

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

3

4 ^{1*}Bradley S. Price, ²Maryam Khodaverdi, ³Adam Halasz, ⁴Brian Hendricks, ²Wesley Kimble,

5 ⁴Gordon S. Smith, ^{2,5}Sally L. Hodder

6

7

8 ¹ Management Information Systems Department, West Virginia University, Morgantown, West
9 Virginia

10

11 ² West Virginia Clinical and Translational Science Institute, Morgantown, West Virginia

12

13 ³ School of Mathematics and Data Science, West Virginia University, Morgantown, West Virginia

14

15 ⁴ Department of Epidemiology and Biostatistics, West Virginia University, Morgantown, West
16 Virginia

17

18 ⁵ West Virginia University School of Medicine, Morgantown, West Virginia

19

20 * Corresponding Author

21 E-Mail: brad.price@mail.wvu.edu (BP)

22 Abstract

23 During the COVID-19 pandemic, West Virginia developed an aggressive SARS-CoV-2 testing
24 strategy which included utilizing pop-up mobile testing in locations anticipated to have near-
25 term increases in SARS-CoV-2 infections. In this study, we describe and compare two methods
26 for predicting near-term SARS-CoV-2 incidence in West Virginia counties. The first method, R_t
27 Only, is solely based on producing forecasts for each county using the daily instantaneous
28 reproductive numbers, R_t . The second method, ML+ R_t , is a machine learning approach that
29 uses a Long Short-Term Memory network to predict the near-term number of cases for each
30 county using epidemiological statistics such as R_t , county population information, and time
31 series trends including information on major holidays, as well as leveraging statewide COVID-19
32 trends across counties and county population size. Both approaches used daily county-level
33 SARS-CoV-2 incidence data provided by the West Virginia Department Health and Human
34 Resources beginning April 2020. The methods are compared on the accuracy of near-term
35 SARS-CoV-2 incidence predictions by county over 17 weeks from January 1, 2021- April 30,
36 2021. Both methods performed well (correlation between forecasted number of cases and the
37 actual number of cases week over week is 0.872 for the ML+ R_t method and 0.867 for the R_t
38 Only method) but differ in performance at various time points. Over the 17-week assessment
39 period, the ML+ R_t method outperforms the R_t Only method in identifying larger spikes. We also
40 find that both methods perform adequately in both rural and non-rural predictions. Finally, we
41 provide a detailed discussion on practical issues regarding implementing forecasting models for
42 public health action based on R_t , and the potential for further development of machine learning
43 methods that are enhanced by R_t .

44 Introduction

45 Rural communities in the United States (US) have been heavily impacted by the novel coronavirus
46 (SARS-CoV-2) pandemic. SARS-CoV-2 related deaths have occurred disproportionately among
47 rural areas of the US, and negative impacts on health and economic well-being have been
48 described to be more severe among rural populations (Bradford, Coe, Enomoto, & White, 2020)
49 (Mueller, McConnell, & Burow, 2021) (Cyr, Etchin, & Guthrie, 2019). Persons living in rural
50 communities often have multiple barriers to health care and laboratory diagnostic testing due to
51 geographic, transportation, and cost.

52

53 Early in the COVID-19 pandemic, the state of West Virginia (WV) provided county-specific data
54 on SARS-CoV-2 testing results so that daily instantaneous reproductive numbers (R_t) could be
55 calculated for each WV county to indicate viral transmission dynamics. An aggressive SARS-CoV-
56 2 testing strategy was implemented that included static as well as mobile testing units. The Rapid
57 Acceleration of Diagnostics in Underserved Populations (RADx-UP), funded by the National
58 Institutes of Health, provided the opportunity to deliver pop-up mobile testing in those areas
59 predicted to have the greatest increases in SARS-CoV-2 incidence. The objective was to increase
60 testing in those communities most likely to have a near-term (within 7-10 days) increase in
61 COVID-19 cases, thereby potentially providing early identification of SARS-CoV-2 infected
62 persons who may then quarantine more rapidly in an effort to blunt the anticipated increase in
63 new cases.

64

65 Two strategies to predict near-term increases in SARS-CoV-2 cases were developed using recent
66 county-specific incidence of infections and R_t – one method is a dynamical algorithm-based
67 prediction using R_t and the serial interval while the second method uses a long short-term
68 memory (LSTM) machine learning strategy. The goal was to recommend counties of outbreak for
69 targeted testing. Here we compare accuracy of the two methods to predict short-term increases
70 in county-specific SARS-CoV-2 incidence and discuss conditions favoring one method or the
71 other.

72

73 Data and Methods

74 Data

75 To obtain estimates of near-term increases in SARS-CoV-2 cases, we deployed the likelihood-
76 based model underlying the EpiEstim package in R and developed in Cori et al. (Cori, Ferguson,
77 Fraser, & Cauchemez, 2013) and Thompson et al. (Thompson, et al., 2019). using software
78 provided by Imperial College London (Mishra & Valka, 2020) Two methods were employed: 1)
79 the R_t Only method, a forecast based on the reproduction number and associated serial interval
80 that predicts the future R_t that is then extrapolated to estimate the number of future cases; 2) a
81 Long Short-Term Memory (LSTM) machine learning model (ML+ R_t) that utilizes the reproduction
82 number from the R_t Only method as an input, but also utilizes total cases and population, among
83 other inputs, to predict the total number of cases for a given period of time.

84

85 We received daily reports of all daily COVID-19 polymerase chain reaction (PCR) and antigen
86 testing results conducted in WV since March 2020 directly from the WV Department of Health
87 and Human Resources (WVDHHR). Noteworthy is that all SARS-COV-2 testing data are required
88 to be reported to WVDHHR. Information for each unique patient is collected and contains test
89 procurement date, test result date, patient zip code, patient county of residence, testing site
90 name, county where the test is obtained, and test result. As patients who test positive may be
91 tested multiple times, we only consider the first positive tests on a patient. When applying this
92 filter, we consider data obtained from all testing sites (i.e., hospital, clinic, pharmacy, drive-
93 through, mobile van). The number of daily cases for each county is calculated by adding the lab
94 confirmed cases and clinical confirmed cases after filtering out repeated tests or COVID-19
95 diagnoses. This daily incidence data on first diagnosed infection is the basis for calculation of R_t .

97 R_t Only Method: Producing Short Term Predictions

98 Our R_t Only method relies on the methodology used in the EpiEstim package and the underlying
99 modeling approach of Cori et al (Cori, Ferguson, Fraser, & Cauchemez, 2013) and Thompson et
100 al. (Thompson, et al., 2019). This approach relates the daily incidence (number of new cases) to
101 past cases through an instantaneous reproduction number R_t which characterizes the daily
102 dynamics of transmission reflects a multitude of factors relating to individual and group behavior
103 in the community of interest.

104

105 As a brief review, daily infections within a community occur as independent random events

106 drawn from a Poisson distribution. The probability that exactly k cases occur is $p_k = \frac{\Gamma^k}{k!} e^{-\Gamma}$, and

107 the rate parameter Γ coincides with the average daily incidence, $\langle k \rangle = \Gamma$. In the instantaneous
108 R_t framework, the expected incidence on day t is a product of two quantities, the infection
109 potential and the reproduction number, $\Gamma_t = \Lambda_t R_t$. The infection potential Λ_t summarizes the
110 record of past cases in the community and the typical variation of the infectiousness of an
111 individual over time.

112

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

115

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

116 The serial interval distribution w_s reflects the time course of infectiousness of one infected
117 individual at $s = 1, 2, \dots$ days from the primary infection. It encapsulates the relative increase and
118 decrease of infectiousness of an individual, assuming all other conditions in the community
119 remain unchanged. In practice, the serial interval is typically obtained as the normalized
120 ($\sum_{s=1}^{S_{\max}} w(s) = 1$) distribution of time intervals between known infector-infected pairs. Based on
121 studies done by Gostic et al. and Challen et al., we used a distribution extending over 100 days
122 for the serial interval (Gostic, et al., 2020) (Challen, Brooks-Pollock, Tsaneva-Atanasova, & Danon,
123 2020). The infection potential can be understood as the sum of the expected number of
124 infections on day t , due to past cases in the community, under ideal “steady state” conditions,
125 such that over time, each primary case causes exactly one secondary case.

126

127 The time varying reproduction number, R_t , captures conditions of transmission that are external
128 to the infected individuals and reflect community behavior. In this framework, R_t is a random
129 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
130 determined for each day through Bayesian (maximum a posteriori probability) estimation. The
131 parameters of interest are estimated using incidence data up to and including the current day,
132 I_1, I_2, \dots, I_t as follows:

$$133 \quad a_t = \sum_{s=0}^{\tau-1} I_{t-s} + a_{\text{prior}}, \quad \frac{1}{b_t} = \sum_{s=0}^{\tau-1} \Lambda_{t-s} + \frac{1}{b_{\text{prior}}}, \quad \Lambda_t = \sum_{s=1}^{s_{\text{max}}} I_{t-s} w_s$$

134 This estimated R_t distribution applies to the most recent τ days, but it requires the values of $I_{t'}$
135 for $t' \leq t$ going back to $t' = t - s_{\text{max}}$ where s_{max} is the length of the serial interval distribution.
136 For the serial interval w_s we used the discretized gamma distribution with mean and standard
137 deviation of $t_s = 7.0 \pm 4$ days, provided in the software similar to Cori, Ferguson, Fraser, &
138 Cauchemez (2013).

139 For the serial interval w_s we use a gamma distribution with mean and standard deviation of $\tau_s =$
140 6.99 ± 4.02 days, as given by Flaxman, et al., 2020. Following Cori and Thompson's method, we
141 used a prior distribution consistent with mean and standard deviation equal to 5 ($a_{\text{prior}} =$
142 $1, b_{\text{prior}} = 5$).

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

148

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

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

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

$$157 \quad \langle I_{t+1} \rangle_{R_{t+1}} = \Lambda_{t+1}R_{t+1} \rightarrow \langle \langle I_{t+1} \rangle_{R_{t+1}} \rangle_{a,b} = \Lambda_{t+1} \langle R_{t+1} \rangle_{a,b} = \Lambda_{t+1}ab$$

158 For the purpose of predicting a likelihood range for the daily incidence, we use the CDF of R_{t+1} :

$$159 \quad P(\bar{I}_t \in [I_1, I_2]|a, b, \Lambda) = gamcdf\left(\frac{I_2}{\Lambda}|a, b\right) - gamcdf\left(\frac{I_1}{\Lambda}|a, b\right) = \frac{1}{b^a \Gamma(a)} \int_{I_1/\Lambda}^{I_2/\Lambda} R^{a-1} e^{-\frac{R}{b}} dR$$

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

163

164 **Extrapolation over multiple days:** To go beyond the “next” day, we iterate the one-day
 165 prediction, using predicted values to expand the incidence data. One can reasonably extrapolate

166 the current distribution of R_t to $t + 1$ and any number of days in the future. For the short term
 167 (7 day) predictions discussed here, we assumed the value of the most recent available R_t remains
 168 the same over the prediction interval, $\bar{R}_{t+k} = R_t$.

169
 170 The estimated incidence for day $t + 1$ requires the infection potential on that day Λ_{t+1} , which
 171 is computed based on incidence up to the preceding day t .

172
$$\bar{I}_{t+1|t} = \Lambda_{t+1} \bar{R}_{t+1} = \Lambda_{t+1} R_t,$$

173
$$\Lambda_{t+1} = \sum_{s=1}^{s_{max}} I_{(t+1)-s} w_s = I_t w_1 + I_{t-1} w_2 + \dots + I_{(t+1)-s_{max}} w_{s_{max}}$$

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

176
$$\begin{aligned} \bar{I}_{t+1|t} &= \Lambda_{t+1} R_t & \bar{\Lambda}_{t+2|t} &= \sum_{s=2}^{s_{max}} I_{t+1-s} w_s + \bar{I}_{t+1|t} w_1 \\ \bar{I}_{t+2|t} &= \bar{\Lambda}_{t+2} R_t & \bar{\Lambda}_{t+3|t} &= \sum_{s=3}^{s_{max}} I_{t+2-s} w_s + \sum_{s=1}^2 \bar{I}_{t+2-s} w_s \\ \bar{I}_{t+k|t} &= \bar{\Lambda}_{t+k} R_t & \bar{\Lambda}_{t+k|t} &= \sum_{s=k}^{s_{max}} I_{t+k-s} w_s + \sum_{s=1}^k \bar{I}_{t+k-s} w_s \end{aligned}$$

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

182

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

$$187 \quad I_t = I_t^{(\text{local})} + I_t^{(\text{imported})}$$

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

$$191 \quad a_t = \sum_{s=0}^{\sigma-1} I_{t-s}^{(\text{local})} + a_{\text{prior}}, \quad \frac{1}{b_t} = \sum_{s=0}^{\sigma-1} \Lambda_{t-s} + \frac{1}{b_{\text{prior}}}, \quad \Lambda_t = \sum_{s=1}^{s_{\text{max}}} I_{t-s} w_s$$

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

$$194 \quad R_t \sim \text{gampdf}(a_t, b_t), \quad I_t^{(\text{local})} \sim \text{poisspdf}(R_t \Lambda_t) \rightarrow \langle I_t^{(\text{local})} \rangle = \Lambda_t \langle R_t \rangle = \Lambda_t a_t b_t$$

195 By definition, imported cases cannot be predicted in the R_t model; however, we can identify
196 events when the observed number of new cases vastly exceeds the expectation from local
197 transmission. We use this hindsight to improve our estimate of the reproduction number as
198 follows.

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

202 In an initial pass we compute the a_t, b_t parameters for time point t based on the incidence time
203 series $\{I_\tau\}_{\tau=0,1,\dots,t-1}$. We compute the one-day predicted incidence on day t as described above,
204 using the infection potential Λ_t and the distribution of $\bar{R}_t \equiv R_{t-1}$ (so we do not rely on the
205 knowledge of I_t). We take the value corresponding to the upper $\theta = 95\%$ credible interval as a
206 cutoff and identify the part of the incidence that exceeds the cutoff with imported cases.

$$207 \quad I_t^{(\text{local,high})} = \Lambda_t \text{gaminv}(\theta, a_{t-1}, b_{t-1}), \quad I_t^{(\text{imported,est})} = \max(I_t - I_t^{(\text{local,high})}, 0)$$

208 We use the estimated local incidence $I_t^{(\text{local,est})}$ to provide revised estimates for the reproduction
209 number as described above (also consistent with Cori and Thompson's approach). Finally, we use
210 the resulting R_t parameters for the most recent day and the full incidence to provide revised
211 estimates for days $t + 1, t + 2, \dots, t + k$.

212

213 ML+ R_t Method: Using Long Short-Term Memory (LSTM) Network to Forecast Outbreaks

214 As previously mentioned, the LSTM method implemented in this project is meant to build on the
215 widely used R_t Only approach described in the previous section. The novelty of this LSTM
216 approach is that it provides for the input of epidemiological modeling while taking advantage of
217 cutting-edge machine learning techniques. The combination of the two allows the LSTM model
218 to incorporate the epidemiological principles used to produce the R_t estimate while adding
219 additional information such as temporal and demographic information that can be leveraged
220 with traditional machine learning models. Further, the calculation of R_t using the R_t Only method
221 uses independent data sets for each county in turn creating a unique model for each county that
222 does not consider the impact of possible relationships between counties. By contrast, the ML+ R_t

223 approach uses global trends across counties. By training on all the data, we are not only able to
224 take advantage of global trends, but by including spatial information, we are also able to show
225 how these trends impact specific counties.

226
227 Daily county-specific R_t , summary statistic information on the estimated R_t such as standard
228 deviation, confidence intervals, and the probability of $R_t > 1$ are also provided. We include values
229 of R_t computed using both 7 and 14 day intervals. All these factors along with temporal
230 information such as daily information on whether it is a weekend or not, holiday or not, days
231 passed from last major holiday, days to the next major holiday were utilized as inputs for our
232 model.

233
234 As mentioned previously, due to the length of time it takes to receive a test result (lag time), we
235 had to consider the deflated effect on the positive cases when considering test procurement
236 date. We observed an average lag of 3 days for results to achieve close to actual levels. To
237 mitigate the effect of the testing lag we impute day t , $t-1$, $t-2$ with the actual SARS-CoV-2 cases
238 for days $t-3$, $t-4$, $t-5$ respectively.

239
240 We utilize a Long-Short Term Memory (LSTM) recurrent neural network (Hochreiter &
241 Schmidhuber, 1997), implemented in Python with an Adam optimizer, as our model of interest
242 for this analysis, permitting consideration of all available county-specific input information for
243 the past 7 days with a prediction of the number of positive cases for the county as an output.

244 Other advantages of the LSTM approach are the ability to exploit autocorrelation between time
245 points and the utilization of dropout layers to remove redundant information.

246

247 In general, the LSTM models are more complex versions of recursive neural networks (RNNs).

248 The multi-layer LSTM method deployed here follows the framework described in Figure 1 where

249 the input layer is defined by a matrix combining the number of positive cases for county c at time

250 point t , $Y_{c,t}$, and all inputs for county c at time point t . The inputs then move their way through

251 the network (i.e., through the LSTM layer and dense layer) to obtain an output. The output is

252 defined as, $\hat{Y}_{c,t+7}$, the predicted daily number of cases for county c at time point $t+7$. LSTM can

253 be viewed as a network where information between time points is shared. Each LSTM cell,

254 diagramed in Figure 1, shares two pieces of information with other LSTM cells; the current state

255 of the cell, C_t , and output of the cell, h_t , is calculated with the following formulas given input

256 data, x_t :

257
$$\tilde{C}_t = \sigma'(W_c \cdot [h_{t-1}, x_t] + b_c)$$

258
$$g_t^i = i_t \times \tilde{C}_t$$

259
$$g_t^f = f_t \times C_{t-1}$$

260
$$C_t = g_t^f + g_t^i$$

261
$$h_t = o_t \times \sigma' C_t$$

262 Where, w are the weight variables (traditionally thought of like regression coefficients), and b are

263 the bias variables (traditionally thought of as intercept terms). Activation functions, σ and σ' are

264 non-linear transformation functions such as, sigmoid and hyperbolic tangent. A feature of each

265 cell is input, output, and forget gates. These gates are what give the LSTM, the memory property
266 which allows it to account and adjust for auto correlation. We define:

$$267 \quad f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$268 \quad i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$269 \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

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

272

273 Figure 1: The LSTM framework deployed for the proposed ML+R_t method on right, and structure of each LSTM Cell
274 on left.

275

276 In addition, as we cannot guarantee the importance of the inputs (including R_t and associated
277 summary statistics), we add dropout layers which allow for the identification of important inputs.

278 Using these dropout layers, we filtered out inputs that would be considered insignificant in order
279 to detect the important signals coming from the input data and also protect against overfitting.

280

281 Once predictions for a given week were determined, the summary statistics of the results were

282 produced. Summary statistics included: 1) predicted number of positive cases by county, 2)

283 predicted percent change in cases per 100,000 persons by county compared to the previous

284 week, 3) predicted increase in number of cases compared to the previous week, and 4) predicted

285 number of cases relative to the population size.

286

287 Evaluations of Models in Deployment

288 Metrics and Evaluation

289 To evaluate performance of the two methods, the predicted values for new SARS-CoV-2 cases
290 were benchmarked against the actual number of positive cases recorded for each week from
291 January 1, 2021 through April 30, 2021. As a main goal of these new case forecasts was to target
292 areas for diagnostic testing, we viewed each week's prediction as a recommendation. These
293 recommendations were ranked on many several metrics but most predominately on the
294 percentage increase in cases over the previous week. To evaluate the recommendations, we
295 measured the total discounted cumulative gain (DCG) of each method (Järvelin & Kekäläinen,
296 2002). DCG is a commonly used metric in page ranking calculations and is suitable here as the
297 information shared was used similar to page ranking calculations. As a reminder the goal of this
298 analysis is to recommend counties of increased incidences for intervention (i.e., increased SARS-
299 CoV-2 testing), not to predict the actual number of incidence. DCG provides a metric for
300 comparison of differing recommendation methods, which is how both the ML+ R_t and R_t Only
301 are being used. Unlike most metrics used in machine learning such as squared error or absolute
302 error, larger DCG values indicate better performance.

303

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

308

309 To define the first metric let $\hat{y}_{c,t}$ and $y_{c,t}$ represent the number of predicted cases and actual
310 cases over a 7-day period for the c th county at time point t respectively. To keep from biasing
311 the evaluation towards rural areas with a low incidence, we only consider those with $y_{c,t+1} >$
312 10. 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. We
313 defined the Binary Discounted Cumulative Gain (Binary DCG) of a set of rankings at time point t
314 as:

$$315 \quad \sum_{i=1}^q \frac{I(i \in S_t)}{\ln(i+1)}$$

316 where $I(i \in S_t)$ is an indicator of a correct identification of a top 10 ranking in the actual
317 percentage increases, and q is the number of rankings used in the calculation. For example, if
318 $q = 10$, then $BDCG_t$ would only evaluate the top 10 rankings, in our setting this would be the
319 top 10 counties, returned by a method. One may view B-DCG as a weighted identifier to measure
320 the quality of the rankings for purposes of identifying case increases (or spikes) of the top q
321 recommendations.

322
323 As the closeness of the predicted number of cases to the actual case number, i.e., the “quality”
324 of the prediction, we considered a second metric to consider the quality of the prediction rather
325 than just considering a binary outcome. To accomplish this, we define Spike DCG as:

$$326 \quad \sum_{i=1}^q \frac{y_{i,t+1}}{y_{i,t} \ln(i+1)}$$

327 Spike DCG considers the relative size of the spike for the top q recommendations. While Binary
328 DCG investigates the ability of a method to correctly identify the top 10 counties, Spike DCG

329 places value on the recommendations that are produced by identification of larger spikes. This
330 comparison is of great importance as targeted interventions may only have finite resources to
331 deploy so understanding the level of trust and impact expected by the two methods is of
332 importance.

333
334 As both the R_t Only and ML+ R_t methods are used to recommend county level locations for testing,
335 we also want to investigate the quality of the top recommendations, disregarding the order and
336 quality of the ranked predictions. This evaluation gives a sense of the quality of the
337 recommendations produced by the methods, relative to others.

338
339 Finally, as this study is being deployed in a state with many rural areas, we analyzed any
340 differences in methods between rural and non-rural areas. We used the 2013 Rural-Urban
341 Continuum Codes (RUCC) (Rural-Urban Continuum Codes (RUCC)) which define a rural area as a
342 non-metro area with population under 20,000 and is not adjacent to an urban metro area. To
343 assess quality of the predictions provided by each method, we examined correlations between
344 predicted and actual 7-day positive case totals. We also assess the quality of Binary DCG and
345 Spike DCG in both rural and non-rural areas by investigating the performance of ML+ R_t and R_t
346 Only methods among lower population communities with less access to large healthcare systems.
347 Both R_t Only and ML+ R_t methods were deployed each week from January 1, 2021 through April
348 30, 2021 using all available training data beginning in April 2002 for each of the 55 counties in the
349 state of WV, and resulting county recommendations were retained for comparison against the
350 actual number of cases.

351

352 Results

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

360

361 *Figure 2: Number of SARS-COV-2 tests in the state of West Virginia from April 2020-April 2021.*

362 *Figure 3: Number of SARS-COV-2 tests in the state West Virginia from May 2020- July 2020.*

363

364

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

368

369 Figure 5 compares Binary and Spike DCG for the case of recommending 10 counties (q=10) and
370 55 counties (q=55). Both the R_t Only and ML+R_t methods perform well overall but differ in
371 performance at various time points. In the case of Binary DCG the R_t Only method has better
372 performance, and in the case of Spike DCG the ML+R_t method performs better. Over the 17-week

373 assessment period, the ML+R_t method outperforms the R_t Only method in recommendations
374 with regard to all measures except Binary DCG for q=10 (Table 1). These results show that if users
375 are interested in mitigating outbreaks by identifying larger spikes in the Top 10
376 recommendations, as was the goal of this implementation, the ML+R_t method should be used.

377

378 *Figure 4: A comparison of actual 7-day case totals and predicted 7-day cases totals for the ML and R_t methods*

379

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

382

383

384

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

388

389 **Assessing Rural vs Non-Rural Results**

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

394

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

400 *Figure 6: A comparison of Binary DCG and Spike DCG for both rural and non-rural counties.*

401
402 A secondary analysis shows that the ML+R_t method recommends for enhanced SARS-CoV-2
403 testing more non-rural counties than rural counties in the top 10 rankings during January and
404 February when compared to the R_t Only method. The opposite occurs during the March and April
405 time period during which the R_t Only method recommends more non-rural counties in the top
406 10 compared to the ML+R_t methods. When coupled with decreasing number of tests, leading to
407 lower daily incidence this alleviates any concern of bias of the ML method on rural counties.

408

409 Discussion

410 In this study, we deployed two methods to predict short term incidence of SARS-CoV-2 infection
411 for purposes of identifying West Virginia counties that might benefit from enhanced SARS-CoV-2
412 testing. One method, R_t Only, utilizes the Cori model [5], assuming that all positive cases are
413 known. In contrast, the ML+R_t method utilizes R_t as an input value, but bases predictions on an
414 LSTM framework that utilizes other factors such as population size.

415

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

421
422 Each of the methods for incidence prediction have strengths and weaknesses. The R_t Only
423 method only assumes that all positive cases are known. However, in practice, this assumption is
424 unreasonable and highlights some of the problems with applying the standard Cori R_t model to
425 SARS-CoV-2 data. The R_t Only approach relies on the most recent testing data available, and our
426 daily incidence I_t represents the number of positive test results from tests performed on day t .
427 Publicly reported case numbers (Dong, Du, & Gardner, 2020) typically represent the number of
428 positive test results reported on the respective day, but the lag time from test procurement
429 varies. Using the day tests were procured eliminates one additional source of variability and
430 brings our proxy for the “serial interval” closer to the relevant distribution (which would be the
431 infectivity profile – see (Challen, Brooks-Pollock, Tsaneva-Atanasova, & Danon, 2020) (Britton &
432 Scalia Tomba, 2019) (Gostic, et al., 2020)). However, this raises a practical issue in that data for
433 day t is typically incomplete on day t and is reported gradually over several days. To address this
434 issue, we estimate SARS-CoV-2 incidence using data from 3 days prior ($\tau_{\text{report}} =$
435 *incidence at $t - 3$ days*). For example, the weekly total reported on day $t =$ May 12, 2021
436 represents the week ending on May 9, 2021, and it is this incidence that is used to predict SARS-
437 COV-2 incidence for the subsequent 7 days.

438

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

443 1. True “exogenous” cases likely occurred due to county residents traveling for school or
444 holidays. [1][2]. There are numerous anecdotal instances in the media but no consistent
445 methodology or documentation of such cases. Commuters from one county to another
446 or out of state could be susceptible to outbreaks outside of their “home” geographic area.

447 2. “Institutional” or “congregate setting” cases, occur (rather, are identified) over a short
448 time in closed or limited access facilities. Congregate setting outbreaks have somewhat
449 similar features; however, it is not obvious whether individuals infected in congregate
450 settings (e.g., nursing homes) cause new infections in the community as these individuals
451 have limited community access.

452 3. Finally, significant variability over time of test availability and policies (e.g., limited test
453 availability early in the pandemic, prioritizing resources for vaccine rollout to the
454 detriment of testing availability) complicates the role of the observed incidence as an
455 estimator of the true number of infections.

456 4. Severity of a disease leading to hospitalization or other interventions that allows for
457 insight into a group that was not previously being tested.

458

459 To address these issues, we use the Bayesian credible interval to better define the number of
460 imported cases in the R_t Only method. By the iterative fitting technique proposed we are able
461 to better estimate the number of imported cases that will be observed.

462

463 The $ML+R_t$ value suffers from issues with practical implementation as well. The same issues with
464 data quality from testing lags can be found when using any data driven method to forecast cases.

465 In addition, there are known problem of using neural networks and deep learning methods when

466 sample sizes are not extremely large. Our approach which predicts using a model that is trained

467 from all combinations of counties and time points takes advantage of the 55 counties over the

468 365+ days of observed data. Early on in a pandemic it would be unreasonable to think an LSTM

469 or many data driven methods could be used and would be reliable due to a limited number of

470 data point. Therefore, early in the pandemic, our results show the stability of the dynamical

471 model underlying the R_t Only method is reliable once the serial interval could be constructed as

472 the Bayesian approach of the R_t Only method utilizes the serial interval to create an informed

473 prior distribution of spread. For this reason, the LSTM method was not incorporated until October

474 2020 and only presented in this study from January through April 2021, a time period at which

475 the SARS-CoV-2 epidemic in West Virginia was well established and just before the new Delta

476 variant became established (only one case of Delta was identified during the study period). As

477 the $ML+R_t$ method utilizes all data available, it is less predictive during times that diagnostic

478 testing is erratic (e.g., school breaks, testing supply shortages, etc) (Figure A1). The R_t Only

479 method is able to adjust predictions in a quicker time frame Figures 3 and Figure 4 demonstrate

480 a sharp decrease in performance of the $ML+R_t$ method in February at which time there was a

481 sharp decrease is SARS-CoV-2 diagnostic testing. Again, we recommend using the R_t Only
482 approach when drastic changes in testing occur and doing so until testing stabilizes.

483

484 As we have seen during the SARS-COV-2 pandemic, situations are dynamic and models must be
485 built to account for the changing landscape of the data and inputs available. With this in mind,
486 extensions of this work should consider vaccination rates, population distributions, vaccine
487 hesitancy, and baseline testing access to better predict outbreaks and target testing. A
488 combination of vaccine information could account for decrease testing and smaller number of
489 cases in models such as the $ML+R_t$ method can adjust for this new input and do so in ways that
490 cannot be accounted for using the R_t Only method. Furthermore, this could lead to interesting
491 results in both identification of not only outbreaks but areas for potential variants and the
492 possibility to use model averaging techniques to create an optimized rule that utilizes both
493 methods.

494

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

501

502 Limitations

503 This study addressed the West Virginia SARS-CoV-2 epidemic from January – April 2021. At that
504 time, only one case of the Delta variant had been detected, therefore, our models do not address
505 prediction of new SARS-CoV-2 incidence when Delta is the prevalent variant. As the Delta variant
506 has unique epidemiologic characteristics compared to earlier SARS-CoV-2 variants such as a
507 shortened serial interval which influences calculation of R_t , models must be adjusted as new more
508 virulent strains of SARS-CoV-2 appear in the population (Baisheng, et al., 2021).

509

510 Conclusion

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

522

523 Acknowledgements

524 The project described was supported by the National Institute Of General Medical
525 Sciences, 5U54GM104942-04 and 5U54GM104942-05S3. The content is solely the responsibility
526 of the authors and does not necessarily represent the official views of the NIH. The authors
527 would like to thank the West Virginia Department of Health and Human Resources, West
528 Virginia's Governors Joint Inter-Agency Task-Force on COVID-19 Vaccination, and Stacey
529 Whanger. Finally, the authors would like to recognize and thank the men and women who have
530 been on the front lines testing and treating patients during the COVID-19 pandemic.

531

532 References

- 533 Baisheng, L., Aiping, D., Kuibiao, L., Yao, H., Zhencui, L., Qianling, X., & et, a. (2021). Viral
534 infection and transmission in a large well-traced outbreak caused by the Delta SARS-
535 CoV-2 variant. *medRxiv*(Jan 1). doi:10.1101/2021.07.07.21260122
- 536 Bradford, J., Coe, E., Enomoto, K., & White, M. (2020). *Rural Communities: Protecting Rural*
537 *Lives and Health*. Retrieved from McKinsey:
538 [https://www.mckinsey.com/industries/healthcare-systems-and-services/our-](https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/covid-19-and-rural-communities-protecting-rural-lives-and-health)
539 [insights/covid-19-and-rural-communities-protecting-rural-lives-and-health](https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/covid-19-and-rural-communities-protecting-rural-lives-and-health)
- 540 Britton, T., & Scalia Tomba, G. (2019). Estimation in emerging epidemics: Biases and remedies. *J*
541 *R Soc Interface*, 16, 20180670.
- 542 Challen, R., Brooks-Pollock, E., Tsaneva-Atanasova, K., & Danon, L. (2020). *Meta-analysis of*
543 *SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19*

- 544 *reproduction number*. Retrieved November 2020, from medRxiv preprint:
- 545 <https://doi.org/10.1101/2020.11.17.20231548>
- 546 Cori, A., Ferguson, N., Fraser, C., & Cauchemez, S. (2013). A new framework and Software to
- 547 Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of*
- 548 *Epidemiology*, 179(9), 1505-1512.
- 549 Cyr, M., Etchin, A., & Guthrie, B. (2019). Access to specialty healthcare in urban vs. rural US
- 550 populations: a systematic literature review. *BMC Health Serv Res*, 19, 974.
- 551 Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in
- 552 real time. *Lancet*, 20(5), 533-534.
- 553 Flaxman, S., Mishra, S., Gandy, A., Unwin, J., Mellan, T., Coupland, H., . . . Zhu, H. (2020).
- 554 Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe.
- 555 *Nature*, 584(7820), 257-261.
- 556 Gostic, K., McGough, L., Baskerville, E., Abbott, S., Joshi, K., & Tedijanto C. (2020). PRactical
- 557 considerations for measuring the effective reproductive number Rt. *PLoS Computational*
- 558 *Biology*, 16(12), e1008409.
- 559 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8),
- 560 1735-1780.
- 561 Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM*
- 562 *Transactions on Information Systems*, 20(4), 422-446.
- 563 Mishra, S., & Valka, F. (2020, Jan). *ImperialCollegeLondon/covid19model: Nature, 2020*
- 564 <https://www.nature.com/articles/s41586-020-2405-7>. Retrieved Aug 13, 2021, from
- 565 <https://doi.org/10.5281/zenodo.3888697>

566 Mueller, J., McConnell, K., & Burow, P. (2021). Impact of the COVID-19 Pandemic on Rural
567 America. *PNAS*, *118*, e2019378118.

568 *Rural-Urban Continuum Codes (RUCC)*. (n.d.). Retrieved from [https://www.ers.usda.gov/data-](https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/)
569 [products/rural-urban-continuum-codes/](https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/)

570 Thompson, R., Stockwin, J., van Gaalen, R., Polonsky, J., Kamvar, Z., Demarsh, P., . . . Cori, A.
571 (2019). Improved inference of time-varying reproduction numbers during infectious
572 disease outbreaks. *Epidemics*, *29*, 100356.

573 Wallinga, J., & Teunis, P. (2004). Different epidemic curves for severa acute respiratory
574 syndrome reveal simila impacts of control measures. *Americal Journal of Epidemiology*,
575 *160*(6), 509-516.

576

577

578 Tables

579

580 *Table 1: A comparison of total both DCG metrics for recommendations of 10 counties and 55 counties for the ML*
 581 *and 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

582

583

584 *Table 2: A comparison correlation of 7-day positive case totals and 7-day actual case, and both DCG metrics (total)*
 585 *for the ML and 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

586

587

588 Appendix

589

590 Distribution and expectation of daily incidence

591

592 The daily incidence has a Poisson distribution with parameter $\Lambda_t R_t$. R_t is represented as a

593 random variable following a gamma distribution with parameters a, b :

594
$$P(k|R_t, \Lambda_t) = \frac{(\Lambda_t R_t)^k}{k!} e^{-\Lambda_t R_t}$$

595
$$f(R_t|a, b) = \frac{1}{b^a \Gamma(a)} R^{a-1} e^{-\frac{R}{b}}$$

596 where $\Gamma(a)$ is the usual Gamma function defined as:

597
$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt \rightarrow \Gamma(n+1) = n! \text{ (if } n \text{ is a positive integer)}$$

598
$$\Gamma(z+1) = z\Gamma(z)$$

599

600 Denote by $C_{a,b}$ the normalization constant for the Gamma distribution:

601
$$\frac{1}{C_{a,b}} = \int_0^{\infty} dR R^{a-1} e^{-\frac{R}{b}} = b^a \int_0^{\infty} du u^{a-1} e^{-u} = b^a \Gamma(a)$$

602
$$C_{a,b} = \frac{1}{b^a \Gamma(a)}$$

603
$$C_{a+1,b} = \frac{1}{b^{a+1} \Gamma(a+1)} = \frac{1}{ab} C_{a,b}$$

604

605 The PMF of the expected number of cases is obtained by integrating over the values of R_t :

606
$$P(k|\Lambda_t, a, b) = \int_0^{\infty} dR \frac{(\Lambda_t R)^k}{k!} e^{-\Lambda_t R} \cdot C_{a,b} \cdot R^{a-1} e^{-\frac{R}{b}}$$

607 The integrand is proportional to a gamma distribution with parameters $a' = a + k$, $\frac{1}{b'} = \frac{1}{b} + \Lambda_t$

608
$$P(k) = \frac{\Lambda_t^k C_{a,b}}{k!} \int_0^{\infty} dR R^{k+a-1} e^{-R(\Lambda_t + \frac{1}{b})} = C_{a,b} \frac{\Lambda_t^k}{k!} \frac{\Gamma(a+k)}{(\Lambda_t + \frac{1}{b})^{a+k}} = \frac{(\Lambda_t b)^k}{(\Lambda_t b + 1)^{a+k}} \cdot \frac{\Gamma(a+k)}{k! \Gamma(a)}$$

609 Or (use $\Gamma(z+1) = z\Gamma(z)$)

610
$$P(k|a, b) = \frac{1}{(b\Lambda_t + 1)^a} \left(\frac{b\Lambda_t}{b\Lambda_t + 1}\right)^k \prod_{j=1}^k \frac{(a+j)}{j}$$

611

612 The expected number of new infections follows from working out the Gamma-Poisson
613 distribution and coincides with the infection potential multiplied by the expected R

614
$$\langle I_t \rangle = \Lambda_t R_t \rightarrow \langle \langle I_t \rangle (R_t) \rangle_{R_t} = \Lambda_t \langle R_t \rangle = \Lambda_t a b$$

615
$$\langle I \rangle = \sum_{k=1}^{\infty} k \int_0^{\infty} dR \frac{(\Lambda_t R)^k}{k!} e^{-\Lambda_t R} \cdot C_{a,b} \cdot R^{a-1} e^{-\frac{R}{b}} = \int_0^{\infty} dR \Lambda_t R \sum_{\ell=0}^{\infty} \frac{(\Lambda_t R)^{\ell}}{\ell!} e^{-\Lambda_t R} \cdot C_{a,b} \cdot R^{a-1} e^{-\frac{R}{b}} = \Lambda_t C_{a,b} \int_0^{\infty} dR \cdot R^a e^{-\frac{R}{b}}$$

616