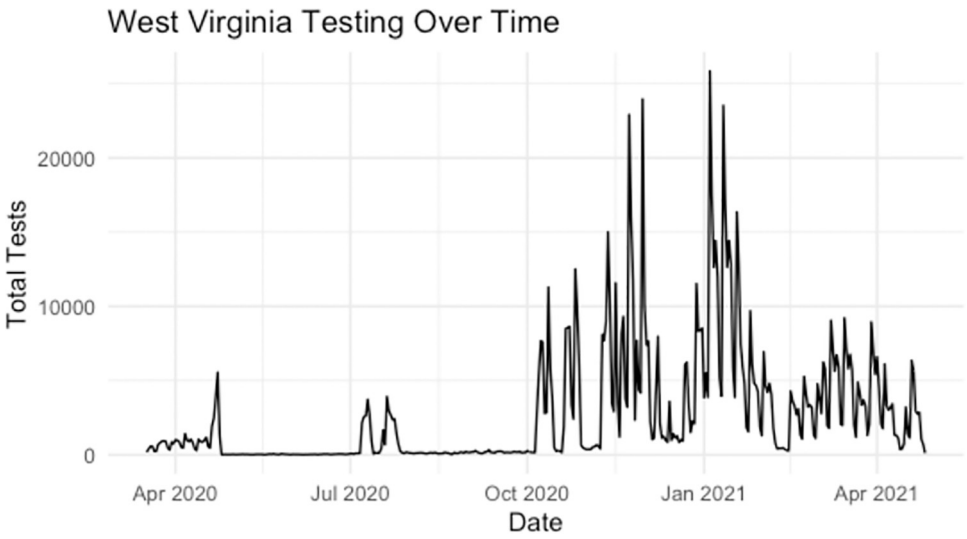

Fig 2. Number of SARS-COV-2 tests in the state of West Virginia from April 2020-April 2021.

https://doi.org/10.1371/journal.pone.0259538.g002

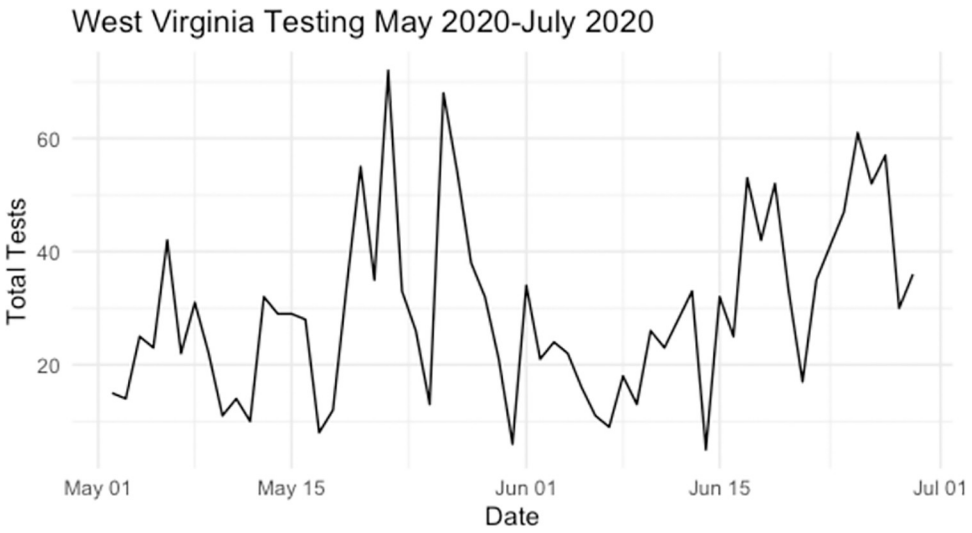

Fig 3. Number of SARS-COV-2 tests in the state West Virginia from May 2020- July 2020.

https://doi.org/10.1371/journal.pone.0259538.g003

Fig 5 compares Binary and Spike DCG for the case of recommending 10 counties ($q = 10$) and 55 counties ($q = 55$). Both the R_t Only and ML+R_t methods perform well overall but differ in performance at various time points. In the case of Binary DCG the Rt Only method has better performance, and in the case of Spike DCG the ML+R_t method performs better. Over the 17-week assessment period, the ML+R_t method outperforms the R_t Only method in recommendations with regard to all measures except Binary DCG for $q = 10$ (Table 1). These results show that if users are interested in mitigating outbreaks by identifying larger spikes in the Top 10 recommendations, as was the goal of this implementation, the ML+R_t method should be used.

A more concerning result is the decrease in both DCG metrics that are seen with regard to both methods over time. Further investigation and analysis showed that during deployment the focus of providers shifted from active testing and contact tracing to vaccination.

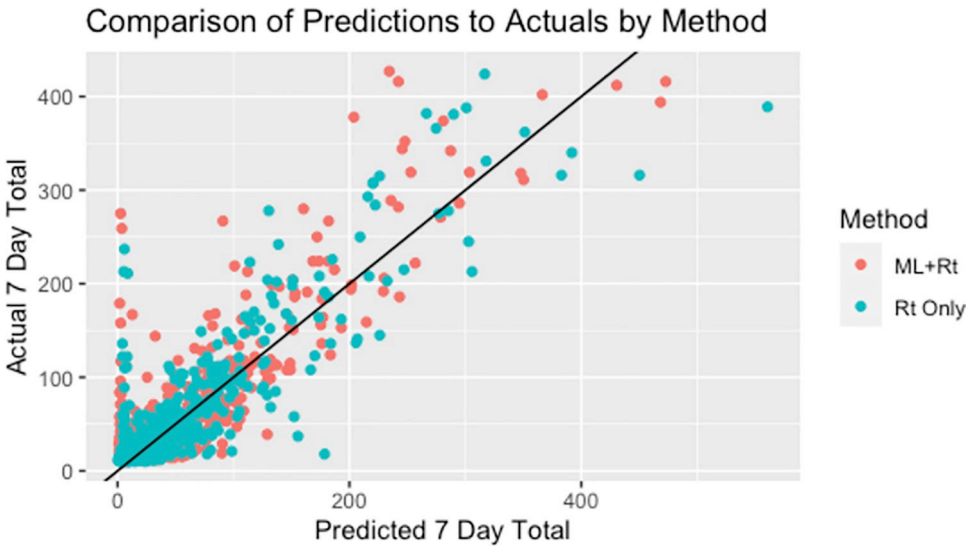

Fig 4. A comparison of actual 7-day case totals and predicted 7-day cases totals for the ML+R_t methods.

https://doi.org/10.1371/journal.pone.0259538.g004

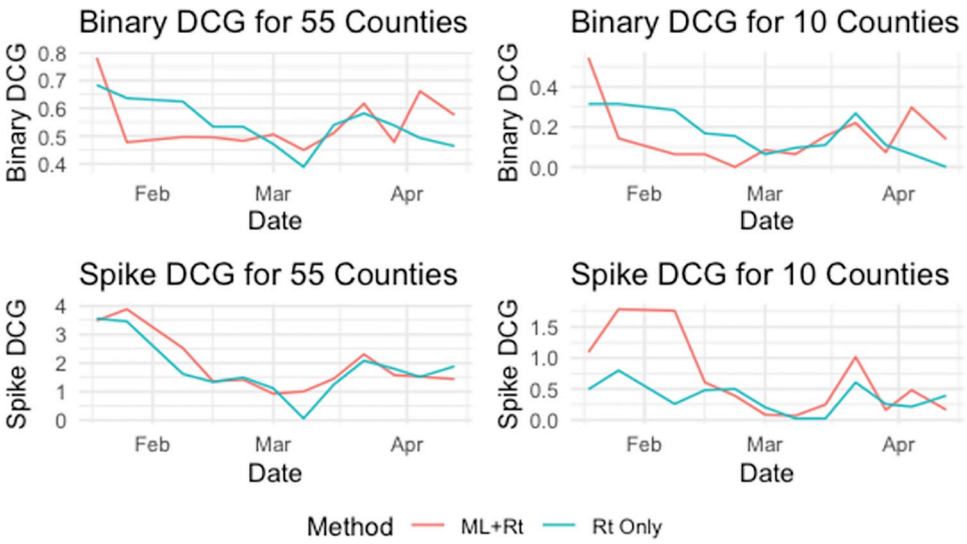

Fig 5. A comparison of the ML+R_t and R_t Only methods with respect to Binary DCG and Spike DCG over the 17-week evaluation period for both 10 and 55 county recommendations.

https://doi.org/10.1371/journal.pone.0259538.g005

Assessing rural vs non-rural results

Critically important is analysis on the performance of the two forecasting strategies in rural compared with more urban counties in WV. Correlations between predicted 7-day positive case totals and actual 7-day positive case totals are higher for non-rural counties than rural counties for both methods (Table 2).

For rural areas, the two methods perform similarly with the ML+R_t method slightly outperforming R_t Only in regard to Spike DCG (Fig 6). For non-rural areas, ML+R_t outperforms Rt Only for both DCG metrics (Table 2). The R_t Only Method performs well when identifying counties in the top 10, but ML+R_t method identifies larger spikes in the top 10 recommendations.

A secondary analysis shows that the ML+R_t method recommends for enhanced SARS-CoV-2 testing more non-rural counties than rural counties in the top 10 rankings during January and February when compared to the R_t Only method. The opposite occurs during the March and April time period during which the R_t Only method recommends more non-rural counties in the top 10 compared to the ML+R_t methods. When coupled with decreasing number of tests, leading to lower daily incidence this alleviates any concern of bias of the ML method on rural counties.

Discussion

In this study, two methods to predict short term incidence of SARS-CoV-2 infection were deployed for purposes of identifying West Virginia counties that might benefit from enhanced

Table 1. A comparison of total both DCG metrics for recommendations of 10 counties and 55 counties for the ML+R_t methods implemented.

		Binary DCG	Spike DCG
55 Counties	ML+R_t	42.50	22.90
	R_t Only	41.83	21.18
10 Counties	ML+R_t	11.88	7.87
	R_t Only	12.59	4.26

https://doi.org/10.1371/journal.pone.0259538.t001

Table 2. A comparison correlation of 7-day positive case totals and 7-day actual case, and both DCG metrics (total) for the ML+R_t methods implemented when viewed by rural and non-rural counties.

		Correlation	Binary DCG	Spike DCG
Rural	ML+R_t	0.690	4.12	0.84
	R_t Only	0.710	6.07	0.76
Non-Rural	ML+R_t	0.867	7.77	7.03
	R_t Only	0.862	6.52	3.50

https://doi.org/10.1371/journal.pone.0259538.t002

SARS-CoV-2 testing. One method, R_t Only, utilizes the Cori model [8], assuming that all positive cases are known. In contrast, the ML+R_t method utilizes R_t as an input value, but bases predictions on an LSTM framework that utilizes other factors such as population size.

Our results demonstrate that both methods perform well. The ML+R_t out performs the R_t only method when it comes to recommending larger spikes in the top recommendations. The implementation of the ML+R_t method is novel as it is utilizing epidemiological underpinnings while exploiting other information such as county population, minimum and maximum values of R_t, variability in R_t, and other information that may, or may not be useful in predicting out breaks.

Each of the methods for incidence prediction have strengths and weaknesses. The R_t Only method only assumes that all positive cases are known. However, in practice, this assumption is unreasonable and highlights some of the problems with applying the standard Cori R_t model to SARS-CoV-2 data. The R_t Only approach relies on the most recent testing data available, and our daily incidence I_t represents the number of positive test results from tests performed on day t. Publicly reported case numbers [18] typically represent the number of positive test results reported on the respective day, but the lag time from test procurement varies. Using the day tests were procured eliminates one additional source of variability and brings our proxy for the "serial interval" closer to the relevant distribution (which would be the infectivity profile–see [11, 12, 19]). However, this raises a practical issue in that data for day t is typically incomplete on day t and is reported gradually over several days. To address this issue,

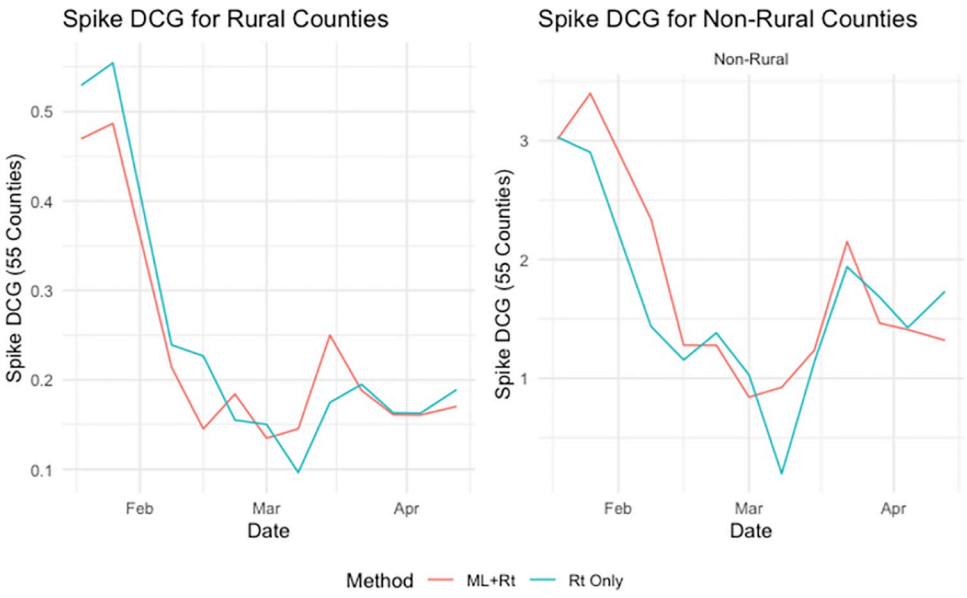

Fig 6. A comparison of Spike DCG for both rural and non-rural counties.

https://doi.org/10.1371/journal.pone.0259538.g006

SARS-CoV-2 incidence are estimated using data from 3 days prior (τ_{report} = *incidence at t*−3 days). For example, the weekly total reported on day *t* = May 12, 2021 represents the week ending on May 9, 2021, and it is this incidence that is used to predict SARS-C0V-2 incidence for the subsequent 7 days.

A second issue with the R_t Only method is that a reliable record of imported cases is not available as they are a theoretical concept in this model. In practical settings, the term "imported" is to be taken in a (very) broad sense. There are a number of situations that have a similar effect.

1. True "exogenous" cases likely occurred due to county residents traveling for school or holidays [3, 4]. There are numerous anecdotal instances in the media but no consistent methodology or documentation of such cases. Commuters from one county to another or out of state could be susceptible to outbreaks outside of their "home" geographic area.
2. "Institutional" or "congregate setting" cases, occur (rather, are identified) over a short time in closed or limited access facilities. Congregate setting outbreaks have somewhat similar features; however, it is not obvious whether individuals infected in congregate settings (e.g., nursing homes) cause new infections in the community as these individuals have limited community access.
3. Finally, significant variability over time of test availability and policies (e.g., limited test availability early in the pandemic, prioritizing resources for vaccine rollout to the detriment of testing availability) complicates the role of the observed incidence as an estimator of the true number of infections.
4. Severity of a disease leading to hospitalization or other interventions that allows for insight into a group that was not previously being tested.

To address these issues, a Bayesian credible interval is used to better define the number of imported cases in the R_t Only method. The proposed iterative fitting technique provides a better estimate of the number of imported cases that will be observed.

The ML+R_t value suffers from issues with practical implementation as well. The same issues with data quality from testing lags can be found when using any data driven method to forecast cases. In addition, there are known problem of using neural networks and deep learning methods when sample sizes are not extremely large. Our approach which predicts using a model that is trained from all combinations of counties and time points takes advantage of the 55 counties over the 365+ days of observed data. Early on in a pandemic it would be unreasonable to think an LSTM or many data driven methods could be used and would be reliable due to a limited number of data point. Therefore, early in the pandemic, our results show the stability of the dynamical model underlying the R_t Only method is reliable once the serial interval could be constructed as the Bayesian approach of the R_t Only method utilizes the serial interval to create an informed prior distribution of spread. For this reason, the LSTM method was not incorporated until October 2020 and only presented in this study from January through April 2021, a time period at which the SARS-CoV-2 epidemic in West Virginia was well established and just before the new Delta variant became established (only one case of Delta was identified during the study period). As the ML+R_t method utilizes all data available, it is less predictive during times that diagnostic testing is erratic (e.g., school breaks, testing supply shortages, etc) (Fig 2). The R_t Only method is able to adjust predictions in a quicker time frame Figs 4 and 5 demonstrate a sharp decrease in performance of the ML+R_t method in February at which time there was a sharp decrease is SARS-CoV-2 diagnostic testing. Again, the R_t Only method is the recommended approach when drastic changes in testing occur and doing so until testing stabilizes.

As seen during the SARS-COV-2 pandemic, situations are dynamic and models must be built to account for the changing landscape of the data and inputs available. With this in mind, extensions of this work should consider vaccination rates, population distributions, vaccine hesitancy, and baseline testing access to better predict outbreaks and target testing. A combination of vaccine information could account for decrease testing and smaller number of cases in models such as the ML+R_t method can adjust for this new input and do so in ways that cannot be accounted for using the Rt Only method. Furthermore, this could lead to interesting results in both identification of not only outbreaks but areas for potential variants and the possibility to use model averaging techniques to create an optimized rule that utilizes both methods. If observed, clinical data on patients (e.g. symptomatic/asymptomatic as percentages) could also be recorded and utilized as an input in the model. Note, that anything utilized in the model must be known during the forecast period, thus information that is dynamic would also have to be forecasted.

The approaches proposed in this work provide a framework for forecasting outbreaks at a local level that utilizes two different approaches. The first is a model based on epidemiological theory, while the second is a machine learning approach that simultaneously considers historic trends and other inputs. Both methods are useful specifically the R_t Only method when data is limited, while the ML+R_t method performs well when data has been collected and a historic perspective can be presented.

Limitations

This study addressed the West Virginia SARS-CoV-2 epidemic from January–April 2021. At that time, only one case of the Delta variant had been detected, therefore, our models do not address prediction of new SARS-CoV-2 incidence when Delta is the prevalent variant. As the Delta variant has unique epidemiologic characteristics compared to earlier SARS-CoV-2 variants such as a shortened serial interval which influences calculation of R_t, models must be adjusted as new more virulent strains of SARS-CoV-2 appear in the population [7]. This study also looks at West Virginia specifically, though the techniques could be applied broadly. As the data in this study is specifically from WVDHHR, it is unable to be compared directly to publicly available sources from other states, limiting the implications of this specific model.

Conclusion

This study provides important information on strategies for predicting near-term increases in SARS-CoV-2 incidence, and hence, for targeting SARS-CoV-2 testing. A new approach is proposed, R_t Only, that utilizes the estimation of the reproduction number to provide recommendations on county-specific areas where outbreaks will likely occur. A second approach is also proposed, ML+R_t, utilizing LSTM models that consider epidemiological statistics such as R_t, county population information, and time series trends including information on major holidays to forecast outbreaks and create county recommendations. Comparison of the two approaches shows the top 10 recommendations produced by the ML+R_t method outperform the R_t Only method over the period of this study. Our data suggest that traditional epidemiological modeling can be enhanced by modern machine learning tools to inform decisions on where to target SARS-CoV2 testing.

Supporting information

S1 Appendix. Distribution and expectation of daily incidence.
(PDF)

Acknowledgments

The authors would like to thank the West Virginia Department of Health and Human Resources, West Virginia's Governors Joint Inter-Agency Task-Force on COVID-19 Vaccination, and Stacey Whanger. Finally, the authors would like to recognize and thank the men and women who have been on the front lines testing and treating patients during the COVID-19 pandemic.

Author Contributions

Conceptualization: Bradley S. Price, Maryam Khodaverdi, Adam Halasz, Brian Hendricks, Wesley Kimble, Gordon S. Smith, Sally L. Hodder.

Data curation: Bradley S. Price, Maryam Khodaverdi, Adam Halasz, Wesley Kimble.

Formal analysis: Bradley S. Price, Maryam Khodaverdi, Adam Halasz.

Funding acquisition: Sally L. Hodder.

Investigation: Adam Halasz, Gordon S. Smith, Sally L. Hodder.

Methodology: Bradley S. Price, Maryam Khodaverdi, Adam Halasz, Brian Hendricks, Wesley Kimble, Gordon S. Smith, Sally L. Hodder.

Project administration: Wesley Kimble, Sally L. Hodder.

Resources: Brian Hendricks, Sally L. Hodder.

Software: Bradley S. Price, Maryam Khodaverdi, Adam Halasz.

Supervision: Sally L. Hodder.

Visualization: Bradley S. Price, Maryam Khodaverdi, Adam Halasz.

Writing – original draft: Bradley S. Price, Maryam Khodaverdi, Adam Halasz, Brian Hendricks, Wesley Kimble, Gordon S. Smith, Sally L. Hodder.

Writing – review & editing: Bradley S. Price, Maryam Khodaverdi, Adam Halasz, Brian Hendricks, Wesley Kimble, Gordon S. Smith, Sally L. Hodder.

References

1. Wiersinga W, Rhodes A, Cheng A, Peacock S, Prescott H. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. JAMA. 2020 Aug; 324(8): p. 782–793. https://doi.org/10.1001/jama.2020.12839 PMID: 32648899
2. Masters N, Shih S, Bukoff AAK, Kobayashi L, Miller A, Harapan H, et al. Social distancing in response to the novel coronavirus (COVID-19) in the United States. PloS One. 2020 Sep; 15(9): p. e0239025. https://doi.org/10.1371/journal.pone.0239025 PMID: 32915884
3. Bradford J, Coe E, Enomoto K, White M. COVID-19 and rural communities: Protecting rural lives and health. In: McKinsey & Company [Internet]. 2020 [cited 2021 Aug 13]. Available from: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/covid-19-and-rural-communities-protecting-rural-lives-and-health.
4. Mueller J, McConnell K, Burow P, Pofahl K, Merdjanoff A, Farrell J. Impact of the COVID-19 Pandemic on Rural America. Proceedings of the National Academy of Sciences. 2021 Jan; 118(1). https://doi.org/10.1073/pnas.2019378118 PMID: 33328335
5. Cyr M, Etchin A, Guthrie B, Benneyan J. Access to specialty healthcare in urban vs. rural US populations: a systematic literature review. BMC health services research. 2019 Dec; 19(1). https://doi.org/10.1186/s12913-019-4815-5 PMID: 31852493
6. Fisman D, Tuite A. Evaluation of the relative virulence of novel SARS-CoV-2 variants: a retrospective cohort study in Ontario, Canada. CMAJ. 2021 Oct. https://doi.org/10.1503/cmaj.211248 PMID: 34610919

7. Li B, Deng A, Li K, Hu Y, Li Z, Xiong Q, et al. Viral infection and transmission in a large well-traced outbreak caused by the Delta SARS-CoV-2 variant. medRxiv. 2021 Jan. https://doi.org/10.1101/2021.07.07.21260122
8. Cori A, Ferguson N, Fraser C, Cauchemez S. A new framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. American Journal of Epidemiology. 2013 Nov; 178(9): p. 1505–1512. https://doi.org/10.1093/aje/kwt133 PMID: 24043437
9. Thompson R, Stockwin J, van Gaalen R, Polonsky J, Kamvar Z, Demarsh P, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics. 2019 Dec; 29: p. 100356. https://doi.org/10.1016/j.epidem.2019.100356 PMID: 31624039
10. Mishra S, Valka F. ImperialCollegeLondon/covid19model: Nature, 2020 https://www.nature.com/articles/s41586-020-2405-7; 2020 [cited 2021 Aug 13]. Database: Zenodo [Internet]. Available from: https://zenodo.org/record/3888697#.YWN5SNrMJPY.
11. Gostic K, McGough L, Baskerville E, Abbott S, Joshi K, C T, et al. Practical considerations for measuring the effective reproductive number Rt. PLoS Computational Biology. 2020 Dec; 16(12): p. e1008409. https://doi.org/10.1371/journal.pcbi.1008409 PMID: 33301457
12. Challen R, Brooks-Pollock E, Tsaneva-Atanasova K, Danon L. Meta-analysis of SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number. MedRxiv: [Preprint]; 2020 [cited 2021 Aug 13]. Available from: https://www.medrxiv.org/content/10.1101/2020.11.17.20231548v2. https://doi.org/10.1101/2020.11.17.20231548
13. Flaxman S, Mishra S, Gandy A, Unwin J, Mellan T, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020 Aug; 584(7820): p. 257–261. https://doi.org/10.1038/s41586-020-2405-7 PMID: 32512579
14. Hochreiter S, Schmidthuber J. Long short-term memory. Neural computation. 1997 Nov; 9(8): p. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 PMID: 9377276
15. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems. 2002 Oct; 20(4): p. 422–446. https://doi.org/10.1145/582415.582418
16. Rural-Urban Continuum Codes (RUCC); [cited 2021 Aug 13]. In: U.S. Department of Agriculture [Internet]. Available from: HYPERLINK https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/.
17. Khodaverdi M. Covid19_County_Prediction Repository; 2021 [cited 2021 Oct 14].: Github [Internet]. Available from: HYPERLINK https://github.com/MKhodaverdi/Covid19_County_Prediction.
18. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet. 2020 May; 20(5): p. 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1 PMID: 32087114
19. Britton T, Scalia Tomba G. Estimation in emerging epidemics: Biases and remedies. Journal of the Royal Society Interface. 2019 Jan; 16(150): p. 20180670. https://doi.org/10.1098/rsif.2018.0670 PMID: 30958162