

1 A modular approach to integrating multiple data sources 2 into real-time clinical prediction for pediatric diarrhea

3 **Ben J. Brintz^{1,6*}, Benjamin Haaland², Joel Howard³, Dennis L. Chao⁴, Joshua L.
4 Proctor⁴, Ashraful I. Khan⁵, Sharia M. Ahmed⁶, Lindsay T. Keegan¹, Tom Greene¹,
5 Adama Mamby Keita⁷, Karen L. Kotloff⁸, James A. Platts-Mills⁹, Eric J. Nelson^{10,11},
6 Adam C. Levine¹², Andrew T. Pavia³, Daniel T. Leung^{6,13*}**

*For correspondence:

ben.brintz@hsc.utah.edu (BB);
Daniel.Leung@utah.edu (DTL)

[†]These authors contributed
equally to this work

7 ¹Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake
8 City, USA; ²Population Health Sciences, University of Utah, Salt Lake City, USA; ³Division of
9 Pediatric Infectious Diseases, University of Utah, Salt Lake City, USA; ⁴Institute of Disease
10 Modeling, Bill and Melinda Gates Foundation, Seattle, USA; ⁵International Centre for
11 Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka, Bangladesh.; ⁶Division of
12 Infectious Diseases, Department of Internal Medicine, University of Utah, Salt Lake City,
13 USA; ⁷Centre Pour le Développement des Vaccins-Mali, Bamako, Mali.; ⁸Division of
14 Infectious Disease and Tropical Pediatrics, University of Maryland, Baltimore, USA;
15 ⁹Division of Infectious Diseases and International Health, University of Virginia,
16 Charlottesville, USA; ¹⁰Departments of Pediatrics, University of Florida, Gainesville, USA;
17 ¹¹Departments of Environmental and Global Health, University of Florida, Gainesville,
18 USA; ¹²Department of Emergency Medicine, Brown University, Providence, USA;
19 ¹³Division of Microbiology and Immunology, Department of Internal Medicine, University
20 of Utah, Salt Lake City, USA

21
22 **Abstract** Traditional clinical prediction models focus on parameters of the individual patient. For
23 infectious diseases, sources external to the patient, including characteristics of prior patients and
24 seasonal factors, may improve predictive performance. We describe the development of a
25 predictive model that integrates multiple sources of data in a principled statistical framework using
26 a post-test odds formulation. Our method enables electronic real-time updating and flexibility,
27 such that components can be included or excluded according to data availability. We apply this
28 method to the prediction of etiology of pediatric diarrhea, where "pre-test" epidemiologic data may
29 be highly informative. Diarrhea has a high burden in low-resource settings, and antibiotics are
30 often over-prescribed. We demonstrate that our integrative method outperforms traditional
31 prediction in accurately identifying cases with a viral etiology, and show that its clinical application,
32 especially when used with an additional diagnostic test, could result in a 61% reduction in
33 inappropriately prescribed antibiotics.

35 Introduction

36 Healthcare providers use clinical decision support tools to assist with patient diagnosis, often to
37 improve accuracy of diagnosis, reduce cost by avoiding unnecessary laboratory tests, and in the case
38 of infectious diseases, deter the inappropriate prescription of antibiotics (*Sintchenko et al. (2008)*).
39 Typically, data entered into these tools is related directly to the patient's individual characteristics,
40 but data sources external to the patient can be informative for diagnosis. For example, climate,
41 seasonality, and epidemiological data inform predictive models for communicable disease incidence
42 (*Colwell (1996)*, *Chao et al. (2019)* *Fine et al. (2011)*). The emergence of advanced computing and

43 machine learning has enabled the incorporation of large data sources in the development of
44 clinical support tools (*Shortliffe and Sepúlveda (2018)*) such as SMART-COP for predicting the need
45 for intensive respiratory support for pneumonia (*Charles et al. (2008)*) or the ALaRMS model for
46 predicting inpatient mortality (*Tabak et al. (2014)*).

47 Clinical decision support tools rely on the availability of information sources and computing
48 at the time of patient encounter. Although increased availability of internet/mobile phones have
49 increased access to information and computing power in low-resource settings, there may be times
50 when connectivity, computing power, or data-collection infrastructure is unavailable. Thus, there is
51 a need to build clinical decision support tools which can flexibly include features of external sources
52 when available, or function without them if unavailable. Methods that enable the dynamic updating
53 of predictive models are advantageous due to potential cyclical patterns of infectious etiologies.
54 Furthermore, with the emergence of point-of-care (POC) tests for clinical decision-making (*Price*
55 *(2001)*), predictive models that are able to integrate results of such diagnostic testing could enhance
56 their usefulness.

57 We develop a novel method for diagnostic prediction which integrates multiple data sources by
58 utilizing a post-test odds formulation with proof-of-concept in antibiotic stewardship for pediatric
59 diarrhea. Our formulation first fits separate models from different sources of data, and then
60 combines the likelihood ratios from each of these independent models into a single prediction.
61 This method allows the multiple components to be flexibly included or excluded. We apply this
62 method to the prediction of diarrhea etiology with data from the Global Enteric Multicenter Study
63 (GEMS) (*Kotloff et al. (2013)*) and assess the performance of this tool, including with the addition of
64 a synthetic diagnostic, using two forms of internal-validation and by showing its potential effect on
65 reducing inappropriate antibiotic use.

66 **Methods**

67 We present our approach to building and assessing a flexible multi-source clinical prediction tool
68 with 1) the data sources, 2) the individual prediction models, 3) the use of the likelihood ratio for
69 integrating predictive models, 4) validation of the method, 5) the impact of an additional diagnostic,
70 and 6) a simulation of conditionally dependent tests.

71 **Data Sources**

72 We apply our post-test odds model using clinical data from GEMS, a prospective, case-control study
73 from 2007-2011 which took place in 7 countries in Africa and Asia. Methods for the GEMS study
74 have been described in detail (*Kotloff et al. (2012)*). Briefly, 9439 children with moderate-to-severe
75 diarrhea were enrolled at local health care centers along with 1 to 3 matched control-children. A fecal
76 sample was taken from each child at enrollment to identify enteropathogens clinical information
77 was collected, including demographic, anthropometric, and clinical history of the child. We used the
78 quantitative real-time PCR-based (qPCR) attribution models developed by *Liu et al. (2016)* in order
79 to best characterize the cause of diarrhea. Our dependent variable was presence or absence of
80 viral etiology, defined as a diarrhea episode with at least one viral pathogen with an episode-specific
81 attributable fraction ($AF_v \geq 0.5$) and no bacterial or parasitic pathogens with an episode-specific
82 attributable fraction. Prediction of viral attribution is clinically meaningful since it indicates that a
83 patient would not benefit from antimicrobial therapy. We defined other known etiologies as having
84 a majority attribution of diarrhea episode by at least one other non-viral pathogen. We exclude
85 patients with unknown etiologies when fitting the model, though it has been previously shown that
86 these cases have a similar distribution of viral predictions using a model with presenting patient
87 information as those cases with known etiologies (*Brintz et al. (2020)*).

88 We obtained weather data local to each site's health centers during the GEMS study using NOAA's
89 Integrated Surface Database (*Smith et al. (2011)*). The incidence of many pathogens, including
90 rotavirus (*Cook et al. (1990)*), norovirus (*Ahmed et al. (2013)*), cholera (*Emch et al. (2008)*), and
91 *Salmonella* (*Mohanty et al. (2006)*), are known to have seasonal patterns, and other analyses have

92 established climatic factors to be associated with diarrheal diseases (*Colwell (1996), Chao et al.*
93 *(2019), Farrar et al. (2019)*). Stations near GEMS sites such as in The Gambia exhibit seasonal
94 patterns (Figure 1). We used daily temperature and rain data weighted most by those weather
95 stations closest to the GEMS sites (Appendix 1).

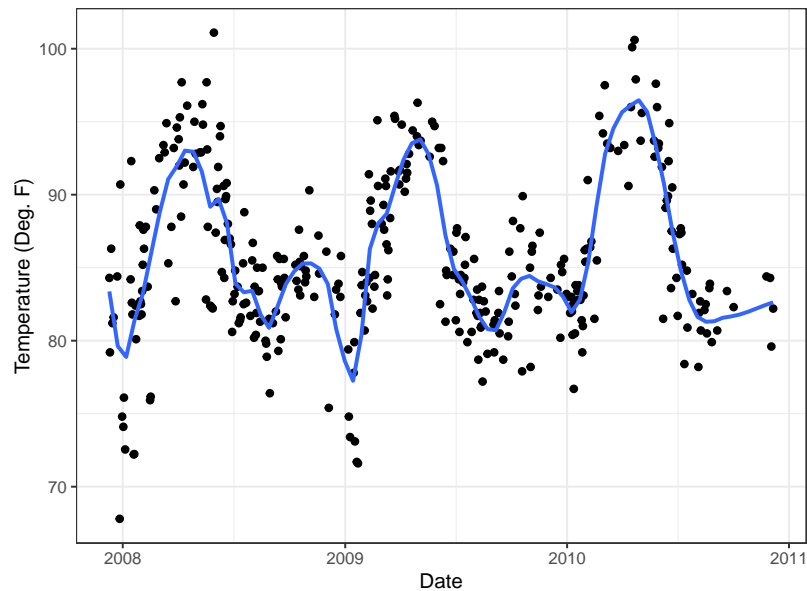


Figure 1. Temperature in The Gambia over study period with (blue) trend line from LOESS (locally estimated scatterplot smoothing)

96 **Construction of Predictive Models**

97 We define each model using the features described in the below sub-sections in an additive logistic
98 regression model. Each model can be trained using a sample of data from a specific country,
99 continent, or all available data.

100 Predictive model A) Presenting Patient

101 The patient model derived from the GEMS data treats each enrolled patient as an observation and
102 uses their available patient data at presentation to predict viral only versus other etiology of their
103 infectious diarrhea. In order to make a parsimonious model, we used the previously published
104 random forests variable importance screening (Brintz, et. al.). Using the screened variables (Table
105 1), we fit a logistic regression including the top five variables that would be accessible to providers at
106 the time of presentation. These include age, blood in stool, vomiting, breastfeeding status, and mid-
107 upper arm circumference (MUAC), an indicator of nutritional status. We note that while variables
108 such as fever and diarrhea duration were shown to be important in previous studies (*Fontana et al.*
109 *(1987)*), adding these variables did not improve performance (*Brintz et al. (2020)*). Additionally,
110 we excluded “Season”, since variables representing it are included in the climate predictive model
111 (discussed below), as well as “Height-for-age Z-score”, another indicator of nutritional status, which
112 would require a less feasible calculation than measurement of MUAC.

Variable Name	Viral Etiology Variance Reduction
Age	51.6
Season	29.0
Blood in stool	26.1
Height-for-age Z-score	24.7
Vomiting	23.0
Breastfeeding	22.0
Mid-upper arm circumference	20.9
Respiratory rate	18.5
Wealth index	18.3
Body Temperature	16.7

Table 1. Rank of Variable Importance by average reduction in the mean squared prediction error of the response using Random Forest regression. Greyed rows are variables that would be accessible for providers in LMICs at the time of presentation.

113 Predictive model B) Climate

114 We use an aggregate (mean) of the weighted (Appendix 1) local weather data over the prior 14 days
 115 to create features that capture site-specific climatic drivers of etiology of infectious diarrhea. By
 116 taking an aggregate, we create a moving average that reflects the seasonality seen in Figure 1. An
 117 example of the aggregate climate data from The Gambia is shown in Figure 1-figure supplement 1.
 118 From the figure, which also shows a moving average of the viral rate, We see that the periods of
 119 higher viral cases of diarrhea tend to have low temperatures and less rain.

120 Predictive model C) Seasonality

121 We include a predictive model with sine and cosine functions as features as explored in *Stolwijk*
 122 *et al. (1999)*. Assuming a periodicity of 365.25 days, we have functions $\sin(\frac{2\pi t}{365.25})$ and $\cos(\frac{2\pi t}{365.25})$. We
 123 show that standardized seasonal sine and cosine curves correlate with a rolling average of daily
 124 viral etiology rates in The Gambia over time (Figure 1-figure supplement 2). These functions can be
 125 used to represent multiple underlying processes that result in a seasonality of viral etiology.

126 **Use of the likelihood ratio to integrate predictive models from multiple data sources**

We integrate predictive models from the multiple sources of data described above using the post-
 test odds formulation. Using Bayes' Theorem, $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$, to construct the post-test odds of
 having a viral etiology,

$$\frac{P(V = 1|T_1 = t_1, T_2 = t_2, \dots, T_k = t_k)}{P(V = 0|T_1 = t_1, T_2 = t_2, \dots, T_k = t_k)} = \frac{P(V = 1, T_1 = t_1, T_2 = t_2, \dots, T_k = t_k)}{P(V = 0, T_1 = t_1, T_2 = t_2, \dots, T_k = t_k)} \quad (1)$$

$$= \frac{P(T_1 = t_1, T_2 = t_2, \dots, T_k = t_k|V = 1) \cdot P(V = 1)}{P(T_1 = t_1, T_2 = t_2, \dots, T_k = t_k|V = 0) \cdot P(V = 0)} \quad (2)$$

$$= \frac{P(V = 1)}{P(V = 0)} \cdot \prod_{j=1}^k \frac{P(T_j = t_j|V = 1)}{P(T_j = t_j|V = 0)} \quad (3)$$

127 where $V = 1$ represents a viral etiology and $V = 0$ represents an other known etiology, T_1, T_2, \dots, T_k
 128 represent the k tests, the distribution of the predictions from one or more predictive models, used
 129 to obtain the post-test odds, and $\frac{P(V=1)}{P(V=0)}$ is the pre-test odds. Note that going from line (2) to line
 130 (3) requires conditional independence between the tests, i.e., that $P(T_i = t_i, T_j = t_j|V = 1) = P(T_i =$
 131 $t_i|V = 1) \cdot P(T_j = t_j|V = 1)$ and $P(T_i = t_i, T_j = t_j|V = 0) = P(T_i = t_i|V = 0) \cdot P(T_j = t_j|V = 0)$ for all i
 132 and j . We test for conditional independence to assess the necessity of making higher-dimensional
 133 kernel density estimates using the *ci.test* function from the *{bnlearn}* package in R (*Scutari (2010)*).
 134 We derive each $P(T_j = t_j|V = 1)$ and $P(T_j = t_j|V = 0)$ using Gaussian kernel density estimates on
 135 conditional predictions from a logistic regression model fit on the training set (*Silverman (1986)*).
 136 The distribution of $P(T_j|V)$ is derived using the kernel density estimator $f(t_j) = \frac{1}{nh} \sum_{i=1}^n K(\frac{t_j - x_i}{h})$

137 where, in our case, $K(x) = \phi(x)$, the standard normal density function, and the bandwidth, h , is
 138 Silverman's 'rule of thumb' and the default chosen in the *density* function in R (Parzen (1962)).

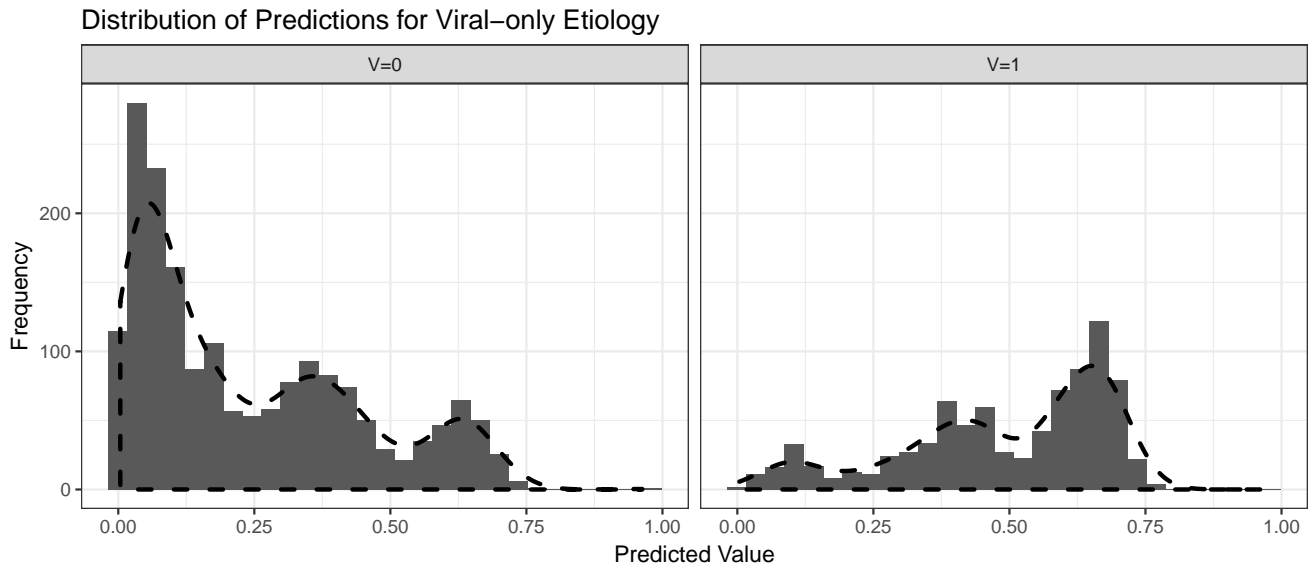


Figure 2. Histograms with overlaid estimated kernel densities (dashed lines) of predicted values obtained from logistic regression on patient training data. The left graph represent other known etiologies and the right graph represent viral etiologies. The dashed lines do not represent standardized density heights so the heights for V=0 and V=1 should not be compared from this graph.

139 Figure 2 shows an example of the frequency of predictions from a logistic regression model
 140 conditional on the viral-only status (V=0 and V=1) determined from attributable fractions. Addi-
 141 tionally, we overlaid the estimated 1-dimensional kernel density. To obtain the value of $\frac{P(T_j=t_j|V=1)}{P(T_j=t_j|V=0)}$,
 142 the predicted odds, from a model's prediction, we divide the kernel density estimate from the
 143 $V = 1$ set (right) by the kernel density estimate from the $V = 0$ set (left). It is feasible to estimate a
 144 multi-dimensional kernel density so that it is not necessary to make the conditional independence
 145 assumption to move from line 2 to line 3 in the equation above. Figure 2-figure supplement 1 shows
 146 an example 2-dimensional contour plot for kernel density estimates of predicted values obtained
 147 from logistic regression on GEMS seasonality and climate data in Mali which we will discuss further
 148 below. The density was created using R function *kde2d* (Venables and Ripley (2002)).

149 Pre-test Odds from Historical Data

150 We calculated pre-test odds using historical rates of viral diarrhea by site and date. We utilize
 151 available diarrhea etiology data for a given date, regardless of year, and site using a moving average
 152 such that pre-test probability π_d for date d is

$$\pi_d = \frac{D_{d-n} + D_{d-n+1} + \dots + D_d + \dots + D_{d+n-1} + D_{d+n}}{k_{d-n} + k_{d-n+1} + \dots + k_d + \dots + k_{d+n-1} + k_{d+n}}$$

$$D_d = \sum_{i=1}^{k_d} D_{di}$$

153 where k_d is the number of observed patients on date d , D_{di} is 1 if the etiology of the patients' diarrhea
 154 is viral and 0 otherwise, and n is the number of days included on both sides of the moving average.
 155 We would expect π_d to represent a pre-test probability of observing a viral diarrhea etiology on date
 156 d . Given that this rate information will likely be unavailable in new sites without established etiology
 157 studies, we provide an alternative formula based on recent patients' presentations (Appendix 2).

158 Validating the method

159 Given the temporal nature of some of the tests we developed, we estimate model performance
160 using within rolling-origin-recalibration evaluation. This method evaluates a model by sequentially
161 moving values from a test set to a training set and re-training the model on all of the training set
162 (*Bergmeir and Benítez (2012)*); for example, we train on the first 70% of the data and test on the
163 remaining 30%, then train on the first 80% of the data and test on the remaining 20%. No data
164 from the training set is used as part of the prediction for the test set. In each iteration of evaluation,
165 predictions on the test set are produced and corresponding measures of performance obtained:
166 the receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC), also known
167 as the C-statistic, along with AUC confidence intervals (*LeDell et al. (2015)*). Figure 3 depicts one
168 iteration of within rolling-origin-recalibration evaluation.

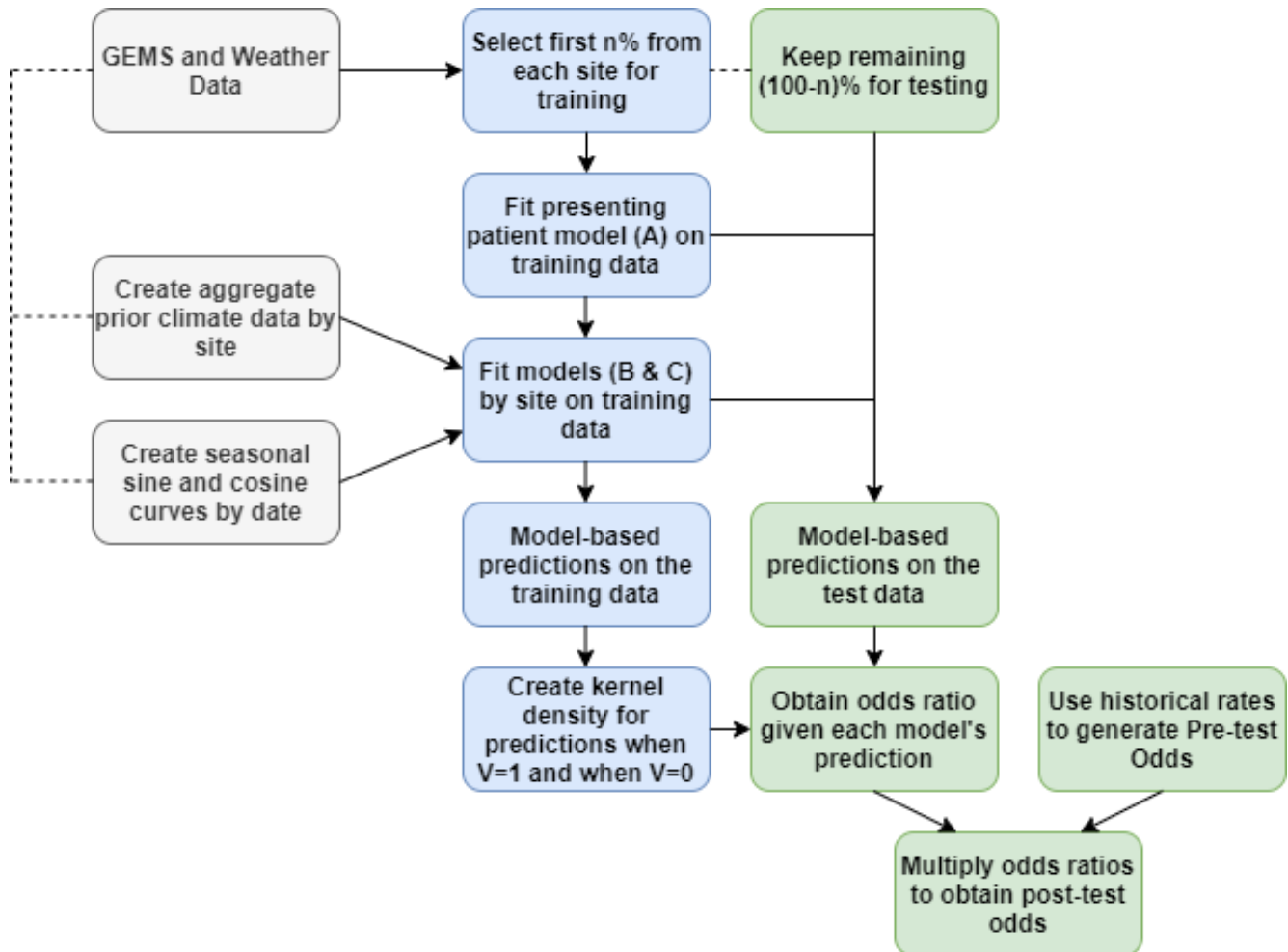


Figure 3. The steps for fitting prediction models and calculating the post-test odds for within rolling-origin-recalibration evaluation.

169 We additionally include a joint density for the climate and seasonal data in which we estimate a
170 2-dimensional kernel density (not shown in Figure 3). This model is called "Joint" in the results to
171 follow. To assess how this model might generalize to a site that was not used for model training, we
172 used a leave-one-site-out validation. By excluding a site and training the model's tests at a higher
173 level, such as on the entire continent, we get an idea of performance at a new site within one of the
174 continents for which we have data. Lastly, we define a threshold for the predicted odds ratio based
175 on the desired specificity of the model. We use this threshold to evaluate the effect of the model on
176 prescription or treatment of patients with antibiotics in the GEMS data.

177 **Modeling the impact of an additional diagnostic test**

178 We include a theoretical diagnostic which indicates viral versus other etiology with a given sensitivity
179 and specificity specifically to show the effect of an additional diagnostic-type test, such as a point-
180 of-care stool test, on the performance of our integrated post-test odds model. We include three
181 scenarios: 1) 70% sensitivity and 95% specificity, 2) 90% sensitivity and 95% specificity, and 3) 70%
182 sensitivity and 70% specificity. In order to estimate the performance of an additional diagnostic
183 test, for each patient in each of 500 bootstrapped samples of our test data, we randomly simulated
184 a test result based on the sensitivity or specificity of the diagnostic test. From the simulated test
185 result, we derive the likelihood ratio of the component directly from the specified sensitivity and
186 specificity of the test. A positive test results in a component likelihood ratio of $\frac{\text{sensitivity}}{1-\text{specificity}}$ and a
187 negative test results in a component likelihood ratio of $\frac{1-\text{sensitivity}}{\text{specificity}}$. We then take an average the
188 measure of performance of the bootstrapped samples.

189 **Simulation of Conditionally Dependent Tests**

We demonstrate the utility of the 2-dimensional kernel density estimate through simulation. In each
iteration of the simulation (100 iterations), we generate 3366 responses from a random Bernoulli
variable Z with a $\frac{1}{3}$ probability of success (the approximate proportion of GEMS cases with a viral
etiology). Then, conditioned on Z we generate predictive variables X and Y such that:

$$X = Z + \sigma \quad (4)$$

$$Y = \gamma X + Z + \sigma \quad (5)$$

190 where σ is a random draw from the standard normal distribution and values of γ ranging from -10
191 to 10 determine the level of conditional dependence between the two predictors conditional on
192 the value of Z . $\gamma = 0$ indicates conditional independence. Using an 80% training set, we derive the
193 kernel density estimate for the likelihood ratio (no pre-test odds included) using X and Y as two
194 separate tests and as a single 2-dimensional test and calculate the AUC from the 20% test set.

195 **Determination of Appropriate Antibiotic Prescription**

196 We demonstrate the clinical usefulness of our models by applying them directly to the prescription
197 of antibiotics. For each version of the model, we determined the threshold of prediction that would
198 amount to attaining a model specificity of 0.90 and 0.95. Since the prediction of a viral only etiology
199 of diarrhea indicates that antibiotics should not be prescribed, we chose these high specificities
200 due to the potential harm or even death that could occur if a patient who needed antibiotics did
201 not receive them. Using the thresholds, we determine which patients our models would correctly
202 predict a viral only etiology of their diarrhea (true positives) as well as patients our model would
203 incorrectly predict a viral only etiology of their diarrhea (false positives).

204 **Results**

205 **Integrative post-test odds models outperformed traditional models for prediction** 206 **of diarrhea etiology**

207 Of the 3366 patients in GEMS with an attributable identified pathogen, 1049 cases were attributable
208 to viral only etiology. We first examined whether our integrative post-test odds model can better
209 discriminate between patients with diarrhea of viral-only etiology and patients with other etiologies
210 than a traditional prediction model which includes only the presenting patient's information. We
211 found that overall, using the AUC as a discrimination metric, the integrative models outperformed
212 (AUC: 0.837 (0.806-0.869)) the traditional model (AUC: 0.809 (0.776-0.842)). Overall, the best per-
213 forming models were ones in which either the seasonal sine and cosine curves, or the prior patient
214 pre-test component alone was added to the presenting patient information with AUC's of 0.83 and
215 0.837 (with 80% training data), respectively (Figure 4). Including additional components, especially

216 including both climate and seasonality (though not as a joint density), appears to reduce the perfor-
217 mance. As expected, a reduced testing set increases the AUC but also increases the variance of the
218 estimate (Figure 4-figure supplement 1).

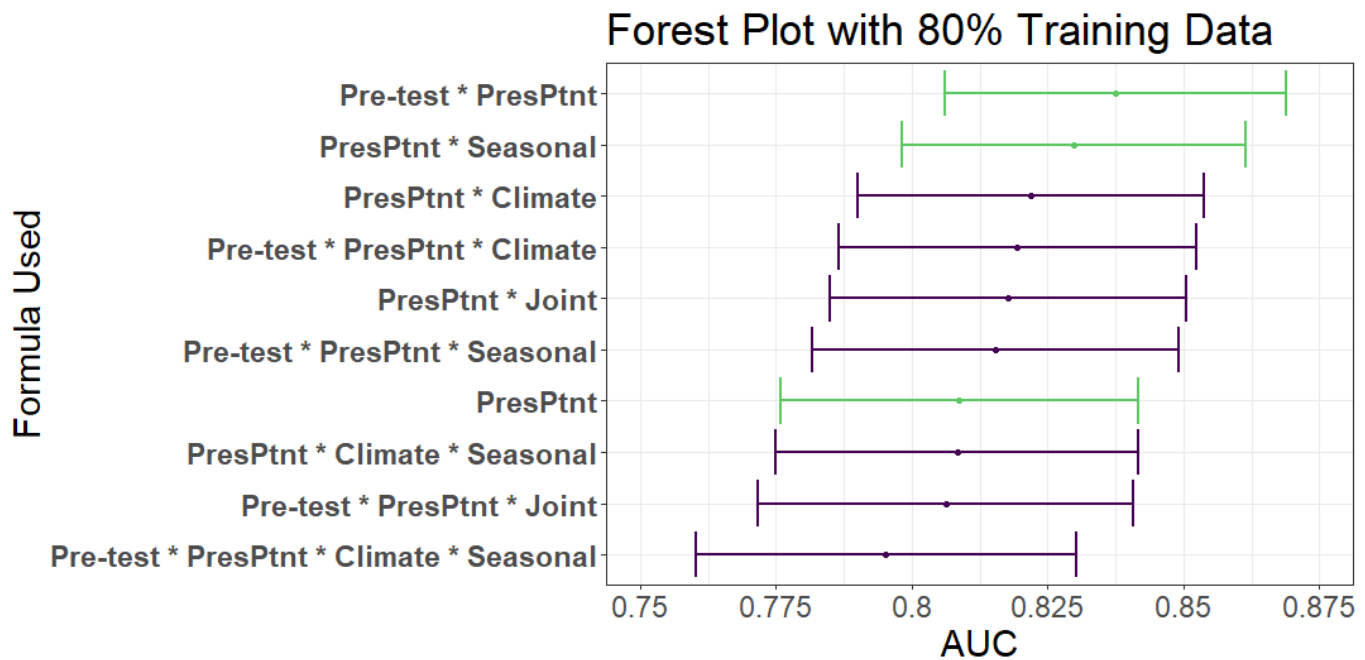


Figure 4. AUC's and confidence intervals for post-test odds used in the 80% training and 20% testing iteration. "PresPtnt" refers to the predictive model using the presenting patient's information. "Pre-test" refers to the use of pre-test odds based on prior patients' predictive models. "Climate" refers to the predictive model using aggregate local weather data. "Seasonal" refers to the predictive model based on seasonal sine and cosine curves. "Joint" refers to the 2-dimensional kernel density estimate from the Seasonal and Climate predictive models.

219 To assess our model's performance more granularly, we then examine performance of the top
220 two predictive models by individual sites. We found that the AUC, with 80% training and 20%
221 testing, varied greatly by site, ranging from 0.63 in Kenya to 0.95 in Bangladesh (Table 2). Of note,
222 the African sites have fewer patients in their testing and training sets than the Asian countries.
223 In leave-one-site-out validation testing, we found that the climate test tends to outperform the
224 seasonality test, and that there were notable differences in c-statistics between sites with the order
225 of performance similar to within rolling-origin-recalibration evaluation (Figure 4-figure supplement
226 2).

Country	Test Set Size	Formula	AUC (95% CI)
Kenya	79	Pre-test * PresPtnt	0.65 (0.53 - 0.77)
		PresPtnt * Seasonal	0.66 (0.54 - 0.78)
		PresPtnt	0.63 (0.51 - 0.75)
Mali	88	Pre-test * PresPtnt	0.74 (0.61 - 0.86)
		PresPtnt * Seasonal	0.78 (0.66 - 0.89)
		PresPtnt	0.75 (0.62 - 0.87)
Pakistan	108	Pre-test * PresPtnt	0.81 (0.72 - 0.89)
		PresPtnt * Seasonal	0.8 (0.72 - 0.88)
		PresPtnt	0.81 (0.73 - 0.89)
India	119	Pre-test * PresPtnt	0.84 (0.76 - 0.91)
		PresPtnt * Seasonal	0.85 (0.78 - 0.92)
		PresPtnt	0.81 (0.74 - 0.89)
The Gambia	80	Pre-test * PresPtnt	0.89 (0.82 - 0.96)
		PresPtnt * Seasonal	0.87 (0.79 - 0.94)
		PresPtnt	0.78 (0.67 - 0.88)
Mozambique	66	Pre-test * PresPtnt	0.88 (0.79 - 0.97)
		PresPtnt * Seasonal	0.9 (0.82 - 0.98)
		PresPtnt	0.77 (0.66 - 0.89)
Bangladesh	141	Pre-test * PresPtnt	0.91 (0.82 - 1)
		PresPtnt * Seasonal	0.93 (0.88 - 0.99)
		PresPtnt	0.95 (0.92 - 0.99)

Table 2. AUC results by site using 80% of data for training and 20% of data for testing of the top two models. PresPtnt refers to the model fit using presenting patient information.

227 Addition of a diagnostic test to integrative models improves discrimination

228 Emerging efforts to develop diagnostic devices, including laboratory assays as well POC tests, have
 229 focused on the performance of the test used in isolation. Here, we consider the use of a diagnostic
 230 device in combination with clinical predictive models. We used the integrative model to examine the
 231 impact that an additional diagnostic would have on discrimination of two of the best performing
 232 models. We show that an additional diagnostic, with varying sensitivity and specificity, would
 233 improve the cross-validated AUC as expected (Table 3). An additional test with a 70% sensitivity and
 234 70% specificity increases the AUC by 3-5%, while a more specific test could increase the AUC by 10%.

Model	Addl. Diag. (Se.,Sp.)	AUC (95% CI)	specificity=0.90		specificity=0.95	
			True +	False +	True +	False +
Pre-test * PresPtnt	None	0.837 (0.806 - 0.869)	90	30	58	15
	(0.7, 0.7)	0.874 (0.846 - 0.902)	101	31	76	15
	(0.7, 0.95)	0.933 (0.913 - 0.952)	132	31	122	15
	(0.9, 0.95)	0.972 (0.959 - 0.984)	154	33	147	17
PresPtnt * Seasonal	None	0.830 (0.798 - 0.861)	69	25	52	11
	(0.7, 0.7)	0.870 (0.842 - 0.897)	100	28	68	14
	(0.7, 0.95)	0.931 (0.912 - 0.951)	130	27	121	16
	(0.9, 0.95)	0.971 (0.959 - 0.984)	155	30	149	18
PresPtnt	None	0.809 (0.776 - 0.842)	66	31	41	15
	(0.7, 0.7)	0.857 (0.827 - 0.886)	97	34	68	16
	(0.7, 0.95)	0.925 (0.904 - 0.946)	129	33	117	18
	(0.9, 0.95)	0.968 (0.955 - 0.981)	154	33	149	18

Table 3. AUC and 95% confidence intervals from 80% training set after adding an additional point-of-care diagnostic test with specified sensitivities (Se.) and specificities (Sp.) to the current patient test and pre-test odds. Additionally, + and - refer to our model indicating a true positive or false positive, respectively, based on the threshold for each model which achieves a 0.90 or 0.95 specificity. Only patients who were prescribed/given antibiotics are included in the count.

235 We next examined ROC curves, which visually demonstrate the effect of additional diagnostics
 236 with varying levels of sensitivity and specificity (Figure 5). We show that a similar level of sensitivity
 237 and specificity is achievable by the model with the pre-test information versus the model with

238 seasonal information. Additionally, the additional diagnostics result in improved overall sensitivity
239 and specificity corresponding to sensitivity and specificity of the diagnostic. The overall sensitivity
240 and specificity of each model is greater than the diagnostic alone.

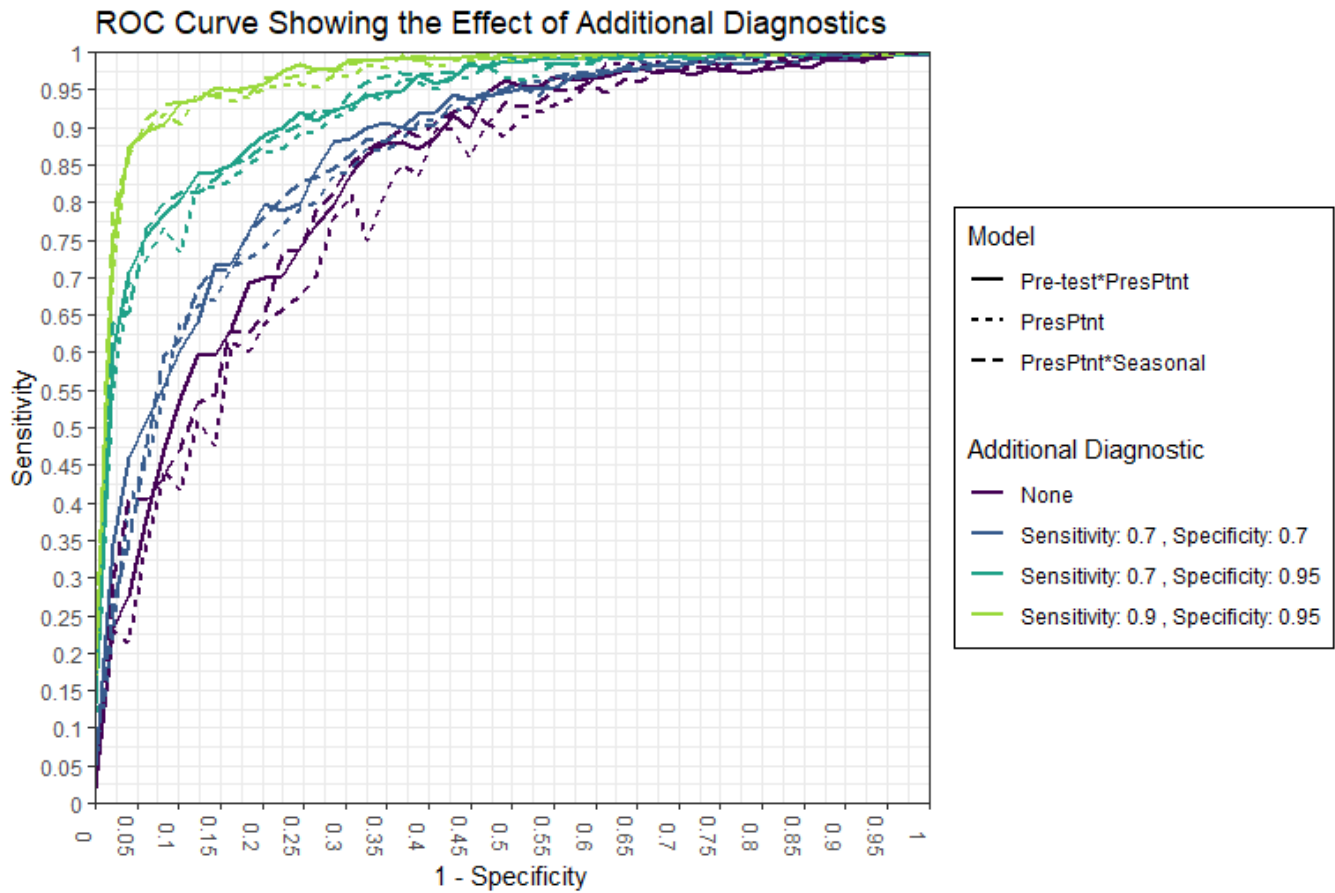


Figure 5. ROC curves from validation from 80% training set. Curves shown for three models with additional diagnostics.

241 **Breaking the conditional independence assumption can be addressed using 2-D**
242 **Kernel Density Estimates**

243 Our integrative post-test odds method assumes the conditional independence of its component
244 tests, and thus we performed simulation of increasingly conditionally dependent components to
245 assess the performance of the method when the assumption is broken. We showed that the AUC
246 of the post-test odds model deteriorates quickly as the conditional independence assumption is
247 violated (Table 4). With no conditional dependence between predictions from models X and Y, the
248 result using 1-dimensional kernel density is comparable to the result with 2-dimensional kernel
249 density model. However, as the conditional correlation between the tests increase to -0.90, the
250 1-dimensional AUC decreases by about 11% while the post-test odds with the 2-dimensional test
251 performs consistently across this range of conditional correlation.

γ	$cor(X, Y Z)$	AUC	
		1D-KDE	2D-KDE
-2.000	-0.894	0.725	0.830
-1.000	-0.709	0.758	0.828
-0.500	-0.446	0.824	0.838
0.000	0.002	0.838	0.836
0.500	0.448	0.836	0.836
1.000	0.708	0.831	0.840
2.000	0.894	0.810	0.836

Table 4. Average AUC's from 1-dimensional and 2-dimensional kernel density estimates (KDE) when the post-test odds conditional independence assumption is broken. The table shows the factor (γ) used to simulate induced conditional dependence between two covariates and their average conditional correlation. Additionally, it shows the average AUC resulting from a post-test odds model where a 1-dimensional kernel density estimate (conditional independence assumed) is generated for each covariate, and a post-test odds model where a 2-dimension joint kernel density estimate is derived for the two covariates.

252 **Clinician use of an integrative predictive model for diarrhea etiology could result**
253 **in large reductions in inappropriate antibiotic prescriptions**

254 Given that one potential application of an integrative predictive model for diarrhea etiology would
255 be as support for clinical decision making for antibiotic use (i.e. antibiotic stewardship), we then
256 examined the impact that the top predictive model would have on prescription of antibiotics by
257 clinicians in GEMS. Of the 3366 patients included in our study, 2653 (79%) were treated with or
258 prescribed antibiotics, 806 (30%) of whom were prescribed to those with a viral-only etiology as
259 determined by qPCR. Here, we examined how use of integrative predictive model could have altered
260 antibiotic use in our sample. Of the 681 patients in the 20% test set, 540 (79%) were prescribed
261 antibiotics, including 166 (30%) with a viral-only etiology. Of those prescribed/given antibiotics
262 the model with pre-test odds, with threshold chosen for an overall specificity of 0.90, identified
263 90 (54%) viral cases as viral, and 30 non-viral cases as viral. With an additional diagnostic with a
264 sensitivity and specificity of 0.70, the same model would on average identify 101 (61%) viral cases
265 as viral with the same 31 non-viral cases identified as viral. Assuming that clinicians would not
266 prescribe antibiotics for those cases identified by the predictive model with the additional diagnostic
267 as viral, we would avoid 90 (54%) and 101 (61%) of inappropriate antibiotic prescriptions in the
268 two scenarios described. The majority of the false positives (30 in both scenarios) were episodes
269 majority attributed to Shigella, ST-EPEC, and combinations of rotavirus with a non-viral pathogen
270 (Table 3-table supplement 1). All of these false positive, with exception of 1 case, had non-bloody
271 diarrhea, and thus would have been deemed as not requiring antibiotics by WHO IMCI guidelines.

272 **Discussion**

273 The management of illness in much of the world relies on clinical decisions made in the absence
274 of laboratory diagnostics. Such empirical decision-making, including decisions to use antibiotics,
275 are informed by variable degrees of clinical and demographic data gathered by the clinician.
276 Traditional clinical prediction rules focus on the clinical data from the presenting patient alone.
277 In this analysis, we present a method that allows flexible integration of multiple data sources,
278 including climate data and clinical or historical information from prior patients, resulting in improved
279 predictive performance over traditional predictive models utilizing a single source of data. Using
280 this formulation, if certain sources of data such as climate or previous patient information are not
281 available (e.g., due to a lack of internet connection or data infrastructure), the prediction can still be
282 made using the other sources. We show that application of such a predictive model, especially with
283 an additional diagnostic test, may translate to reductions in inappropriate antibiotic prescriptions
284 for pediatric viral diarrhea.

285 The global burden of acute infectious diarrhea is highest in low- and middle-income countries
286 (LMICs) in southeast Asia and Africa (*Walker et al. (2013)*), where there is limited access to diagnostic
287 testing. The care of children in these regions could greatly benefit from an accurate and flexible

288 decision making tool. Decisions for treatment are often empiric and antibiotics are over-prescribed
289 (*Rogawski et al. (2017)*), though the majority of cases of diarrhea do not benefit from antibiotic
290 use and also many instances of acute watery diarrhea are self-limiting . For example, 2653 (79%)
291 of the 3366 patients in our study were treated with or prescribed antibiotics. Of these 806 (30%)
292 were prescribed to those with a viral-only etiology. Unnecessary antibiotic use exposes children to
293 significant adverse events including serious allergic reactions and clostridium difficile infection, and
294 contributes to increased antimicrobial resistance. We show that a predictive model can be used to
295 discriminate between those with and without a viral-only etiology and that the inappropriate use of
296 antibiotics can be avoided in 54% cases using our model with no additional diagnostics.

297 We found using within rolling-origin-recalibration evaluation that models which include either
298 the pre-test odds calculated historical rates or the seasonal test were the best at discriminating
299 between viral etiologies and other etiologies, a finding that held true across training and testing
300 set sizes. However, in the leave-one-out validation, models which included the alternate pre-test
301 odds and climate tended to perform the best. This difference is likely due to the generalizeability of
302 the individual tests, i.e, the leave-one-out tests are trained at the continental level and the effect of
303 climate on etiology is intuitively more generalizeable than seasonal curves which are very specific
304 to each location. We found that our integrative model with only the historical (pre-test) information
305 included (without additional diagnostics) would have identified a viral-only etiology in 90 (54%)
306 patients who received antibiotics. We then show that even the use of an additional diagnostic test
307 with modest performance (70% sensitivity and specificity) would further decrease inappropriate
308 antibiotic use by another 11 (for a total of 101, or 61% of) patients. In the context of calls by the
309 WHO for the development of affordable rapid diagnostic tools (RDTs) for antibiotic stewardship
310 (*Declaration (2017)*), our findings suggest that development and evaluation of novel RDTs should not
311 be performed in isolation. Potential for integration of rapid diagnostic tests into clinical prediction
312 algorithms should be considered, though this needs to be balanced with the additional time and
313 resources needed. The incremental improvement in discriminative performance achieved by the
314 addition of an RDT to a clinical prediction algorithm may not be cost-effective in lower resourced
315 settings. Finally, providing this model in the form of a decision support tool to the clinician could
316 translate to reductions in inappropriate use of antibiotics.

317 The novel use of kernel density estimates to derive the conditional tests when calculating the
318 post-test odds enabled a flexibility in model input. While kernel density estimates have been used
319 for conditional feature distributions in Naïve Bayes classifiers (*John and Langley (1995), Murakami
320 and Mizuguchi (2010)*), here we show that they can be used to derive conditional likelihoods for
321 diagnostic tests constituting one or more features, stressing the effect of the overall test on
322 the post-test odds and not individual features. As such, complicated machine learning models
323 can be combined with simple diagnostics as part of the post-test odds. For example, we could
324 have fit neural networks in lieu of logistic regression models, and in addition to these more
325 complicated models, it is possible to incorporate the result of an RDT that make results available to
326 the clinician at the point-of-care. Additionally, our method of using two-dimensional kernel density
327 estimates can also be used to overcome the conditional independence assumption for tests based
328 on potentially interrelated diagnostic information. Densities with higher than two dimensions can
329 be considered, though, computational limitations are likely in both speed and, we expect, accuracy,
330 as the dimensions increase.

331 Our study has a number of limitations. First, a robust training set of both cases and non-
332 cases is required to adequately build the conditional kernel densities. Second, the post-test odds
333 calculation, at the time of prediction, lacks interpretation on a feature level like a logistic regression
334 or decision tree. Although, we do observe the effect of a test on an observation, we cannot see which
335 features caused that effect without diving deeper into the training of the diagnostic tests. Thirdly, the
336 prediction algorithm generated by the post-test odds model using GEMS data was only validated
337 internally, and further studies are need for external validation and field implementation.

338 In conclusion, we have developed a clinical prediction model that integrates multiple sources

339 external to the presenting patient, through use of a post-test odds framework and showed that it
340 improved diagnostic performance. When applied to the etiological diagnosis of pediatric diarrhea,
341 we demonstrate its potential for reducing inappropriate antibiotic use. The flexible inclusion or
342 exclusion of output from its components makes it ideal for decision support in lower-resourced
343 settings, when only certain data may be available due to limitations in information computation
344 or connectivity. Additionally, the ability to incorporate new training data in real-time to update
345 decisions allows the model to improve as more data is collected. Such a predictive model has the
346 potential to improve the management of pediatric diarrhea, including the rational use of antibiotics
347 in lower-resourced settings.

348 Acknowledgments

349 This investigation was supported by the University of Utah Study Design and Biostatistics Center,
350 with funding in part from the National Center for Research Resources and the National Center for
351 Advancing Translational Sciences, National Institutes of Health, through Grant 8UL1TR000105 (to
352 BJB, BH, and TG). Research reported in this publication was supported by the NIAID of the NIH
353 under award number R01AI135114 (to DTL), and the Bill and Melinda Gates Foundation award
354 OPP1198876 (to DTL). The authors would like to thank Bill and Melinda Gates for their active support
355 (JLP and DC) of the Institute for Disease Modeling and their sponsorship through the Global Good
356 Fund.

357 References

- 358 **Ahmed SM**, Lopman BA, Levy K. A systematic review and meta-analysis of the global seasonality of norovirus.
359 *PLoS one*. 2013; 8(10):e75922.
- 360 **Bergmeir C**, Benítez JM. On the use of cross-validation for time series predictor evaluation. *Information Sciences*.
361 2012; 191:192–213.
- 362 **Brintz BJ**, Howard J, Haaland B, Platts-Mills JA, Greene T, Levine A, Nelson E, Pavia A, Kotloff K, Leung DT. Clinical
363 predictors for etiology of acute diarrhea in children in resource-limited settings. *medRxiv*. 2020; .
- 364 **Chao DL**, Roose A, Roh M, Kotloff KL, Proctor JL. The seasonality of diarrheal pathogens: A retrospective study
365 of seven sites over three years. *PLoS neglected tropical diseases*. 2019; 13(8):e0007211.
- 366 **Charles PG**, Wolfe R, Whitby M, Fine MJ, Fuller AJ, Stirling R, Wright AA, Ramirez JA, Christiansen KJ, Waterer
367 GW, et al. SMART-COP: a tool for predicting the need for intensive respiratory or vasopressor support in
368 community-acquired pneumonia. *Clinical Infectious Diseases*. 2008; 47(3):375–384.
- 369 **Colwell RR**. Global climate and infectious disease: the cholera paradigm. *Science*. 1996; 274(5295):2025–2031.
- 370 **Cook S**, Glass R, LeBaron C, Ho MS. Global seasonality of rotavirus infections. *Bulletin of the World Health
371 Organization*. 1990; 68(2):171.
- 372 **Declaration D**. CIDRAP Antimicrobial Stewardship Project POLICY UPDATE, October 2017. *Policy*. 2017; .
- 373 **Emch M**, Feldacker C, Islam MS, Ali M. Seasonality of cholera from 1974 to 2005: a review of global patterns.
374 *International journal of health geographics*. 2008; 7(1):31.
- 375 **Farrar DS**, Awasthi S, Fadel SA, Kumar R, Sinha A, Fu SH, Wahl B, Morris SK, Jha P. Seasonal variation and
376 etiologic inferences of childhood pneumonia and diarrhea mortality in India. *eLife*. 2019; 8.
- 377 **Fine AM**, Brownstein JS, Nigrovic LE, Kimia AA, Olson KL, Thompson AD, Mandl KD. Integrating spatial epidemi-
378 ology into a decision model for evaluation of facial palsy in children. *Archives of pediatrics & adolescent
379 medicine*. 2011; 165(1):61–67.
- 380 **Fontana M**, Zuin G, Paccagnini S, Ceriani R, Quaranta S, Villa M, Principi N. Simple clinical score and laboratory-
381 based method to predict bacterial etiology of acute diarrhea in childhood. *The Pediatric infectious disease
382 journal*. 1987; 6(12):1088–1091.
- 383 **John GH**, Langley P. Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the Eleventh
384 conference on Uncertainty in artificial intelligence* Morgan Kaufmann Publishers Inc.; 1995. p. 338–345.

- 385 **Kotloff KL**, Blackwelder WC, Nasrin D, Nataro JP, Farag TH, van Eijk A, Adegbola RA, Alonso PL, Breiman RF,
386 Golam Faruque AS, et al. The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and
387 young children in developing countries: epidemiologic and clinical methods of the case/control study. *Clinical*
388 *infectious diseases*. 2012; 55(suppl_4):S232–S245.
- 389 **Kotloff KL**, Nataro JP, Blackwelder WC. Burden and aetiology of diarrhoeal disease in infants and young children
390 in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The*
391 *Lancet*. 2013; 382(9888):209–222.
- 392 **LeDell E**, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area
393 under the ROC curve estimates. *Electronic journal of statistics*. 2015; 9(1):1583.
- 394 **Liu J**, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C, Operario DJ, Uddin J, Ahmed S, Alonso PL, et al. Use of
395 quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the
396 GEMS case-control study. *The Lancet*. 2016; 388(10051):1291–1301.
- 397 **Mohanty S**, Renuka K, Sood S, Das B, Kapil A. Antibioqram pattern and seasonality of Salmonella serotypes in a
398 North Indian tertiary care hospital. *Epidemiology & Infection*. 2006; 134(5):961–966.
- 399 **Murakami Y**, Mizuguchi K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction
400 of protein–protein interaction sites. *Bioinformatics*. 2010; 26(15):1841–1848.
- 401 **Parzen E**. On estimation of a probability density function and mode. *The annals of mathematical statistics*.
402 1962; 33(3):1065–1076.
- 403 **Price CP**. Point of care testing. *Bmj*. 2001; 322(7297):1285–1288.
- 404 **Rogawski ET**, Platts-Mills JA, Seidman JC, John S, Mahfuz M, Ulak M, Shrestha SK, Soofi SB, Yori PP, Mduma E,
405 et al. Use of antibiotics in children younger than two years in eight countries: a prospective cohort study.
406 *Bulletin of the World Health Organization*. 2017; 95(1):49.
- 407 **Scutari M**. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*. 2010;
408 35(3):1–22. doi: [10.18637/jss.v035.i03](https://doi.org/10.18637/jss.v035.i03).
- 409 **Shortliffe EH**, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *Jama*. 2018;
410 320(21):2199–2200.
- 411 **Silverman BW**. Density estimation for statistics and data analysis, vol. 26. CRC press; 1986.
- 412 **Sintchenko V**, Coiera E, Gilbert GL. Decision support systems for antibiotic prescribing. *Current opinion in*
413 *infectious diseases*. 2008; 21(6):573–579.
- 414 **Smith A**, Lott N, Vose R. The integrated surface database: Recent developments and partnerships. *Bulletin of*
415 *the American Meteorological Society*. 2011; 92(6):704–708.
- 416 **Stolwijk A**, Straatman H, Zielhuis G. Studying seasonality by using sine and cosine functions in regression
417 analysis. *Journal of Epidemiology & Community Health*. 1999; 53(4):235–238.
- 418 **Tabak YP**, Sun X, Nunez CM, Johannes RS. Using electronic health record data to develop inpatient mortality
419 predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). *Journal of the American Medical*
420 *Informatics Association*. 2014; 21(3):455–463.
- 421 **Venables WN**, Ripley BD. *Modern Applied Statistics with S*. Fourth ed. New York: Springer; 2002. <http://www.stats.ox.ac.uk/pub/MASS4>, ISBN 0-387-95457-0.
- 422
- 423 **Walker CLF**, Rudan I, Liu L, Nair H, Theodoratou E, Bhutta ZA, O'Brien KL, Campbell H, Black RE. Global burden
424 of childhood pneumonia and diarrhoea. *The Lancet*. 2013; 381(9875):1405–1416.

425 **Supplemental Figures and Table**

426 **Figure 1 Supplements**

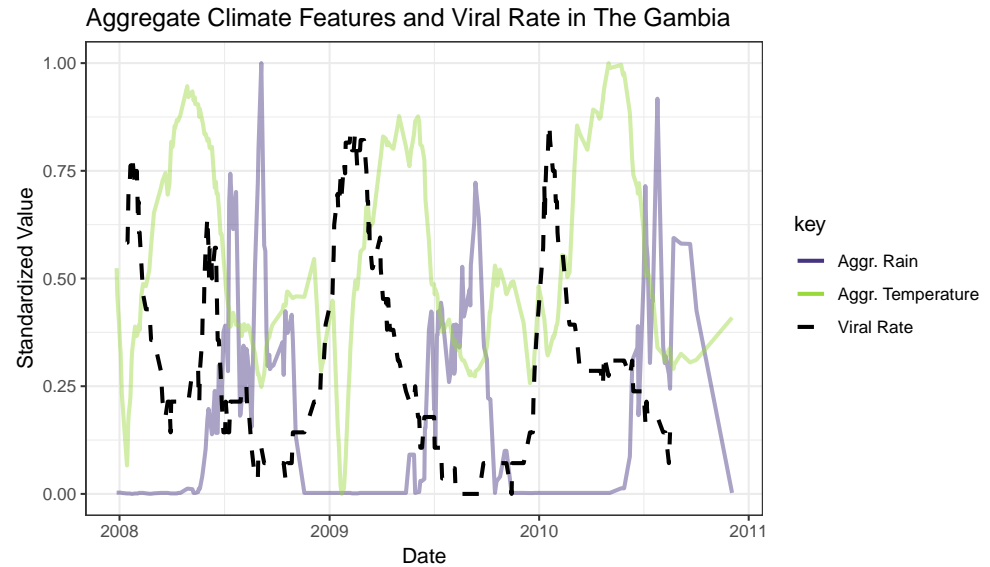


Figure 1. The black line represents a 2-week rolling average of daily viral etiology rates over time. The purple and green lines represent the prior two week average of daily rain and temperature averages.

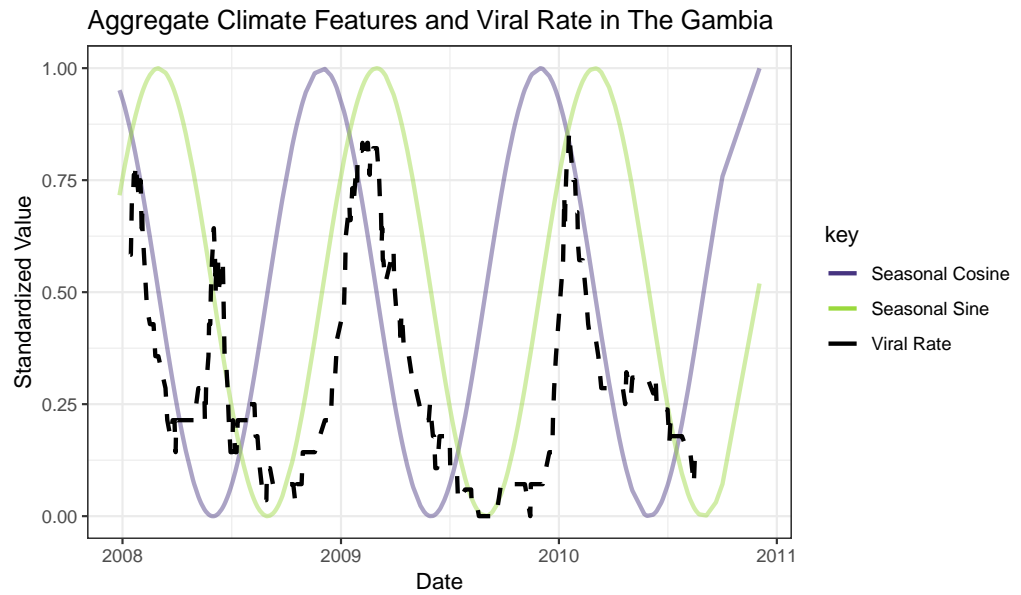


Figure 2. The black line represents a 2-week rolling average of daily viral etiology rates over time. The purple and green lines represent the prior two week average of daily rain and temperature averages.

427 **Figure 2 Supplements**

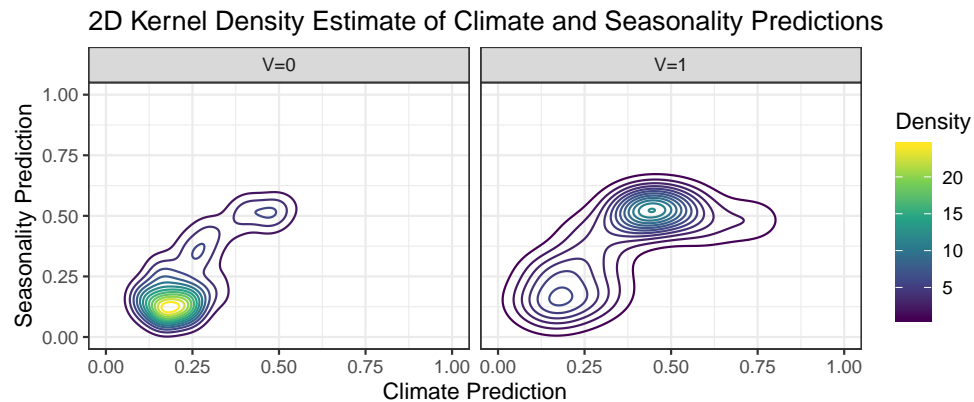


Figure 1. Contour plots of 2-dimensional kernel densities of predicted values obtained from logistic regression on GEMS climate and seasonality data in Mali. The right graph represents viral etiologies and the left graph represents other known etiologies.

428 **Figure 4 Supplements**

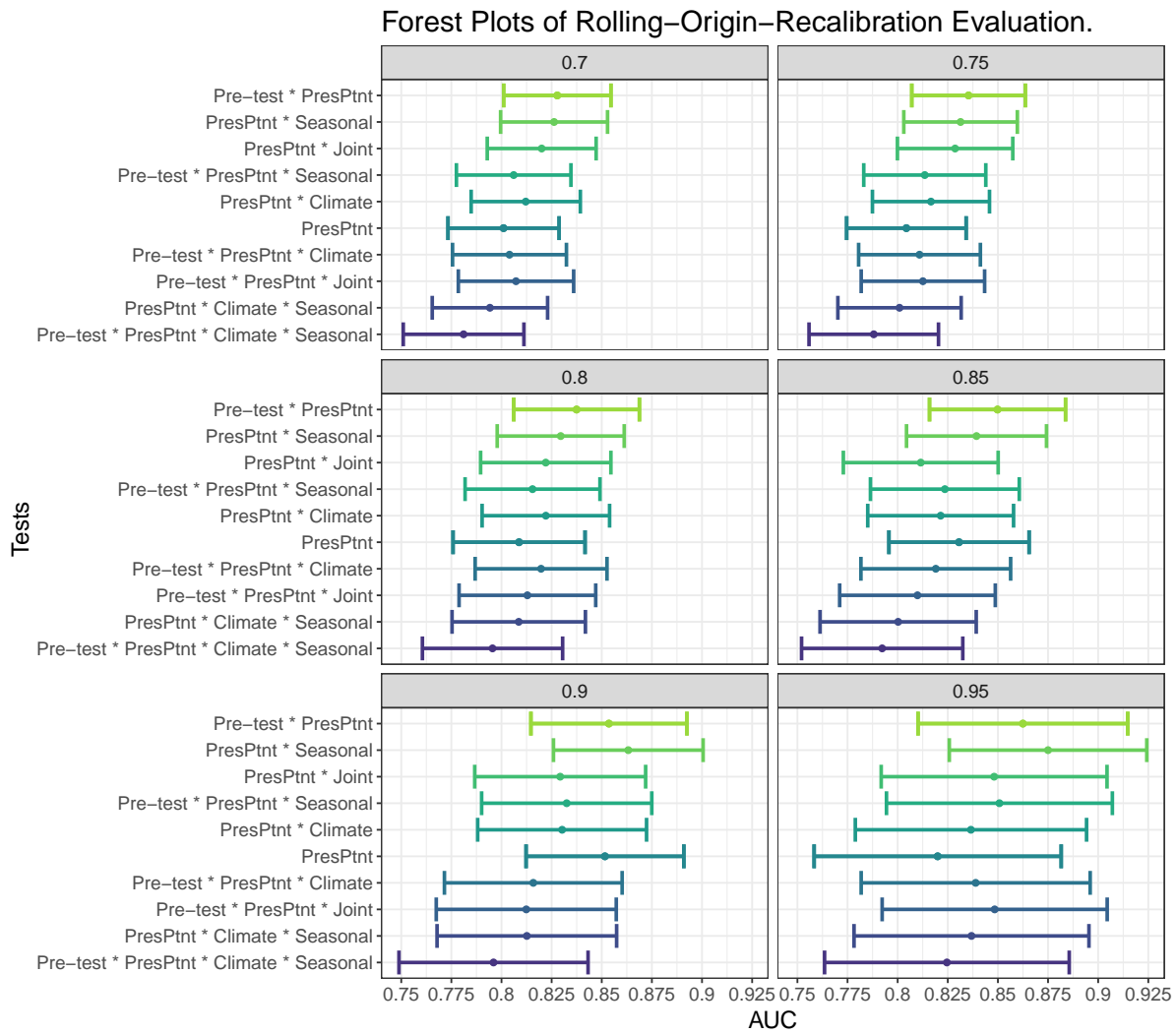


Figure 1. AUC's and confidence intervals for tests used in within rolling-origin-recalibration evaluation. Individual plot titles show the proportion of data used in training.

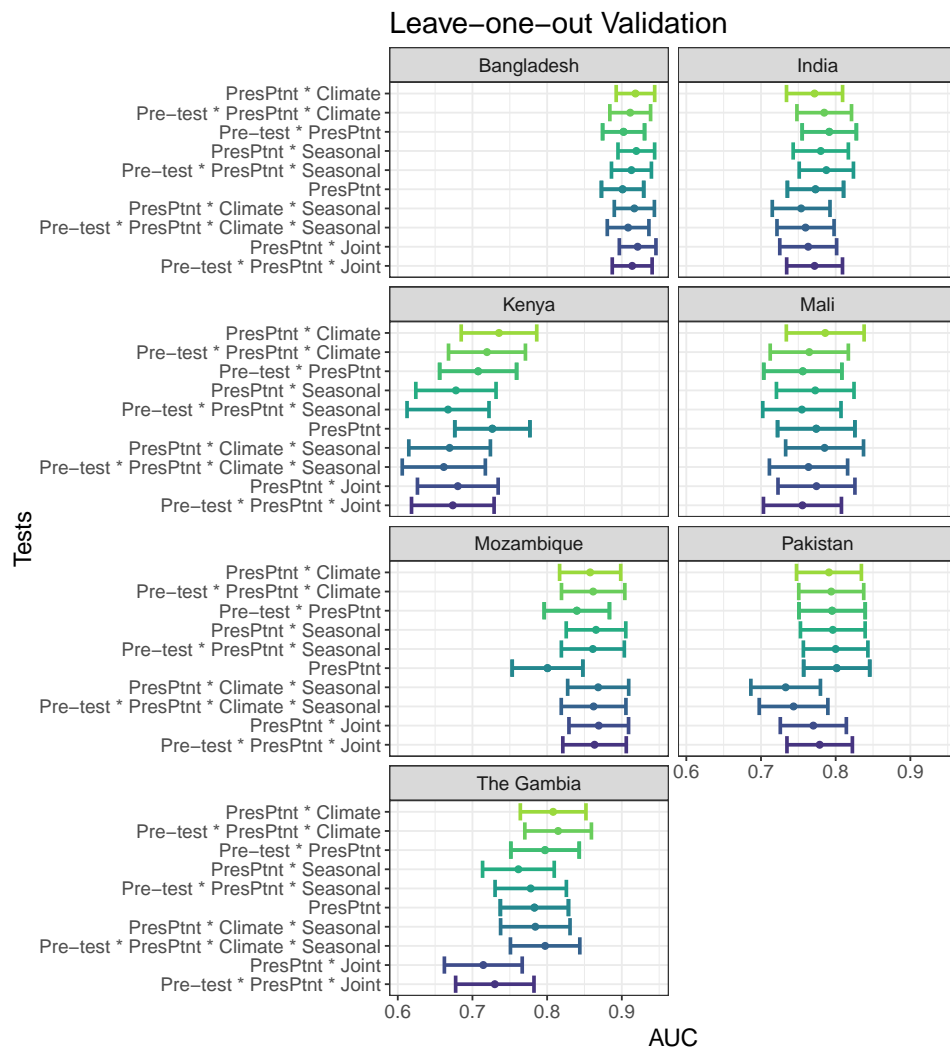


Figure 2. AUC's and confidence intervals for tests used in the leave-one-site-out evaluation. Pre-test refers to the use of prior patient predictions. Individual plot titles show the site left out of training.

429 **Table 3 Supplements**

Pathogen(s)	Model					
	Pre-test * PresPtnt		PresPtnt * Seasonal		PresPtnt	
	Sp.=0.90	Sp.=0.95	Sp.=0.90	Sp.=0.95	Sp.=0.90	Sp.=0.95
ST-ETEC	5	3	4	2	5	2
Shigella/EIEC	5	1	5	0	4	1
Cryptosporidium	4	2	3	2	9	5
Cryptosporidium+Rotavirus	3	1	2	1	3	3
H. pylori+Rotavirus	3	2	3	2	1	0
Rotavirus+TEPEC	2	1	1	1	1	0
C. jejuni/C. coli+Rotavirus	1	1	1	1	1	1
TEPEC	1	1	1	1	1	1
Adenovirus 40/41+Shigella/EIEC	1	0	0	0	1	0
Rotavirus+ST-ETEC	1	1	2	0	0	0
Rotavirus+Shigella/EIEC	1	1	1	0	0	0
salmonella	1	0	0	0	0	0
Astrovirus+TEPEC	1	0	0	0	0	0
Norvirus GII+Shigella/EIEC	1	1	1	1	0	0
Astrovirus+Shigella/EIEC	0	0	1	0	0	0
C. jejuni/C. coli+Crypto.	0	0	0	0	1	0
Cryptosporidium+ST-ETEC	0	0	0	0	1	0
Adenovirus 40/41+ST-ETEC	0	0	0	0	1	1
Adenovirus 40/41+Crypto.	0	0	0	0	1	1
H. pylori+Shigella+V. cholerae	0	0	0	0	1	0
	30	15	25	11	31	15

Table 1. Frequency table of pathogens in which the post-test odds formulation with varying specificity (Sp.) chosen have false positives.

430 Appendix 1

431 **Weighted Weather Station Data**

432 Daily local weather information was constructed based on data from weather stations within
433 200km of the site of interest. We chose 200km because one our sites, Mozambique, does not
434 have any stations nearer than 180 km. We then collect the temperature and rain info from
435 the top 5 closest weather stations and take a weighted average where they are weighted
436 inversely by distance so that the closer weather stations will have more effect on the average.
437 For instance, for temperature on day d across the 5 closest weather stations: $T_d = \frac{\sum_{i=1}^5 T_{di} \cdot d_i^{-1}}{\sum_{i=1}^5 d_i^{-1}}$
438 where T_{di} is the average temperature for weather station i on day d and d_i is the distance
439 from weather station i .

440 Appendix 2

441 Pre-test Odds from Prior Patient Predictions for Prediction in New Sites

442 We calculated pre-test odds by combining past predictions from predictive model A, the
443 presenting patient model. By taking a weighted average of the recently predicted odds
444 of viral etiology, we attempt to capture recent local trends in diarrhea pathogens, such as
445 localized outbreaks. This is similar to heuristic decision making historically used by clinicians.
446 We aggregated the odds calculated from the presenting patient model on their probability
447 scale for each site over the past d days such that pre-test probability π_d for day d is

$$447 \pi_d = \frac{P_{d-n+1} \cdot w_1 + P_{d-n+2} \cdot w_2 + \dots + P_d \cdot w_n}{448 w_1 + w_2 + \dots + w_n}$$

$$449 P_d = \frac{1}{k} \sum_{i=1}^k P_{di}$$

450 where P_{di} are the $i = 1, \dots, k$ current patient predictions converted from the odds scale to
451 the probability scale on day d and n is the number of prior days included in the calculation.
452 Provided the greatest weights are put on the most recent predictions, we would expect an
453 influx of certain symptoms related to a viral etiology to be represented by π_d .
454