

1 **Integration of population-level data sources into an individual-level clinical prediction model for dengue**  
2 **virus test positivity**

3  
4 **Short title: Population-level data for clinical prediction**

5  
6 RJ Williams<sup>1</sup>, Ben J. Brintz<sup>1,2</sup>, Gabriel Ribeiro Dos Santos<sup>3</sup>, Angkana Huang<sup>3,4</sup>, Darunee Buddhari<sup>4</sup>, Surachai  
7 Kaewhiran<sup>5</sup>, Sapon Iamsirithaworn<sup>5</sup>, Alan L. Rothman<sup>6</sup>, Stephen Thomas<sup>7</sup>, Aaron Farmer<sup>4</sup>, Stefan Fernandez<sup>4</sup>,  
8 Derek A T Cummings<sup>8,9</sup>, Kathryn B Anderson<sup>4,7</sup>, Henrik Salje<sup>\*3</sup>, Daniel T. Leung<sup>\*1,10</sup>

- 9  
10 1. Division of Infectious Diseases, Department of Internal Medicine, University of Utah, Salt Lake City,  
11 USA.  
12 2. Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, USA.  
13 3. Department of Genetics, University of Cambridge, United Kingdom.  
14 4. Department of Virology, Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand.  
15 5. Ministry of Public Health, Nonthaburi, Thailand  
16 6. Institute for Immunology and Informatics and Department of Cell and Molecular Biology, University of  
17 Rhode Island, Providence, USA.  
18 7. Department of Microbiology and Immunology, SUNY Upstate Medical University, Syracuse, USA.  
19 8. Department of Biology, University of Florida, Gainesville, USA.  
20 9. Emerging Pathogens Institute, University of Florida, Gainesville, USA.  
21 10. Division of Microbiology and Immunology, Department of Pathology, University of Utah, Salt Lake City,  
22 USA

23  
24 \*co-corresponding authors:

25 Daniel T. Leung, MD  
26 University of Utah, USA  
27 [Daniel.Leung@utah.edu](mailto:Daniel.Leung@utah.edu)

28  
29 Henrik Salje, PhD  
30 University of Cambridge, UK  
31 [hs743@cam.ac.uk](mailto:hs743@cam.ac.uk)

32  
33 Keywords: Clinical Prediction, Dengue Virus, Acute Febrile Illness  
34

35 **Abstract**

36 The differentiation of dengue virus (DENV) infection, a major cause of acute febrile illness in tropical regions,  
37 from other etiologies, may help prioritize laboratory testing and limit the inappropriate use of antibiotics.  
38 While traditional clinical prediction models focus on individual patient-level parameters, we hypothesize that  
39 for infectious diseases, population-level data sources may improve predictive ability. To create a clinical  
40 prediction model that integrates patient-extrinsic data for identifying DENV among febrile patients presenting  
41 to a hospital in Thailand, we fit random forest classifiers combining clinical data with climate and population-  
42 level epidemiologic data. In cross validation, compared to a parsimonious model with the top clinical  
43 predictors, a model with the addition of climate data, reconstructed susceptibility estimates, force of infection  
44 estimates, and a recent case clustering metric, significantly improved model performance.  
45

## 46 Introduction

47 Acute febrile illness (AFI) is a common reason for seeking healthcare in low- and middle-income countries  
48 (LMICs) (1). Determination of AFI etiology is often limited by diagnostic testing capacity, given the wide  
49 spectrum of potential infectious agents. Inappropriate use of testing and treatment resources may result in  
50 poor outcomes, such as the high case fatality rates seen in admitted AFI patients (5-20%) (2-7). Dengue virus  
51 (DENV) is a major cause of AFI in LMICs, accounting for an estimated 390 million infections, 96 million  
52 illnesses, 2 million severe cases, and 21,000 deaths per year (8). The differentiation between dengue and  
53 other common causes of febrile illness is important to avoid misdiagnosis, which can lead to delays in initiation  
54 of effective treatment, and inappropriate use of antibiotics (9). Due to the lack of pathognomonic clinical  
55 features that reliably distinguish dengue from other febrile illnesses, virological or serological laboratory  
56 confirmation is required for definitive diagnosis. While multiplexed tests that can quickly identify the  
57 causative pathogen are ideal, they are often unavailable in LMICs due to cost and insufficient laboratory  
58 infrastructure. Even rapid, point-of-care tests may be cost-prohibitive in LMICs (10). Accurate and cost-  
59 effective tools to better determine etiology of fever at the point-of-care are greatly needed to guide the use of  
60 diagnostics and therapeutics, conserving scarce healthcare resources.

61  
62 Clinical Decision-Support Systems (CDSS) incorporating prediction models may offer a solution to better  
63 management of infectious diseases in low resource settings. CDSSs, such as applications on smartphone  
64 devices, can gather data from a range of online sources and implement sophisticated clinical prediction  
65 models that would be impractical for clinicians to calculate manually. CDSS have proven effective at improving  
66 therapeutic management and reducing unnecessary diagnostic tests in both high-income countries (HICs) (11)  
67 and LMICs (12-14). In Bangladesh, an electronic CDSS was shown to improve clinical dehydration assessment  
68 and WHO diarrhea guideline adherence, as well as reduce non-indicated antibiotic use in children under five  
69 by 29% (12). Traditional predictive models generally incorporate clinical information that is obtained solely  
70 from the presenting patient. Predictive models that incorporate additional information – such as seasonal or  
71 climate predictors, location-specific historical prevalence, characteristics of prior patients – have been shown  
72 to increase diagnostic accuracy and limit inappropriate antibiotic use (14-16).

73  
74 The underlying probability of being infected by DENV varies by both space and time. The risk of DENV  
75 transmission depends on conditions that promote mosquito breeding, including when temperatures are  
76 warmer (17-19), and the risk of infection is influenced by local population immunity, as large outbreak years  
77 are typically followed by periods of low transmission (20-22). As most DENV transmission is highly focal, it  
78 means that population susceptibility profiles can be spatially heterogeneous at any time (21, 23-25). Thus, our  
79 objective is to develop an improved clinical prediction model for dengue by integrating temporal and spatial  
80 (location-specific) parameters including climate data, clustering of recent cases, and population susceptibility  
81 estimates derived from seroprevalence or hospital data in the surrounding community. We demonstrate the  
82 potential for integrating location- and population-specific data sources into clinical prediction models. This  
83 approach has the potential to inform the development of improved tools to aid clinicians in diagnostic and  
84 therapeutic decision making for patients presenting with suspected dengue

## 86 Methods

### 87 Location

88 Kamphaeng Phet is a province in north-central Thailand that is located 350 km north of Bangkok and has a  
89 population of 725,000 people in a mostly rural and semirural setting (26, 27). We used data collected from  
90 patients presenting to Kamphaeng Phet Provincial Hospital (KPPH), a large, tertiary care hospital in the  
91 province to identify clinical predictors that could discriminate between DENV-infected and uninfected patients  
92 (26, 27).

93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122

#### Hospital-based suspected dengue patient data

We used data on over 12,000 patients presenting to KPPH with suspected dengue between August 2007-December 2021. The data was collected by the United States Army Medical Directorate-Armed Forces Research Institute of Medical Sciences (USAMD-AFRIMS). As DENV testing in this hospital is provided free of charge and this is a highly DENV-endemic region, individuals will be tested for DENV infection if there is any suspicion of dengue, however minor. This provides an excellent test case to understand whether individual or location-specific risk factors are associated with testing positive for DENV.

For all suspected dengue cases, we used demographic and clinical information including patient age, sex, home village, admission diagnosis, date of admission, presenting symptoms, and DENV PCR status. The following signs and symptom were recorded as binary variables: fever, chills, malaise, rhinitis, rash, sore throat, seizure, cough, nuchal rigidity, eye pain, nausea, headaches, vomiting, joint pain, abnormal movements, anorexia, myalgias, diarrhea, dark urine, abdominal pain, and bleeding. DENV infection was evaluated using RT-PCR. We recorded the residence of each patient to the district (Amphoe) level using detailed base maps of the region.

#### Climate variables using National Oceanic and Atmospheric Administration (NOAA) data

Climate and seasonal factors such as temperature, precipitation, and humidity influence vector populations and DENV transmission (17-19, 28). We employed the R package GSODR to gather climate data from the central most NOAA weather station in the province of Kamphaeng Phet, Thailand, which included mean daily temperature, precipitation, dewpoint, relative humidity, sea level pressure, visibility, and windspeed. To better reflect seasonal trends, we aggregated data in 14-day increments prior to the day of the DENV infection prediction. As climate can alter vector feeding behavior (19, 29), we used aggregated climate predictors in the two weeks prior to case presentation. Additionally, climate in the months prior to outbreaks can influence both vector population dynamics as well as viral replication (19, 28). To determine the appropriate lag time for each climate variable, we constructed a random forest classifier with climate variables lagged at one, two, and three months. Using the R package, "vip", we calculated each Variable of Importance by AUC and used the best performing lag time for each climate variable.

### Estimates of temporal changes in population susceptibility using national surveillance system data

We estimate population susceptibility data using age-specific case data from the national surveillance system using data from Kamphaeng Phet province only. We note that most of the cases in this dataset are suspected DENV cases (i.e., without confirmatory testing). We have previously developed models to explicitly link underlying infection risks to the observed age distribution of cases by age and year to estimate annual age-specific force of infection in provinces of Thailand up until 2017 (30). The estimates can be used to reconstruct the buildup of immunity in populations by age. Here, we reconstruct population susceptibilities in Kamphaeng Phet going into each year, using only data prior to the year, to mimic the real-world use, where only prior years' data is available. As dengue disease severity is greatest for secondary infections, we consider two alternative formulations to define susceptibility to disease. Firstly, we consider complete susceptibility, where we use the estimates of the proportion of individuals of an age group and year that are completely seronaive. Second, we consider the proportion of individuals of an age group and year that have experience one prior infection, and are therefore at risk of increased risk of severe disease.

### Estimates of spatial differences in the underlying force of infection using seroprevalence data from a cohort study

To estimate underlying spatial differences in the force of infection in the province, we make use of a DENV cohort study in the region, where healthy individuals of all ages from throughout Kamphaeng Phet province have provided blood (31). The cohort is ongoing. We use data from samples collected during baseline blood draws, that occurred between 2015 and 2021. Hemagglutination inhibition assays were used to characterize immunity to the four DENV serotypes; individuals were considered seropositive if they had a titer of 10 or greater to any serotype. We have previously used this seroprevalence data to estimate the underlying mean force of infection, and the proportion of the population that are susceptible to DENV infection in different subdistricts in the province (32). Here, we use this subdistrict specific estimates to characterize underlying heterogeneity in the force of infection in the province. As the cohort data comes from 2015-2021, however, much of the hospital case data we are working with comes from prior to the cohort, we are assuming that the force of infection is stable in time within any location.

### Spatial clustering of positive cases based on prior patients presenting to the hospital

The local clustering of positive cases from a single area, may signal local ongoing transmission. To assess for a temporal and spatial relationship between cases, we stratified cases that presented to KPP hospital by both district and province and then summed the number of positive cases in the 30 days prior to presentation divided by the total cases over the study period from that area.

### Statistical Analysis and Modeling

We fit random forest classifiers to predict DENV infection. Random forests are a machine learning method which constructs a multitude of decision trees and averages over them to obtain a prediction robust to nonlinearities and interactions between covariates, and has been widely applied to biomedical sciences for both classification and regression (33, 34).

We initially identified the subset of clinical symptoms that were most informative of true infection status. To do this we fit random forest models using only clinical predictors and then used the R package "vip" to calculate the Variable of Importance by AUC for each clinical variable. We determined a variable's importance by calculating the change in AUC after permuting, or randomly shuffling each predictor. To attempt to achieve the most parsimonious prediction rule (i.e., the best predictive model requiring the fewest variables to be input by clinicians), we fit random forest and logistic regression models using training data with consecutively increasing clinical predictor set sizes based on the order of importance and applied this to the test set to

170 determine the smallest model with the best performance. Next, we incorporated the patient extrinsic factors.  
171 We fit each random forest classifier using 1000 decision trees and used the default number of variables to be  
172 randomly considered at each node split (*mtry* = square root of number of candidate variables). In the  
173 construction of our predictive models, we input climate predictors, age, susceptibility estimates, and the case  
174 clustering metric as continuous variables and we input the optimized clinical predictors as binary presence or  
175 absence categorical variables. Missing predictor data was imputed using the R package 'RandomForest'.  
176

177 We used logistic regression for each predictor to create a univariate comparison between DENV-positive and  
178 DENV-negative cases. We fit multiple logistic regression models to compare the performance of parsimonious  
179 models with a random forest classifier using the same number of predictors.  
180

181 To assess predictive performance for both random forest and logistic regression models, we used repeated  
182 cross-validation using 80% training/20% testing splits with 100 iterations. No testing data was used when  
183 training the model. In each iteration, predictions on the test set were produced and corresponding measures  
184 of performance obtained. To determine overall model performance, we averaged the area under the receiver  
185 operator characteristic curve (AUC) and confidence intervals for the 100 iterations. To determine statistical  
186 significance between models we used a bootstrap method over 100 iterations, which involves resampling the  
187 data with replacement multiple times, creating bootstrap samples. For each bootstrap sample, receiver  
188 operating characteristic (ROC) curves were generated and the differences between the curves were  
189 computed. All analyses were completed using R version 4.2.0, and model development/validation was  
190 completed in accordance with the TRIPOD checklist (Supplement Table S1).  
191

#### 192 Ethical considerations

193 This study was approved by the institutional review boards of the Thai Ministry of Public Health and Walter  
194 Reed Army Institute of Research (WRAIR #2119), and the University of Utah (IRB\_00150106)  
195

#### 196 **Results**

197 Of the 12,833 participants in the clinical data set, 5731 (45%) were confirmed to have DENV infection by PCR.  
198 DENV-positive patients were significantly younger (18 vs 22 years,  $p < 0.001$ , Table 1). Nearly all cases (97.8%)  
199 came from the 11 districts within Kamphaeng Phet province (Table 1). There was no significant difference  
200 between the probability of testing positive for males and females ( $p = 0.07$ ); no other genders were reported.  
201 The probability of testing positive differed substantially by age, ranging from 26% for those < 4 years to 58%  
202 for those 15-19 years of age (Table 2). Patients between the ages of 10-14 years, 15-19 years, and 5-9 years  
203 comprised the largest proportion of cases (23%, 18%, 16% respectively) while older patients comprised a  
204 much smaller proportion of cases (30-34 years 5%, 35-39 years 4%).  
205

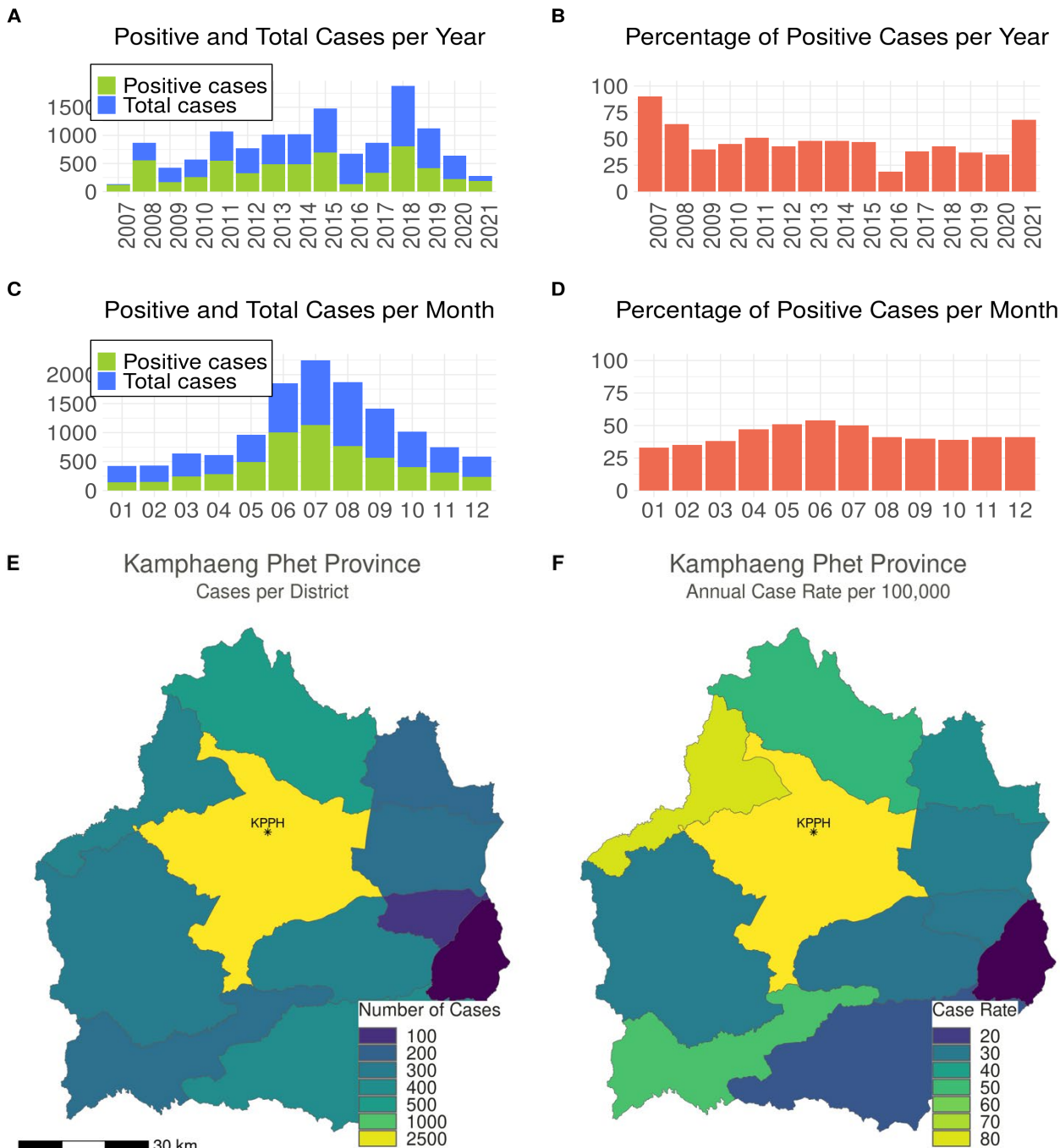
	Overall, N = 12,833 <sup>1</sup>	DENV Negative, N = 7,102 <sup>1</sup>	DENV Positive, N = 5,731 <sup>1</sup>	p-value <sup>2</sup>
Age (mean, sd)	21 (15)	22 (18)	18 (11)	<0.001
Female	6,401 (50)	3,491 (49)	2,910 (51)	0.068
<b>Symptoms</b>				
Cough	4,741 (37)	3,057 (43)	1,684 (29)	<0.001
Nausea	6,227 (49)	3,051 (43)	3,176 (55)	<0.001
Fever	11,467 (89)	6,129 (86)	5,338 (93)	<0.001
Headache	9,146 (71)	4,797 (68)	4,349 (76)	<0.001
Rhinitis	2,165 (17)	1,455 (20)	710 (12)	<0.001
Pharyngitis	3,534 (28)	2,113 (30)	1,421 (25)	<0.001
<b>Location</b>				
<u>District</u>				<0.001
Bueng Samakkhi	226 (1.8)	166 (2.3)	60 (1.0)	
Khanu Woralaksaburi	910 (7.1)	522 (7.4)	388 (6.8)	
Khlung Khlung	733 (5.7)	397 (5.6)	336 (5.9)	
Khlung Lan	945 (7.4)	645 (9.1)	300 (5.2)	
Kosamphi Nakhon	750 (5.8)	407 (5.7)	343 (6.0)	
Lan Krabue	556 (4.3)	333 (4.7)	223 (3.9)	
Mueang Kamphaeng Phet	5,780 (45)	2,910 (41)	2,870 (50)	
Pang Sila Thong	571 (4.4)	324 (4.6)	247 (4.3)	
Phran Kratai	1,186 (9.2)	684 (9.6)	502 (8.8)	
Sai Ngam	609 (4.7)	363 (5.1)	246 (4.3)	
Sai Thong Watthana	288 (2.2)	178 (2.5)	110 (1.9)	
<u>Province</u>				
Kamphaeng Phet	12,554 (97.8)	6,929 (97.5)	5,625 (98.2)	

<sup>1</sup>Mean (SD); n (%), <sup>2</sup>Wilcoxon rank sum test; Pearson's Chi-squared test

**Table 1:** Age, gender, and top discriminative symptoms by DENV positivity. Locations listed are the eleven provinces in Kamphaeng Phet.

We found that there were significant differences in the clinical symptoms between DENV positive and negative patients. Table 1 lists the top discriminative symptoms between the groups based on random forest and logistic regression. The most common symptom reported was fever, followed by headache. In univariate analysis, we found that individuals with fever, chills, malaise, retro-orbital pain, nausea, headache, and vomiting were significantly more likely to test positive for DENV, and individuals with cough, rhinitis, pharyngitis were significantly less likely to test positive for DENV (Supplementary Table S2).

When we examined the proportion of positive cases to total cases by year and month, we found that both total and positive cases significantly increased in the months between June and September ( $p < 0.001$ ). The proportion of positive cases differed substantially by year ( $p < 0.001$ ), ranging from 19% in 2016 to 90% in 2017. The period of lowest test-positivity in 2016 and 2017, coincided with the Zika virus epidemic in the country (Figure 1).

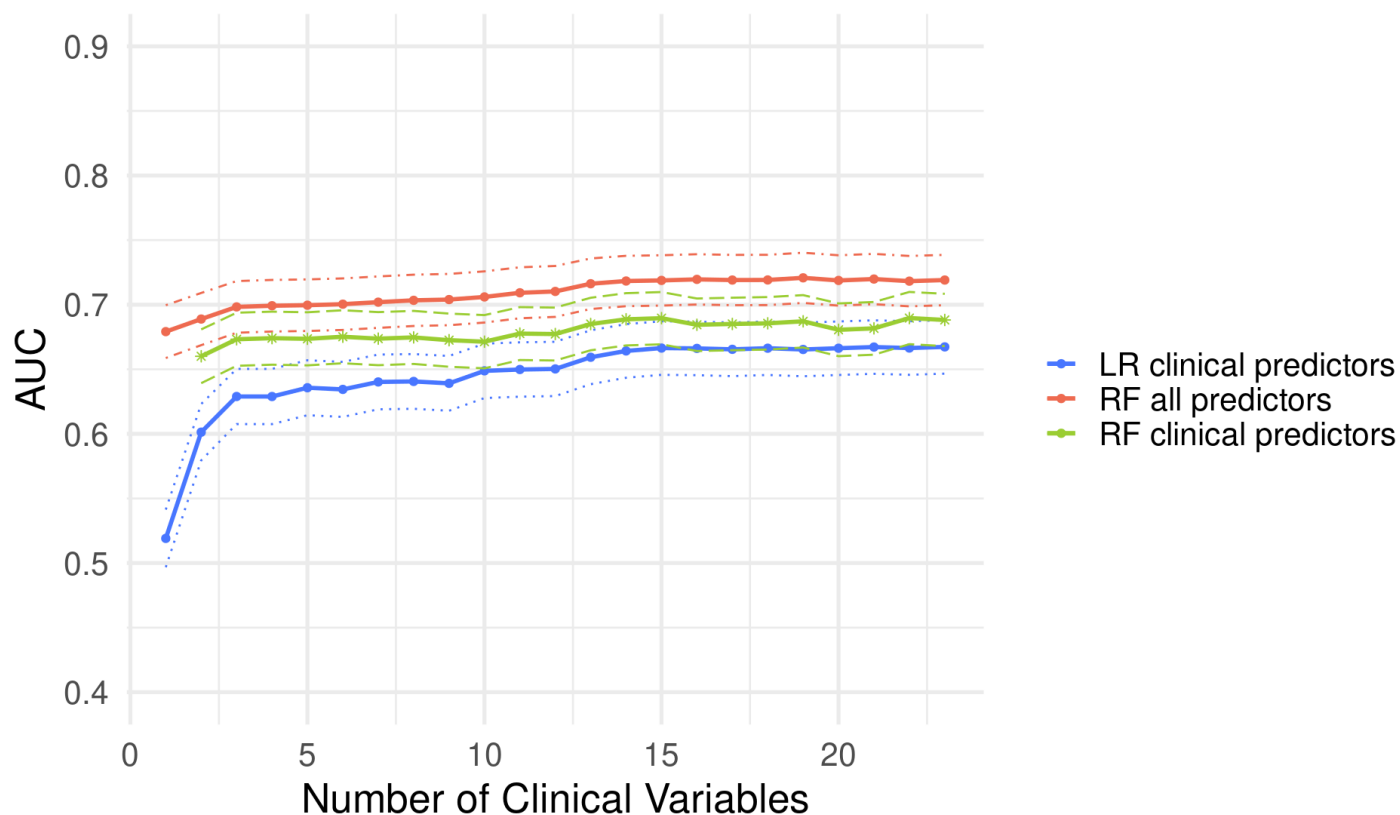


**Figure 1.** Dengue virus (DENV) cases at Kamphaeng Phet Provincial Hospital (KPPH), Thailand, 2007-2021. The number of DENV cases (green) over total cases (blue) as proportion of AFI cases by year (A) and month (C) and the percentage of positive cases by year (B) and month (D) over the study period. A map of Kamphaeng Phet Province and its 11 districts. Colors indicate the number of positive cases (E) and the annual case rate per 100,000 persons (F) within each district between 2007-2021.

### Model performance evaluation using only clinical predictors and parsimonious variable selection

We first assessed the performance of the model using a traditional clinical prediction model which only includes the presenting patient's information. A random forest classifier using all 23 clinical features resulted in an average AUC of 69.5% (95%CI: 67.5-71.5) from repeated cross-validation. To determine the optimal number of variables for a parsimonious prediction model, we used a random forest classifier to analyze the

236 improvement in model performance with each additional clinical variable included. Figure 2 shows the  
 237 improvement in AUC with each additional variable using two random forest classifiers – one with all other  
 238 predictors and the other using only clinical data – as well as a logistic regression model using only clinical  
 239 variables. Performance levelled off with three clinical variables: age, cough, and nausea. Using a model with  
 240 only these three predictors, we achieve an average AUC of 67.0% (95%CI: 65.0-69.1). Supplementary Table S3  
 241 shows the relative frequency of these variables by age group. We demonstrate the direction and magnitude of  
 242 the effect of the top predictors by generating partial dependence plots from random forest and logistic  
 243 regression classifiers (Supplementary Figure S1).  
 244



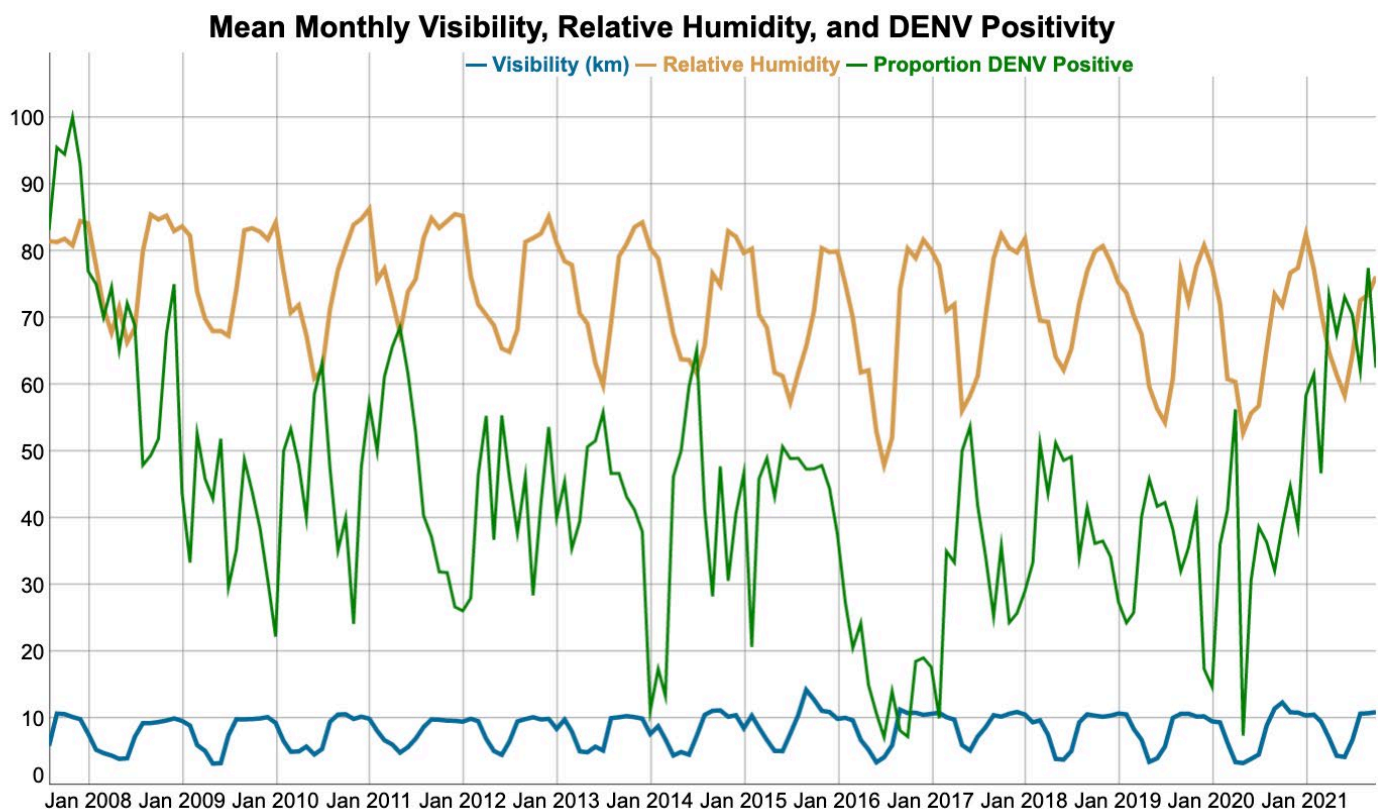
245  
 246  
 247 **Figure 2.** Average AUC and 95% CIs from cross-validation (100 iterations) for Random Forest (RF) and Logistic  
 248 Regression (LR) models. The red line indicates an RF model with all other predictors (climate, reconstructed  
 249 susceptibilities estimates, force of infection estimates, prior patients) included. The green line indicates an RF  
 250 model which includes only clinical predictors. The blue line indicates an LR model with only clinical predictors  
 251 included. The dotted lines indicate CIs.  
 252

253 **Addition of climate data to the clinical parameters model resulted in an improved area under the receiver  
 254 operating characteristic curve**

255 Next, we fit models using climate data. To appropriately adjust lag time for each climate variable, we fit a  
 256 random forest classifier using only climate variables and assessed the Variables of Importance by AUC. A  
 257 random forest model with recent and lagged aggregated climate data without clinical predictors resulted in an  
 258 AUC of 58.7% (95% CI: 56.5-60.9). We found the best performing climate variables were visibility, relative  
 259 humidity, wind speed, and precipitation, all lagged by 3 months. We examined the relationship between the  
 260 top two performing climate predictors - visibility and relative humidity - with the proportion of positive cases  
 261 each month (Figure 3). For each climate predictor, Supplementary Table S4 lists the odds ratio and compares

262 the mean of each predictor by DENV-positive or negative groups. When combined with the top three clinical  
263 variables, climate data performed similarly (AUC of 67.2%, 95%CI:65.2-69.3) as clinical data alone (AUC of  
264 67.0%, 95%CI: 65.0-69.1) (median p = 0.60, 2% p-values <0.05). However, when climate data was combined  
265 with all other predictors, model performance improved from an AUC of 68.4% (95% CI: 66.4-70.4) to and AUC  
266 of 70.0% (95% CI: 67.9-71.0; median p = 0.07, 45% p-values < 0.05). To assess whether integrating more  
267 location specific climate data would improve performance, we fit models using climate data from each case's  
268 home district, however, model performance did not noticeably change. Table 2 shows the AUCs for the clinical  
269 base model, compared to the base model plus the inclusion of additional data sources.

270  
271



272  
273  
274  
275  
276

**Figure 3.** The monthly relative humidity (orange) and visibility (blue) in Thailand over the study period, compared with rates of DENV (green). For each case, we gathered the nearest NOAA weather station's climate data, lagged by three months, and averaged that data for each month.

Model	AUC (%)	95% CI
Clinical*Climate*RS*Fol*Cluster	70.0	67.9-71.9
Clinical*Climate*RS*Cluster	69.5	67.5-71.5
Clinical*Climate*Fol*Cluster	69.2	67.2-71.2
Clinical*Climate*Cluster	68.8	66.8-70.8
Clinical*Climate*RS*Fol	68.7	66.7-70.7
Clinical*Cluster	68.7	66.7-70.7
Clinical*Fol*Cluster	68.5	66.5-70.6
Clinical*Climate*RS	68.4	66.4-70.5
Clinical*RS*Fol*Cluster	68.4	66.4-70.4
Clinical*RS*Cluster	68.2	66.1-70.2
Clinical*Climate*Fol	68.1	66.1-70.1
Clinical*Fol	67.7	65.7-69.8
Clinical*RS*Fol	67.6	65.5-69.6
Climate*RS*Fol*Cluster	67.5	65.5-69.6
Clinical*RS	67.5	65.4-69.5
Clinical*Climate	67.2	65.2-69.3
Clinical	67.0	65-69.1
Climate*RS*Cluster	66.8	64.8-68.9
Climate*RS	65.7	63.6-67.8
RS*Cluster	65.7	63.6-67.7
RS	65.6	63.5-67.7
Climate*Fol*Cluster	64.7	62.6-66.8
Climate*Cluster	60.5	58.3-62.7
Climate	58.7	56.5-60.9
Cluster	56.4	54.2-58.6
Fol	57.0	54.8-59.2

278

279

280

281

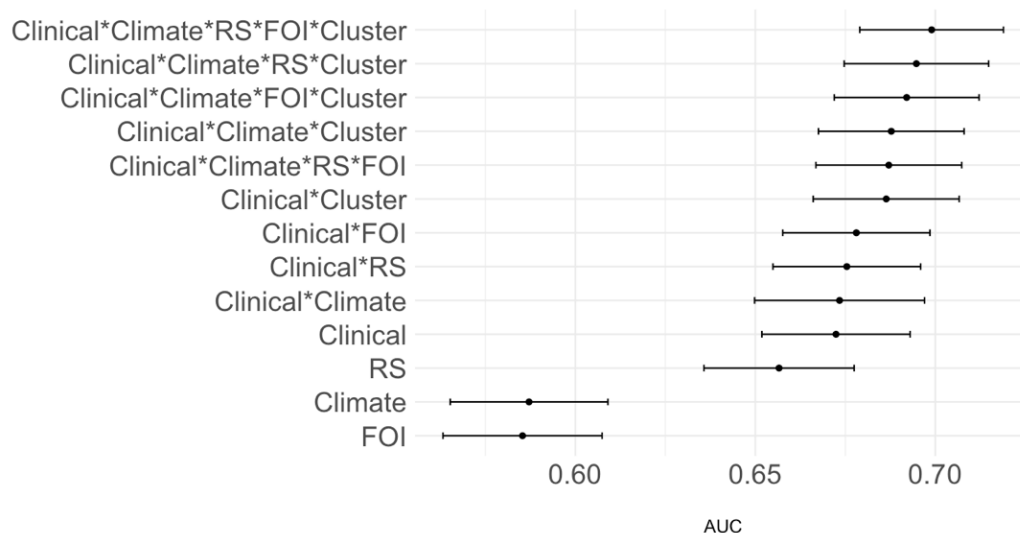
282

283

284

**Table 2.** The AUCs and confidence intervals by base model, compared to base model plus inclusion of additional data sources. 'Clinical' indicates the inclusion of the top three clinical predictors, 'Climate' indicates the inclusion of climate predictors, 'RS' indicates the inclusion of reconstructed susceptibility estimates derived using national surveillance data, 'FOI' indicates the inclusion of force of infection estimates derived using cohort data, 'Cluster' indicates the recent case cluster metric.

## AUC by Model



**Figure 4.** The AUCs and confidence intervals by base model, compared to base model plus inclusion of additional data sources. ‘Clinical’ indicates the inclusion of the top three clinical predictors, ‘Climate’ indicates the inclusion of climate predictors, ‘RS’ indicates the inclusion of reconstructed susceptibility estimates derived using national surveillance data, ‘FOI’ indicates the inclusion of force of infection estimates derived using cohort data, ‘Cluster’ indicates the recent case cluster metric.

### **Addition of reconstructed susceptibility (RS) estimates to the clinical parameters model resulted in an improved area under the receiver operating characteristic curve.**

Using historical hospital case data from the province, we obtained estimates of the size of the susceptible population by age for each year (across all subdistricts in the province). In our predictive model we used the prior year’s RS estimates. Using logistic regression, we found secondary RS estimates performed better than primary RS estimates [60.7% (95%CI: 58.6–62.9) vs 52.3% (95%CI:50.1–54.6)]. When added to a random forest classifier with climate and/or clinical predictors, the inclusion of RS estimates consistently resulted in higher AUCs (Table 2). When added to the top 3 clinical parameters alone, RS estimates non-significantly improved model performance from an AUC of 67.0% (95%CI: 65.0–68.8) to an AUC of 67.5% (95%CI: 65.4–69.5), (median  $p=0.40$ , 9%  $p$ -values < .05). Finally, a model including all predictors resulted in higher AUCs than a model without RS (median  $p=0.09$ , 32%  $p$ -values < 0.05).

### **Addition of subdistrict-specific Force of Infection (Fol) estimates to the clinical parameters model resulted in an improved area under the receiver operating characteristic curve.**

We incorporated Fol estimates for each age by subdistrict using data from a local cohort study. This assumes that the underlying differences in the force of infection are constant in time. Using logistic regression, Fol estimates had an AUC of 57.0% (95%CI: 54.8–59.2). The inclusion of Fol estimates lead to increases in AUC when added to the top clinical predictors, when added to clinical predictors and climate data, and when added to clinical predictors, climate predictors, and RS estimates (Table 2). When included with all other predictors, a model with Fol estimates non-significantly improved performance compared to a model without Fol estimates (median  $p=0.30$ , 23%  $p$ -values < 0.05)

### **Addition of the case clustering metric to the clinical parameters model resulted in an improved area under the receiver operating characteristic curve.**

316 Finally, we fit a model that assessed for clustering of recent cases based on prior patients presenting to the  
317 KPP hospital. Using logistic regression, we found the case clustering metric (the number of positive cases in  
318 the subdistrict over last 30 days divided by the total number of cases from that subdistrict in the study period)  
319 had an AUC of 56.4% (95%CI: 54.2-58.6). We found that the use of the case clustering metric consistently  
320 improved model performance. Stratifying by the finer spatial size of subdistrict consistently outperformed  
321 models with prior patients stratified by province. When added to the top performing clinical variables, model  
322 performance significantly improved (median  $p=0.02$ , 60% of  $p$ -values  $<0.05$ ). When compared to a model  
323 with all predictors except cluster of recent cases, the inclusion of this predictor significantly improved model  
324 performance (median  $p=0.007$ , 79%  $p$ -values  $<0.05$ ).

325  
326 Finally, when comparing a model including all predictors with a model including only the top clinical predictors  
327 model performance improved from an AUC of 67.0% (95%CI: 65.0-69.1) to an AUC of 70.0% [(95%CI: 67.9-  
328 71.9) (median  $p=0.006$ , 87%  $p$ -values  $<0.05$ )]. Our model had a sensitivity of 55.3%, a specificity of 70.2%, a  
329 positive predictive value (PPV) of 60.0%, and a negative predictive value (NPV) of 66.1%.

### 330 Discussion

331 Insufficient diagnostic testing capacity in LMICs necessitates innovative approaches to support clinical  
332 decision-making. Here, we present a predictive model for DENV infection that integrates multiple sources of  
333 information both intrinsic and extrinsic to the patient, including climate data, clinical data, seroprevalence-  
334 based susceptibility estimates, and historical information from prior patients, which results in improved  
335 predictive performance. While the model with all predictors included did significantly outperform the base  
336 parsimonious model with only clinical predictors (median  $p=0.006$ , 87%  $p$ -values  $<0.05$ ), whether the  
337 additional 3.0% improvement in AUC is clinically useful may be case- and clinician-dependent. Certain  
338 components of our model require data from sero-surveillance, which may not be accessible in all communities.  
339 However, simplifying the model by including only the top clinical predictors and the case cluster metric alone  
340 results in an AUC decrease of only 1.3%. These metrics are more readily obtainable and, notably, do not  
341 necessitate laboratory resources. Nevertheless, we believe that the results demonstrate a proof-of-concept  
342 that seroprevalences-based susceptibility estimates and climate data can be used to improve predictive  
343 performance and may be useful to augment prediction in other communicable diseases.

344  
345 There is a lack of information on the deficiency of testing capacity both in Thailand and globally in LMICs.  
346 Accurately quantifying the true extent of diagnostic testing deficiencies is challenging as LMICs often lack  
347 robust national surveillance systems. In a Brazilian study between the years 2010-2019, where every  
348 suspected case of dengue was recorded in a national surveillance database, only 11% of the 350,000 cases of  
349 suspected dengue infection were ultimately tested(35). If we extrapolate the results from Brazil to other  
350 LMICs, as much as 90% of dengue like illness may go undiagnosed, highlighting the need for tools to bridge the  
351 diagnostic gap.

352  
353 In contrast to most dengue diagnostic models, which rely in part on laboratory data, our model relies solely on  
354 clinical indicators, making it accessible to clinicians without laboratory resources. For reference, when  
355 compared to a multiple regression model from Honduras that used only clinical predictors, our model had a  
356 lower sensitivity (55% vs 86%) and PPV (60% vs 75%) and a higher specificity (70% vs 27%) and NPV (66%  
357 vs 44%) (36). Models that integrate laboratory values, such as complete blood count, and hepatic function tests,  
358 tend to perform better than models using only clinical predictors, such as a Bayesian network model from  
359 Thailand (sensitivity 74%, specificity 79%, PPV 75%, NPV 79%) (37), a multiple regression model from Sri  
360 Lanka (sensitivity of 49%, a specificity of 85%, PPV of 70%, NPV of 70%) (38), and a multiple regression  
361 model from Brazil (sensitivity 80%, specificity 71%) (39).

362

363  
364  
365  
366  
367  
368  
369  
370  
371  
372

DENV transmission can exhibit significant temporal and geographical heterogeneity even at fine spatial scales, with variations observed even among neighboring villages (27, 40, 41). We thus used patient-extrinsic (location-specific) data sources in our models. The improvement in AUC with finer spatial units suggests that population-level spatial heterogeneity exists at the district level and can be applied to individual-level clinical prediction. We expect further improvements in predictive performance if finer-scale location became routinely available for case data, such as to the community level. The improvement with the use of either the province or district level case clustering metric highlights the utility of temporal predictors in clinical prediction DENV models.

373  
374  
375  
376  
377  
378  
379

Spatial heterogeneity in dengue incidence may be explained in part by micro-climates, which can modify transmission dynamics at small scales. For example, within urban heat islands, temperature variations of up to 10°C compared to other city areas may create conditions more conducive to dengue transmission in cooler temperatures (42). We collected all climate data from the provincial weather station in Kamphaeng Phet. We attempted to integrate climate data at a more localized level, however several subdistricts either do not have weather stations or weather station data was incomplete. When fitting models using data from all districts in Kamphaeng Phet, however, we found similar results.

380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396

Transmission of DENV occurs in a seasonal pattern, and several climate variables have been found to increase DENV transmission and/or vector populations (17-19, 28, 29). While prior studies have demonstrated associations between climate variables like average precipitation, relative humidity, temperature, and windspeed, with varying lag times between 0-3 months, and dengue incidence (43-46), it is important to note that our predictive-based analytic framework is not intended to examine causal or associative relationships between climate variables and the outcome of dengue incidence. Our findings suggest that site-specific climate variables aid in site-specific models to predict DENV infection. While visibility has not been found to be associated with dengue incidence, we found it was the most important climate predictor. It is plausible that visibility serves as a proxy indicator for an underlying factor that does impact dengue incidence, such as air pollution, which has been postulated as a contributing factor (47, 48). Appropriate lag times would need to be tuned to different sites. For use in a clinical decision support tool, the most recent climate variables could be gathered from online weather sources, based on smartphone-based detection of GPS location. An optimal utilization of this model would be through a smartphone application, as there is a scarcity of electronic medical record availability in LMICs. This would necessitate access to a smart phone device and internet connection; however, clinicians and frontline healthcare workers increasingly have access to smartphone devices, even in remote areas of LMICs (49).

397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408

There were significant differences between DENV-positive and -negative patients in 16 of the 22 clinical symptoms collected on presentation, consistent with features known to distinguish dengue from other illnesses (50, 51). To minimize clinician input requirements (52), we used random forest regression to identify the optimal variables to derive a parsimonious model. We were able to achieve near-optimal performance with only three clinical variables – age, nausea, and cough. Numerous multivariable models based on clinical presentation have been developed to identify dengue infection in patients with AFI. In a review of published logistic regression prediction models, rash and/or petechiae was the most frequently identified predictor (four of seven models) to discriminate between DENV-positive and DENV-negative patients. When evaluated, the absence of cough was found to be a predictor in 33% of models. Nausea, which was evaluated in four logistic regression models, did not achieve significance in any model. Our results differ from those found in many logistic regression models and align with more intricate models for DENV diagnosis. Models employing deep

neural networks (53), random forest (54), and gradient boosting (XGBoost) (55) noted that age was the best clinical discriminative predictor. These models did not include cough or nausea as variables for assessment. We found that with the input of as little as one clinical variable – age – along with other predictors can provide useful clinical information (AUC 67.9%, 95%CI: 65.6-70.0), especially in cases where other symptoms cannot be easily obtained, such as in nonverbal or comatose patients.

We show that reconstructed susceptibility estimates, which reflect the transmission dynamics of disease and the susceptible proportion of a population, improve individual level clinical prediction on their own. However, there are several factors that make the use of reconstructed susceptibility estimates problematic and we favor the use of other location-specific predictors. First, reconstructed susceptibility estimates may be more difficult to obtain across different settings. Moreover, reconstructed susceptibility estimates may not serve as a reliable indicator of protection against DENV, as they represent a mixed concept – immunity may reflect protection due to herd immunity or may indicate increased risk of dengue infection, as higher levels of immunity may reflect higher viral circulation of the multiple DENV serotypes with significant immunologic cross-reactivity. Finally, it should be noted that reconstructed susceptibility estimates are themselves derived from a model and so should be considered with caution.

Our study has several limitations. First, our model was constructed using data from a single center and testing was limited to patients suspected of having dengue infection, potentially hindering the model's generalizability to a broader population. Similarly, as there was inherent heuristic bias in the patients selected for testing, the clinical components of the model reflect this specific population, meaning other important predictors of dengue infection, such as fever, were already included in the clinician's decision making. Our results were limited to internal cross-validation; further studies for external validation are necessary. Finally, our assessment of the use of spatial dynamics in DENV transmission was limited as cases were only matched to each district rather than sub-district or village. In the future, models that integrate cases based on a finer spatial scale may better assess the role of a patient's residing location in prediction. Despite these limitations, we demonstrate that predictive models that include patient-extrinsic location-specific elements can improve prediction and allow for parsimonious models that minimize clinician input and should be considered in future work on clinical prediction and decision support tools.

## References

1. J. Osborn, T. Roberts, E. Guillen, O. Bernal, P. Roddy, S. Ongarello, A. Sprecher, A.-L. Page, I. Ribeiro, E. Piriou, A. Tamrat, R. de la Tour, V. B. Rao, L. Fleवाद, T. Jensen, L. McIver, C. Kelly, S. Dittrich, Prioritising pathogens for the management of severe febrile patients to improve clinical care in low- and middle-income countries. *BMC Infectious Diseases* **20**, 117 (2020).
2. N. Prasad, D. R. Murdoch, H. Reyburn, J. A. Crump, Etiology of Severe Febrile Illness in Low- and Middle-Income Countries: A Systematic Review. *PloS one* **10**, e0127962-e0127962 (2015).
3. D. R. Feikin, B. Olack, G. M. Bigogo, A. Audi, L. Cosmas, B. Aura, H. Burke, M. K. Njenga, J. Williamson, R. F. Breiman, The Burden of Common Infectious Disease Syndromes at the Clinic and Household Level from Population-Based Surveillance in Rural and Urban Kenya. *PLOS ONE* **6**, e16085 (2011).
4. L. K. Archibald, M. O. den Dulk, K. J. Pallangyo, L. B. Reller, Fatal Mycobacterium tuberculosis bloodstream infections in febrile hospitalized adults in Dar es Salaam, Tanzania. *Clin Infect Dis* **26**, 290-296 (1998).
5. K. Chheng, M. J. Carter, K. Emary, N. Chanpheaktra, C. E. Moore, N. Stoesser, H. Putschat, S. Sona, S. Reaksmey, P. Kitsutani, B. Sar, H. R. van Doorn, N. H. Uyen, L. Van Tan, D. Paris, S. D. Blacksell, P.

- Amornchai, V. Wuthiekanun, C. M. Parry, N. P. J. Day, V. Kumar, A Prospective Study of the Causes of Febrile Illness Requiring Hospitalization in Children in Cambodia. *PLOS ONE* **8**, e60634 (2013).
6. J. A. Crump, A. B. Morrissey, W. L. Nicholson, R. F. Massung, R. A. Stoddard, R. L. Galloway, E. E. Ooi, V. P. Maro, W. Saganda, G. D. Kinabo, C. Muiruri, J. A. Bartlett, Etiology of Severe Non-malaria Febrile Illness in Northern Tanzania: A Prospective Cohort Study. *PLOS Neglected Tropical Diseases* **7**, e2324 (2013).
7. F. N. Ssali, M. R. Kanya, F. Wabwire-Mangen, S. Kasasa, M. Joloba, D. Williams, R. D. Mugerwa, J. J. Ellner, J. L. Johnson, A prospective study of community-acquired bloodstream infections among febrile adults admitted to Mulago Hospital in Kampala, Uganda. *J Acquir Immune Defic Syndr Hum Retrovirol* **19**, 484-489 (1998).
8. S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, S. I. Hay, The global distribution and burden of dengue. *Nature* **496**, 504-507 (2013).
9. J. A. Crump, S. Gove, C. M. Parry, Management of adolescents and adults with febrile illness in resource limited areas. *Bmj* **343**, d4847 (2011).
10. P. Yager, G. J. Domingo, J. Gerdes, Point-of-Care Diagnostics for Global Health. *Annual Review of Biomedical Engineering* **10**, 107-144 (2008).
11. T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. R. Coeytaux, G. Samsa, V. Hasselblad, J. W. Williams, M. D. Musty, L. Wing, A. S. Kendrick, G. D. Sanders, D. Lobach, Effect of Clinical Decision-Support Systems. *Annals of Internal Medicine* **157**, 29-43 (2012).
12. S. Bilal, E. Nelson, L. Meisner, M. Alam, S. Al Amin, Y. Ashenafi, S. Teegala, A. F. Khan, N. Alam, A. Levine, Evaluation of Standard and Mobile Health-Supported Clinical Diagnostic Tools for Assessing Dehydration in Patients with Diarrhea in Rural Bangladesh. *The American journal of tropical medicine and hygiene* **99**, 171-179 (2018).
13. F. F. Tuon, J. Gasparetto, L. C. Wollmann, T. P. Moraes, Mobile health application to assist doctors in antibiotic prescription - an approach for antibiotic stewardship. *Braz J Infect Dis* **21**, 660-664 (2017).
14. S. C. Garbern, E. J. Nelson, S. Nasrin, A. M. Keita, B. J. Brintz, M. Gainey, H. Badji, D. Nasrin, J. Howard, M. Taniuchi, J. A. Platts-Mills, K. L. Kotloff, R. Haque, A. C. Levine, S. O. Sow, N. H. Alam, D. T. Leung, External validation of a mobile clinical decision support system for diarrhea etiology prediction in children: A multicenter study in Bangladesh and Mali. *Elife* **11**, (2022).
15. A. M. Fine, J. S. Brownstein, L. E. Nigrovic, A. A. Kimia, K. L. Olson, A. D. Thompson, K. D. Mandl, Integrating Spatial Epidemiology Into a Decision Model for Evaluation of Facial Palsy in Children. *Archives of Pediatrics & Adolescent Medicine* **165**, 61-67 (2011).
16. E. J. Nelson, A. I. Khan, A. M. Keita, B. J. Brintz, Y. Keita, D. Sanogo, M. T. Islam, Z. H. Khan, M. M. Rashid, D. Nasrin, M. H. Watt, S. M. Ahmed, B. Haaland, A. T. Pavia, A. C. Levine, D. L. Chao, K. L. Kotloff, F. Qadri, S. O. Sow, D. T. Leung, Improving Antibiotic Stewardship for Diarrheal Disease With Probability-Based Electronic Clinical Decision Support: A Randomized Crossover Trial. *JAMA Pediatr* **176**, 973-979 (2022).
17. M. Chan, M. A. Johansson, The incubation periods of Dengue viruses. *PLoS One* **7**, e50972 (2012).
18. D. M. Watts, D. S. Burke, B. A. Harrison, R. E. Whitmire, A. Nisalak, Effect of temperature on the vector efficiency of *Aedes aegypti* for dengue 2 virus. *Am J Trop Med Hyg* **36**, 143-152 (1987).
19. R. Barrera, M. Amador, A. J. MacKay, Population dynamics of *Aedes aegypti* and dengue as influenced by weather and human behavior in San Juan, Puerto Rico. *PLoS Negl Trop Dis* **5**, e1378 (2011).
20. G. S. Ribeiro, G. L. Hamer, M. Diallo, U. Kitron, A. I. Ko, S. C. Weaver, Influence of herd immunity in the cyclical nature of arboviruses. *Curr Opin Virol* **40**, 1-10 (2020).

- 501 21. V. Romeo-Aznar, L. Picinini Freitas, O. Gonçalves Cruz, A. A. King, M. Pascual, Fine-scale heterogeneity  
502 in population density predicts wave dynamics in dengue epidemics. *Nat Commun* **13**, 996 (2022).
- 503 22. J. Lourenço, M. Recker, Natural, persistent oscillations in a spatial multi-strain disease system with  
504 application to dengue. *PLoS Comput Biol* **9**, e1003308 (2013).
- 505 23. W. T. Lai, C. H. Chen, H. Hung, R. B. Chen, S. Shete, C. C. Wu, Recognizing spatial and temporal  
506 clustering patterns of dengue outbreaks in Taiwan. *BMC Infect Dis* **18**, 256 (2018).
- 507 24. M. I. Estupiñán Cárdenas, V. M. Herrera, M. C. Miranda Montoya, A. Lozano Parra, Z. M. Zaraza  
508 Moncayo, J. P. Flórez García, I. Rodríguez Barraquer, L. Villar Centeno, Heterogeneity of dengue  
509 transmission in an endemic area of Colombia. *PLoS Negl Trop Dis* **14**, e0008122 (2020).
- 510 25. K. T. Thai, N. Nagelkerke, H. L. Phuong, T. T. Nga, P. T. Giao, L. Q. Hung, T. Q. Binh, N. V. Nam, P. J. De  
511 Vries, Geographical heterogeneity of dengue transmission in two villages in southern Vietnam.  
512 *Epidemiol Infect* **138**, 585-591 (2010).
- 513 26. P. Kerdpanich, S. Kongkiatngam, D. Buddhari, S. Simasathien, C. Klungthong, P. Rodpradit, B.  
514 Thaisomboonsuk, T. Wongstitwilairoong, T. Hunsawong, K. B. Anderson, S. Fernandez, A. R. Jones,  
515 Comparative Analyses of Historical Trends in Confirmed Dengue Illnesses Detected at Public Hospitals  
516 in Bangkok and Northern Thailand, 2002-2018. *Am J Trop Med Hyg* **104**, 1058-1066 (2020).
- 517 27. P. Bhoomboonchoo, R. V. Gibbons, A. Huang, I. K. Yoon, D. Buddhari, A. Nisalak, N. Chansatiporn, M.  
518 Thipayamongkolgul, S. Kalanarooj, T. Endy, A. L. Rothman, A. Srikiatkachorn, S. Green, M. P.  
519 Mammen, D. A. Cummings, H. Salje, The spatial dynamics of dengue virus in Kamphaeng Phet,  
520 Thailand. *PLoS Negl Trop Dis* **8**, e3138 (2014).
- 521 28. S. Flores Ruiz, S. Cabrera Romo, A. Castillo Vera, A. Dor, Effect of the Rural and Urban Microclimate on  
522 Mosquito Richness and Abundance in Yucatan State, Mexico. *Vector Borne Zoonotic Dis* **22**, 281-288  
523 (2022).
- 524 29. T. W. Scott, P. H. Amerasinghe, A. C. Morrison, L. H. Lorenz, G. G. Clark, D. Strickman, P. Kittayapong, J.  
525 D. Edman, Longitudinal studies of *Aedes aegypti* (Diptera: Culicidae) in Thailand and Puerto Rico: blood  
526 feeding frequency. *J Med Entomol* **37**, 89-101 (2000).
- 527 30. A. T. Huang, S. Takahashi, H. Salje, L. Wang, B. Garcia-Carreras, K. Anderson, T. Endy, S. Thomas, A. L.  
528 Rothman, C. Klungthong, A. R. Jones, S. Fernandez, S. Iamsirithaworn, P. Doung-Ngern, I. Rodriguez-  
529 Barraquer, D. A. T. Cummings, Assessing the role of multiple mechanisms increasing the age of dengue  
530 cases in Thailand. *Proceedings of the National Academy of Sciences* **119**, e2115790119 (2022).
- 531 31. K. B. Anderson, D. Buddhari, A. Srikiatkachorn, G. D. Gromowski, S. Iamsirithaworn, A. L. Weg, D. W.  
532 Ellison, L. Macareo, D. A. T. Cummings, I.-K. Yoon, A. Nisalak, A. Ponlawat, S. J. Thomas, S. Fernandez, R.  
533 G. Jarman, A. L. Rothman, T. P. Endy, An Innovative, Prospective, Hybrid Cohort-Cluster Study Design to  
534 Characterize Dengue Virus Transmission in Multigenerational Households in Kamphaeng Phet,  
535 Thailand. *American Journal of Epidemiology* **189**, 648-659 (2020).
- 536 32. G. A.-O. Ribeiro Dos Santos, D. Buddhari, S. Iamsirithaworn, D. Khampaen, A. Ponlawat, T. Fansiri, A.  
537 Farmer, S. Fernandez, S. Thomas, I. Rodriguez Barraquer, A. Srikiatkachorn, A. T. Huang, D. A. T.  
538 Cummings, T. Endy, A. L. Rothman, H. A.-O. Salje, K. A.-O. Anderson, Individual, Household, and  
539 Community Drivers of Dengue Virus Infection Risk in Kamphaeng Phet Province, Thailand.
- 540 33. A. Sarica, A. Cerasa, A. Quattrone, Random Forest Algorithm for the Classification of Neuroimaging  
541 Data in Alzheimer's Disease: A Systematic Review. *Front Aging Neurosci* **9**, 329 (2017).
- 542 34. S. Y. Peng, Y. C. Chuang, T. W. Kang, K. H. Tseng, Random forest can predict 30-day mortality of  
543 spontaneous intracerebral hemorrhage with remarkable discrimination. *Eur J Neurol* **17**, 945-950  
544 (2010).
- 545 35. G. Ribeiro Dos Santos, B. Durovni, V. Saraceni, T. I. Souza Riback, S. B. Pinto, K. L. Anders, L. A. Moreira,  
546 H. Salje, Estimating the effect of the wMel release programme on the incidence of dengue and

- 547 chikungunya in Rio de Janeiro, Brazil: a spatiotemporal modelling study. *Lancet Infect Dis* **22**, 1587-  
548 1595 (2022).
- 549 36. E. Fernández, M. Smieja, S. D. Walter, M. Loeb, A predictive model to differentiate dengue from other  
550 febrile illness. *BMC Infect Dis* **16**, 694 (2016).
- 551 37. C. Sa-Ngamuang, P. Haddawy, V. Luvira, W. Piyaphanee, S. Iamsirithaworn, S. Lawpoolsri, Accuracy of  
552 dengue clinical diagnosis with and without NS1 antigen rapid test: Comparison between human and  
553 Bayesian network model decision. *PLoS Negl Trop Dis* **12**, e0006573 (2018).
- 554 38. C. K. Bodinayake, L. G. Tillekeratne, A. Nagahawatte, V. Devasiri, W. Kodikara Arachchi, J. J. Strouse, O.  
555 M. Sessions, R. Kurukulasoorya, A. Uehara, S. Howe, X. M. Ong, S. Tan, A. Chow, P. Tummalapalli, A. D.  
556 De Silva, T. Østbye, C. W. Woods, D. J. Gubler, M. E. Reller, Evaluation of the WHO 2009 classification  
557 for diagnosis of acute dengue in a large cohort of adults and children in Sri Lanka during a dengue-1  
558 epidemic. *PLoS Negl Trop Dis* **12**, e0006258 (2018).
- 559 39. R. P. Daumas, S. R. Passos, R. V. Oliveira, R. M. Nogueira, I. Georg, K. B. Marzochi, P. Brasil, Clinical and  
560 laboratory features that discriminate dengue from other febrile illnesses: a diagnostic accuracy study in  
561 Rio de Janeiro, Brazil. *BMC Infect Dis* **13**, 77 (2013).
- 562 40. A. C. Restrepo, P. Baker, A. C. Clements, National spatial and temporal patterns of notified dengue  
563 cases, Colombia 2007-2010. *Trop Med Int Health* **19**, 863-871 (2014).
- 564 41. I. K. Yoon, A. Getis, J. Aldstadt, A. L. Rothman, D. Tannitisupawong, C. J. Koenraadt, T. Fansiri, J. W.  
565 Jones, A. C. Morrison, R. G. Jarman, A. Nisalak, M. P. Mammen, Jr., S. Thammapalo, A. Srikiatkachorn,  
566 S. Green, D. H. Libraty, R. V. Gibbons, T. Endy, C. Pimgate, T. W. Scott, Fine scale spatiotemporal  
567 clustering of dengue virus transmission in children and *Aedes aegypti* in rural Thai villages. *PLoS Negl*  
568 *Trop Dis* **6**, e1730 (2012).
- 569 42. R. Misslin, O. Telle, E. Daudé, A. Vaguet, R. E. Paul, Urban climate versus global climate change-what  
570 makes the difference for dengue? *Ann N Y Acad Sci* **1382**, 56-72 (2016).
- 571 43. N. Abdullah, N. C. Dom, S. A. Salleh, H. Salim, N. Precha, The association between dengue case and  
572 climate: A systematic review and meta-analysis. *One Health* **15**, 100452 (2022).
- 573 44. S. Hossain, M. M. Islam, M. A. Hasan, P. B. Chowdhury, I. A. Easty, M. K. Tusar, M. B. Rashid, K. Bashar,  
574 Association of climate factors with dengue incidence in Bangladesh, Dhaka City: A count regression  
575 approach. *Heliyon* **9**, e16053 (2023).
- 576 45. C. A. Ouattara, T. I. Traore, S. Traore, I. Sangare, C. Z. Meda, L. G. B. Savadogo, Climate factors and  
577 dengue fever in Burkina Faso from 2017 to 2019. *J Public Health Afr* **13**, 2145 (2022).
- 578 46. N. Singh, R. K. Mall, T. Banerjee, A. Gupta, Association between climate and infectious diseases among  
579 children in Varanasi city, India: A prospective cohort study. *Sci Total Environ* **796**, 148769 (2021).
- 580 47. H. C. Lu, F. Y. Lin, Y. H. Huang, Y. T. Kao, E. W. Loh, Role of air pollutants in dengue fever incidence:  
581 evidence from two southern cities in Taiwan. *Pathog Glob Health* **117**, 596-604 (2023).
- 582 48. M. A. F. Carneiro, B. Alves, F. S. Gehrke, J. N. Domingues, N. Sá, S. Paixão, J. Figueiredo, A. Ferreira, C.  
583 Almeida, A. Machi, E. Savóia, V. Nascimento, F. Fonseca, Environmental factors can influence dengue  
584 reported cases. *Rev Assoc Med Bras (1992)* **63**, 957-961 (2017).
- 585 49. T. J. Betjeman, S. E. Soghoian, M. P. Foran, mHealth in Sub-Saharan Africa. *International Journal of*  
586 *Telemedicine and Applications* **2013**, 482324 (2013).
- 587 50. D. J. Gubler, Dengue and dengue hemorrhagic fever. *Clin Microbiol Rev* **11**, 480-496 (1998).
- 588 51. H. Tissera, P. Samaraweera, M. de Boer, S. Gandhi, L. Malvaux, S. Mehta, P. Palihawadana, V.  
589 Vantomme, R. Paris, A. Schmidt, The Burden of Acute Febrile Illness Attributable to Dengue Virus  
590 Infection in Sri Lanka: A Single-Center 2-Year Prospective Cohort Study (2016-2019). *Am J Trop Med*  
591 *Hyg* **106**, 160-167 (2021).

- 592 52. S. Richardson, K. L. Dauber-Decker, T. McGinn, D. P. Barnaby, A. Cattamanchi, R. Pekmezaris, Barriers  
593 to the Use of Clinical Decision Support for the Evaluation of Pulmonary Embolism: Qualitative Interview  
594 Study. *JMIR Hum Factors* **8**, e25046 (2021).
- 595 53. T. S. Ho, T. C. Weng, J. D. Wang, H. C. Han, H. C. Cheng, C. C. Yang, C. H. Yu, Y. J. Liu, C. H. Hu, C. Y.  
596 Huang, M. H. Chen, C. C. King, Y. J. Oyang, C. C. Liu, Comparing machine learning with case-control  
597 models to identify confirmed dengue cases. *PLoS Negl Trop Dis* **14**, e0008843 (2020).
- 598 54. A. S. Fathima, D. Manimeglai, Analysis of significant factors for dengue infection prognosis using the  
599 random forest classifier. *Int J Adv Comput Sci Appl* **6**, 240-245 (2015).
- 600 55. D. K. Ming, N. M. Tuan, B. Hernandez, S. Sangkaew, N. L. Vuong, H. Q. Chanh, N. V. Chau, C. P.  
601 Simmons, B. Wills, P. Georgiou, The diagnosis of dengue in patients presenting with acute febrile illness  
602 using supervised machine learning and impact of seasonality. *Frontiers in Digital Health* **4**, 849641  
603 (2022).
- 604  
605

## 606 **Acknowledgments**

607 **Funding:** Research reported in this publication was supported by the United States National Institutes of  
608 Health under award number R01AI135114 (to DTL), K24AI166087 (to DTL), 1R01AI175941 (to KA, HS) and  
609 P01AI034533 (to ALR and KBA), the Military Infectious Disease Research Program (MIDRP), and the European  
610 Research Council (No. 804744, to HS). RJW is funded by the National Institute of Health, through Utah  
611 Stimulating Access to Research in Residency (StARR) under award 1R38HL167282-01.

612 Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its  
613 presentation and/or publication. The opinions or assertions contained herein are the private views of the  
614 author, and are not to be construed as official, or as reflecting true views of the Department of the Army or  
615 the Department of Defense. The investigators have adhered to the policies for protection of human subjects  
616 as prescribed in AR 70-25.

617

Section/Topic	Item	Checklist Item	Page
<b>Title and abstract</b>			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
<b>Introduction</b>			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	3-4
<b>Methods</b>			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	4-5
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	4-5
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	4
	5b	Describe eligibility criteria for participants.	4
	5c	Give details of treatments received, if relevant.	N/A
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	4
	6b	Report any actions to blind assessment of the outcome to be predicted.	N/A
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	4-6
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	N/A
Sample size	8	Explain how the study size was arrived at.	4
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	6
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	6
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	6
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	6
Risk groups	11	Provide details on how risk groups were created, if done.	N/A
<b>Results</b>			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	7
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	7
Model development	14a	Specify the number of participants and outcome events in each analysis.	7
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	N/A
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	9-13
	15b	Explain how to use the prediction model.	9-13
Model performance	16	Report performance measures (with CIs) for the prediction model.	9-13
<b>Discussion</b>			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	15
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	14-15
Implications	20	Discuss the potential clinical use of the model and implications for future research.	14-15
<b>Other information</b>			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	19-22

Funding	22	Give the source of funding and the role of the funders for the present study.	15
---------	----	-------------------------------------------------------------------------------	----

**Supplementary Table S1. TRIPOD checklist.**

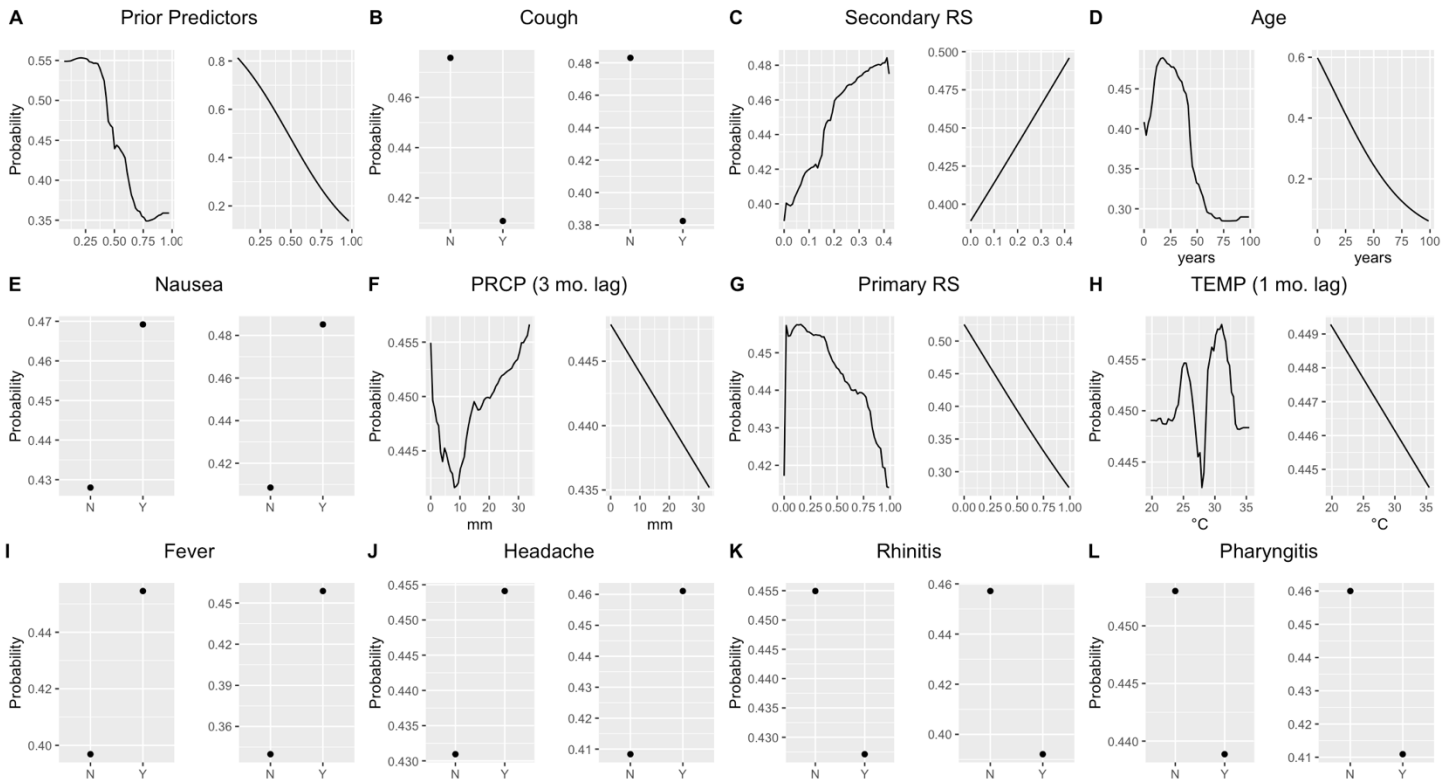
619  
620

Clinical Predictors	DENV Negative	DENV Positive	OR	95% CI
Fever			2.15	1.87-2.47
No	971 (14%)	391 (6.8%)		
Yes	6,127 (86%)	5,337 (93%)		
Nausea			1.66	1.53-1.79
No	4,044 (57%)	2,543 (44%)		
Yes	3,054 (43%)	3,185 (56%)		
Headache			1.53	1.4-1.67
No	2,304 (32%)	1,371 (24%)		
Yes	4,794 (68%)	4,357 (76%)		
Emesis			1.50	1.39-1.62
No	4,126 (58%)	2,754 (48%)		
Yes	2,972 (42%)	2,974 (52%)		
Malaise			1.33	1.23-1.44
No	3,687 (52%)	2,569 (45%)		
Yes	3,411 (48%)	3,159 (55%)		
Anorexia			1.33	1.22-1.44
No	4,611 (65%)	3,340 (58%)		
Yes	2,487 (35%)	2,388 (42%)		
Abdominal Pain			1.27	1.17-1.38
No	4,815 (68%)	3,570 (62%)		
Yes	2,283 (32%)	2,158 (38%)		
Myalgias			1.26	1.16-1.36
No	3,638 (51%)	2,607 (46%)		
Yes	3,460 (49%)	3,121 (54%)		
Chills			1.20	1.11-1.3
No	4,139 (58%)	3,078 (54%)		
Yes	2,959 (42%)	2,650 (46%)		
Retro-orbital Pain			1.16	1.06-1.28
No	5,506 (78%)	4,284 (75%)		
Yes	1,592 (22%)	1,444 (25%)		
Hemorrhage			1.16	1.05-1.29
No	5,951 (84%)	4,677 (82%)		
Yes	1,147 (16%)	1,051 (18%)		
Diarrhea			1.06	0.97-1.16
No	5,401 (76%)	4,297 (75%)		
Yes	1,697 (24%)	1,431 (25%)		
Arthralgias			1.05	0.96-1.15

No	5,198 (73%)	4,134 (72%)		
Yes	1,900 (27%)	1,594 (28%)		
Rash			1.02	0.93-1.12
No	5,507 (78%)	4,426 (77%)		
Yes	1,591 (22%)	1,302 (23%)		
Age			0.98	0.98-0.98
	23 (18)	18 (11)		
Dark Urine			0.86	0.75-1
No	6,502 (92%)	5,306 (93%)		
Yes	596 (8.4%)	422 (7.4%)		
Seizure			0.83	0.69-0.99
No	6,704 (94%)	5,461 (95%)		
Yes	394 (5.6%)	267 (4.7%)		
Abnormal Movement			0.79	0.67-0.93
No	6,622 (93%)	5,423 (95%)		
Yes	476 (6.7%)	305 (5.3%)		
Nuchal Rigidity			0.77	0.63-0.94
No	6,764 (95%)	5,516 (96%)		
Yes	334 (4.7%)	212 (3.7%)		
Pharyngitis			0.76	0.7-0.83
No	4,978 (70%)	4,322 (75%)		
Yes	2,120 (30%)	1,406 (25%)		
Jaundice			0.63	0.51-0.78
No	6,761 (95%)	5,552 (97%)		
Yes	337 (4.7%)	176 (3.1%)		
Cough			0.55	0.51-0.6
No	4,039 (57%)	4,044 (71%)		
Yes	3,059 (43%)	1,684 (29%)		
Rhinitis			0.55	0.49-0.61
No	5,641 (79%)	5,017 (88%)		
Yes	1,457 (21%)	711 (12%)		

**Supplementary Table S2.** The relative frequencies, odds ratios, and confidence intervals for each clinical variable by DENV positivity.

623  
624  
625  
626



**Supplementary Figure S1.** Partial Dependency Plots for the top performing variables for predicting DENV infection by AUC. For each predictor, the graph on the left shows the partial dependency for a random forest model and the partial dependency for a logistic regression model is shown on the right. 'Y' indicates presence of the symptom and 'N' indicates absence of a symptom. 'PRCP' refers to precipitation, 'TEMP' refers to the environmental temperature, 'RS' refers to reconstructed susceptibility estimates.

	Overall, N = 12,826 <sup>1</sup>	0-4 years, N = 954 <sup>1</sup>	5-9 years, N = 2,033 <sup>1</sup>	10-14 years, N = 2,971 <sup>1</sup>	15-19 years, N = 2,271 <sup>1</sup>	20-24 years, N = 1,174 <sup>1</sup>	25-29 years, N = 875 <sup>1</sup>	30-34 years, N = 624 <sup>1</sup>	35-39 years, N = 448 <sup>1</sup>	40+ years, N = 1,476 <sup>1</sup>	p-value <sup>2</sup>
<b>Nausea</b>											<0.001
Y	6,239 (49)	341 (36)	952 (47)	1,515 (51)	1,239 (55)	646 (55)	447 (51)	320 (51)	208 (46)	571 (39)	
<b>Cough</b>											<0.001
Y	4,743 (37)	514 (54)	840 (41)	1,034 (35)	790 (35)	413 (35)	282 (32)	212 (34)	146 (33)	512 (35)	

<sup>1</sup>n (%)

<sup>2</sup>Pearson's Chi-squared test

**Supplementary Table S3.** The relative frequency of the top performing clinical variables stratified by age group. 'Y' indicates presence of the symptom and 'N' indicates absence of a symptom.

Climate Predictors (months lagged)	DENV Negative Mean (sd)	DENV Positive Mean (sd)	OR	95% CI
---------------------------------------	----------------------------	----------------------------	----	--------

DEWPT	23.3°C (2.2)	23.7°C (1.8)	1.10	1.08-1.12
TEMP (1)	28.4°C (1.7)	28.6°C (1.5)	1.08	1.05-1.11
DEWPT (1)	23.3°C (2.3)	23.6°C (1.9)	1.08	1.06-1.1
VISIB	9.3 km (2.4)	9.5 km (2.1)	1.05	1.03-1.06
TEMP	28.3°C (1.6)	28.4°C (1.4)	1.04	1.01-1.06
PRCP	4.9 mm (4.5)	5.4 mm (4.6)	1.02	1.02-1.03
RH	75.3 (9.0)	76.6 (8.1)	1.02	1.01-1.02
RH (3)	70.7 (9.3)	69.8 (8.8)	0.99	0.98-0.99
PRCP (3)	3.7 mm (4.1)	3.2 mm (3.9)	0.97	0.96-0.98
SLP (1)	1008.1 mbar (2.9)	1007.8 mbar (2.6)	0.96	0.94-0.97
SLP	1008.1 mbar (2.9)	1007.7 mbar (2.7)	0.94	0.93-0.96
VISIB (3)	8.3 km (2.8)	7.7 km (2.9)	0.93	0.92-0.94
WDSP	0.6 m/s (0.3)	0.6 m/s (0.3)	0.74	0.65-0.85
WDSP (3)	0.7 m/s (0.3)	0.7 m/s (0.3)	0.73	0.64-0.83

**Supplementary Table S4.** The mean, standard deviation, odds ratio, and 95% CI intervals for each climate predictor.

642  
643