Assessing generalizability of a dengue classifier across multiple datasets

Bingqian Lu[1☯], Yanni Li[1☯], Ciaran Evans[1*]

1 Department of Statistical Sciences, Wake Forest University, Winston-Salem, NC, USA

☯These authors contributed equally to this work.
* evansc@wfu.edu

Abstract

Background

Early diagnosis of dengue fever is important for individual treatment and monitoring disease prevalence in the population. To assist diagnosis, previous studies have proposed classification models to detect dengue from symptoms and clinical measurements. However, there has been little exploration of whether existing models can be used to make predictions for new populations.

Methods

We trained logistic regression models on five publicly available dengue datasets from previous studies, using three explanatory variables identified as important in prior work: age, white blood cell count, and platelet count. These five datasets were collected at different times in different locations, with a variety of disease rates and patient ages. A model was trained on each dataset, and predictive performance was evaluated on both the original (training) dataset, and the other (test) datasets from different studies.

Results

In-sample area under the receiver operating characteristic curve (AUC) values for the logistic regression models ranged from 0.74 to 0.89, while out-of-sample AUCs ranged from 0.55 to 0.89. Matching age ranges in training/test datasets increased AUC values and balanced the sensitivity and specificity. Adjusting the predicted probabilities to account for differences in dengue prevalence improved calibration in 20/28 training-test pairs.

Conclusions

The in-sample performance of the logistic regression model was consistent with previous dengue classifiers, suggesting the chosen model is a good choice in a variety of settings and has decent overall performance. However, adjustments are required to make predictions on new datasets. Practitioners can use existing dengue classifiers in new settings but should be careful with different patient ages and disease rates.

Author summary

Dengue fever is an acute mosquito-borne infection with a substantial and growing disease burden. Early diagnosis of dengue fever is important for treatment and disease monitoring, but gold-standard tests are not always readily available. To supplement diagnostic tests, previous studies have investigated the use of classifiers trained on common diagnostic measurements such as symptoms, blood work, and demographic variables. However, comparisons of existing methods are limited, and in particular there has been little exploration of whether existing models can be used to make predictions for new populations. Generalizability of existing models to new settings could save the substantial time and effort required to collect new data and fit a model, but this generalizability may be limited by differences between populations. In this study, we assess performance of logistic regression models on five publicly available dengue datasets from previous studies, using three explanatory variables identified as important in prior work: age, white blood cell count, and platelet count. Our results show that it can be possible for practitioners to use existing models in new settings, but care is needed to account for differences in patient demographics and dengue prevalence.

Introduction

Dengue fever is an acute mosquito-borne infection worldwide, most common in tropical and subtropical areas, with a substantial and growing disease burden. In countries such as Vietnam, Malaysia, and Colombia, more severe dengue cases were reported in the last decades (1; 2), and the number of dengue cases doubled every 10 years from 1990 to 2013 (3). In 2019, dengue was observed in Afghanistan for the first time (4), and the World Health Organization (WHO) has also reported that dengue has been recorded in more than 100 countries worldwide and is now spreading to Europe (5; 6). Furthermore, accurate estimation of the number of cases per year is complicated by low levels of reporting, a lack of inexpensive diagnostic methods, and incompatible comparative analysis; these issues lead to underestimates of dengue cases, particularly in developing countries, and the annual number of cases may be as high as 390 million (5; 7).

Dengue diagnosis

Dengue is characterized by symptoms such as high fever ($> 40°$C), muscle pain, severe headache, plasma leakage, bleeding diathesis, and skin rash (6). The standard guidelines for dengue diagnosis are given by the WHO (6), but the symptoms used for diagnosis overlap with several other diseases, such as yellow fever, Zika, and West Nile virus. It can therefore be difficult to identify dengue in the first few days of illness (8), and previous research suggests that the diagnostic guidelines need further modifications (9; 10).

Beyond general symptoms, the WHO recommends examining viral RNA and antibodies for dengue detection. Viral RNA detection offers the advantage of providing faster results, typically within 24 to 48 hours. However, viral RNA detection tools require more expensive equipment and complicated processes, which limit their use in many developing countries impacted by dengue fever (11). Furthermore, dengue antibody detection is not suitable for early diagnosis of dengue infection, particularly during primary infections. IgM antibodies are only detectable in approximately 50% of patients between day 3 and 5 from the onset of fever. The percentage of patients with detectable IgM antibodies increases to 80% by day 5 and reaches 99% by day 10. So the performance of antibody detection is the worst in the early stages of the disease (12), which is problematic because early diagnosis of dengue can be critical for timely

treatment. Dengue shock syndrome (DSS), the most common life-threatening complication of dengue caused by a serious loss of blood in the circulatory system, usually onsets between the 4th and 6th day of illness, and the mortality rate can be up to 20% (8; 13). Thus, identifying dengue in the early stage is crucial for impacted countries.

According to the WHO, the gold standard of dengue diagnosis is composed of RT-PCR viral detection, IgM/IgG antibodies and NS1 detection (14). As antigens can be detected earlier, rapid antigen tests are often used together with antibody detection in early diagnosis (15), and the NS1 rapid test is generally considered very sensitive in the early stages of dengue (15). In practice, both NS1 and IgM/IgG antibodies rapid tests are widely used for diagnosis, depending on factors like the patient's clinical presentation, the illness's timing, and the availability of the test (16). In a study comparing the different rapid tests' diagnostic performance, simultaneous NS1 and IgM/IgG antibodies tests gave the best performance with 92% sensitivity and 96% specificity (17). Another comparison study among RT-PCR viral detection, IgM/IgG antibodies, and NS1 detection suggests that viral RNA and NS1 antigen detection were more suitable than IgM antibodies in the early days after symptom onset. However, starting from days 4-5, IgM detection became the most effective diagnostic method. On days 6-7 and in later samples, both NS1 antigen and IgM detection outperformed viral RNA detection as diagnostic markers for dengue infection (18).

Classification methods for diagnosis

While a combination of antigen and antibody tests performs well for detecting dengue, these rapid tests are not always available. To supplement rapid tests, previous studies have investigated the use of classifiers trained on common diagnostic measurements such as symptoms, blood work, and demographic variables. For example, Tuan *et al.* (8), Ngim *et al.* (2), and Andries *et al.* (19) recorded counts of white blood cells, platelets, and lymphocytes. Physical symptoms recorded include rashes, body temperature, bleeding, arthralgia, asthenia, cough, and headache, and demographic variables include sex and age (2; 8; 20; 21).

These symptoms, demographics, and laboratory measurements have been employed by previous studies to predict whether an ill patient actually has dengue. Tuan *et al.* (8) used logistic regression to classify patients, and found through model selection that age, white blood cell count, and platelet count were the most important predictors. Their final model has a sensitivity of 74.8% and a specificity of 76.3% for detecting dengue in children, and their results indicate that combining their classifier with rapid tests can slightly improve the performance of the rapid test. Ngim *et al.* (2) also applied logistic regression, using a penalized model to decrease small-sample bias and improve estimation. Variables were selected using stepwise selection, similar to Tuan *et al.* (8), and the final model had an area under the receiver operating characteristic (ROC) curve (AUC) of 0.82, a sensitivity of 78.4%, and a specificity of 74.6%.

Beyond logistic regression, a variety of other models have been proposed, with performance similar to Tuan *et al.* (8) and Ngim *et al.* (2). For example, Tanner *et al.* (22) employed a C4.5 decision tree classifier, using predictors such as platelet count, white blood cell count, lymphocyte count, body temperature, hematocrit count, and neutrophil count, and reported a specificity of 80.2% and a sensitivity of 78.2%. Park et al. (23) used structural equation models (SEMs), with variables also including white blood cell count, platelet cells count, age, lymphocyte counts, and hematocrit; their final model had an AUC of 0.84 at fever day -3 for dengue prediction.

Limitations of previous studies

In a recent systematic review, Neto *et al.* (24) found limited comparisons of existing methods for dengue classification in the literature, and a dearth of available data from previous studies that would allow researchers to reproduce and compare prior results. Comparisons of previous work are further complicated by the fact that each dataset contains different explanatory variables, so a model from one study may not be usable with new data if one or more variables in the model is not recorded. Additionally, previous studies were conducted in a wide variety of locations and at different times, and it is unclear whether models built in one location and time can generalize to a new setting. To the best of our knowledge, no previous work has assessed the generalizability of dengue classification models.

Contributions

The purpose of this paper is to evaluate whether a dengue classifier trained on one dataset can make useful predictions on a new dataset. Our work focuses on logistic regression models, which have been widely used in previous work, which has also demonstrated their competitive performance compared to other classifiers. We have identified five publicly available datasets from previous dengue studies, which all contain commonly used explanatory variables like age, white blood cell count, and platelet count. To assess generalizability, we conduct an extensive comparison of in-sample and out-of-sample performance for models built on each of the five datasets. We further explore possible reasons a classifier may fail to generalize, and investigate adjustments to make predictions more generalizable.

If dengue classifiers generalize to new data, researchers could adapt existing models for new locations, saving the substantial time and effort required to collect new data and fit a model. If classifiers do not generalize, however, then existing models have limited utility beyond the specific data used to train them.

Methods

Datasets

Data was used from five different papers on predicting dengue, which made their datasets publicly available. Each dataset was collected in a different location and contains varying numbers of patients and features. All raw data can be found in our GitHub repository (`https://github.com/ciaran-evans/dengue-generalizability`).

Dataset 1

Tuan *et al.* (8) collected data on 5726 patients, aged 1–15, who presented with possible dengue fever at seven hospitals in southern Vietnam between 2010 and 2012, with 340 patients from the smallest site and 1589 from the largest site. Of these patients, 1698 (29.7%) were diagnosed with dengue using a combination of RT-PCR, IgM serology, and NS1 detection by ELISA. Patients were eligible for the study if their fever and symptom history was less than 72 hours. Blood samples, personal information, and symptoms were collected from each patient at the time of enrolment. A total of 35 variables were recorded, including white blood cell count, platelet count, temperature, height, and weight, and symptoms such as vomiting, rashes, skin bleeding, and mucosal bleeding. We omitted two patients with missing values for white blood cell count and platelet count.

Dataset 2

Gasem *et al.* (25) collected data on 1486 fever patients (body temperature at least 38°C) aged 1–98 from eight different sites in Indonesia between July 2013 and June 2016, with 65 patients from the smallest site and 267 from the largest site. Of these patients, 467 (31.4%) were diagnosed with dengue. A total of 51 variables were recorded on the symptoms, laboratory results, and final diagnosis for each subject, and the site at which each subject was recorded. Diagnoses included Streptococcus, dengue, Chikungunya, influenza, and RSV, among others. To focus on identifying dengue fever, we grouped all non-dengue diagnoses together as one category for analysis. We omitted one patient with a missing value for platelet count.

Dataset 3

Saito *et al.* (21) collected data on 1573 patients aged 1–85, with a fever lasting at most 21 days and suspected of bloodstream infection, in the Philippines from 2015 to 2019. Of these patients, 256 (16.3%) were diagnosed with dengue. A total of 284 variables were recorded on the symptoms, laboratory results, medical history, and the diagnosis from laboratory tests for each subject. We omitted 21 patients with missing values for platelet count or white blood cell count, leaving 1552 rows in the data.

Dataset 4

Park *et al.* (23) collected data on 257 children aged 6 months to 15 years from Thailand, with fever less than 72 hours. Of these patients, 156 (60.7%) were diagnosed with dengue; of the 156 dengue patients, 51 were diagnosed with dengue hemmorhagic fever, and nine were diagnosed with dengue shock syndrome. A total of 21 variables were recorded, including the patient's age, sex, final diagnosis, and laboratory results. Each of the laboratory results was recorded twice at different time points, one of which was four days before patients had a fever (fever day -3), and the other was two days before patients had a fever (fever day -1). In the dataset, age is recorded as a two-year range (e.g., 12–13); we converted the range to a single numerical value by taking the midpoint of the range (e.g., 12.5).

Dataset 5

Ngim *et al.* (2) collected data on 368 adult patients with possible dengue infection in Malaysia in 2018. Of these patients, 167 (45.4%) were ultimately diagnosed with dengue using a laboratory test. A total of 82 variables were recorded, including dengue status, white blood cell count, platelet count, age, and symptoms such as fatigue, headache, and chills.

Choice of explanatory variables

There are many possible explanatory variables available in each dataset, and most of the available variables do not appear in all five datasets. However, previous research suggests that a small number of features can be used for dengue classification, so to assess generalizability we restricted our attention to some of the most common and widely-used features available across studies.

Using variable selection, Tuan *et al.* simplified a full model containing 17 variables with several interaction terms to a simpler, final model with only three explanatory variables: patient age, white blood cell count, and platelet count (8). Their final model with these three variables had a senstivity of 74.8%, a specificity of 76.3%, and an AUC of 0.829. Similarly, Cavailler *et al.* (26) created a logistic regression model with five

variables: positive tourniquet test, absence of upper respiratory infection symptoms, platelet count, white blood cell count, and liver transaminases (sensitivity: 75.7%, specificity: 65.7%, AUC: 0.71). Tuan *et al.*'s work is also similar to the model created by Ho *et al.* (27), who created a logistic regression model with patient age, temperature, white blood cell count, and platelet count, and reported an AUC of 0.846. Ho *et al.* also compared the performance of their small logistic regression model to a larger model with 18 explanatory variables, and saw a similar AUC of 0.841. Ngim *et al.* (2) also chose a relatively small logistic regression model with seven variables, including white blood cell count and platelet count.

To assess generalizability of dengue classification, we focused on logistic regression with three explanatory variables which are available in all five of the datasets we studied: age, white blood cell count, and platelet count. These explanatory variables were also identified as important dengue predictors by Park *et al.* (23) and Potts *et al.* (28).

Model choice: logistic regression

We chose a logistic regression model to predict dengue status because logistic regression is familiar to many researchers, is relatively easy to interpret, works well with a small number of explanatory variables, and has been widely used in prior studies (2; 8; 26; 27).

Furthermore, logistic regression has competitive performance with other classification methods. For example, Tuan *et al.* compared logistic regression with classification and regression trees (CART) and random forests, and chose logistic regression for their final model (8). Ho *et al.* reported similar results when comparing logistic regression with decision trees and deep neural networks; the respective AUCs of the three methods were 0.797, 0.794, and 0.810 (27). The AUCs for structural equation models (SEMs) reported by Park *et al.* were similar, ranging from 0.73 to 0.94 depending on the time of illness (23).

Logistic regression was fit using R version 4.2.1 (29). All code for model fitting and analysis can be found in our GitHub repository (`https://github.com/ciaran-evans/dengue-generalizability`).

Assessing in-sample performance

Before assessing generalizability of dengue predictions to different datasets, we summarized the in-sample performance of the logistic regression model (with age, white blood cell count, and platelet count as explanatory variables) on each of the five datasets. Summarizing in-sample performance allows us to better understand the expected performance of the model, which helps develop a baseline for comparison across datasets.

In-sample performance was assessed by fitting a logistic regression model on each dataset, and evaluating the model's performance on the same dataset. Cross-validation was used to evaluate performance on new observations from each dataset. If data were collected from multiple sites, cross-validation was performed by holding out each site in turn, as in Tuan *et al.* (8). Otherwise, 10-fold cross-validation was used. Predictive performance was measured using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the ROC curve (AUC); these performance metrics were calculated using the predicted probabilities from cross-validation for each observation.

Sensitivity, specificity, PPV, and NPV all require that the predicted probabilities produced by the logistic regression model are converted into binary predictions, which requires a choice of threshold (AUC, on the other hand, considers *all* possible thresholds for binarization). If a particular threshold was chosen in the original study from which we obtained the data (e.g., Tuan *et al.* (8) used a threshold of 0.33), we used that

threshold for consistency. If no threshold was chosen, we selected the threshold which maximized the geometric mean of sensitivity and specificity (30; 31).

Assessing model generalizability

To assess generalizability of the logistic regression model with age, white blood cell count, and platelet count as explanatory variables, we considered all possible pairs of the available datasets, with one of the datasets as the training data and one of the remaining datasets as the test data. The logistic regression model was fit on the training set, and applied to the test set to calculate a predicted probability of dengue for each patient in the test set. We then calculated AUC, sensitivity, specificity, positive predictive value, and negative predictive value for the test set using these predicted probabilities. For metrics which required a threshold for binary classification, we considered both the thresholds for the training and test data from our assessment of in-sample performance.

In practice, we usually only have the training data to train and assess the model's performance, which means the model would make predictions and calculate the metrics using the threshold from the training data when we apply it to new data. Given information on both the training and test datasets, we examined thresholds from both the training and test datasets when calculating the sensitivity and specificity. Exploring both thresholds provides insight into whether the choice of threshold is generalizable.

Restricting age ranges

It is well known that statistical methods perform poorly when asked to extrapolate to new data beyond the range of the training sample. As age is an explanatory variable in our logistic regression model, and the range of patient ages differs substantially between datasets, we re-assessed model generalizability after restricting the age range in test sets.

When the training set was Dataset 1 or Dataset 4 (age < 16 years), and the test set was Dataset 2 or Dataset 3, we restricted the age range of the test set to also be below 16 years. When the training set was Dataset 5 (age > 16 years), and the test set was Dataset 2 or Dataset 3, we restricted the age range of the test set to also be above 16 ears. We re-calculated all performance metrics for predictions on the age-restricted test sets, and compared the results to predictions on the unrestricted test sets. When re-calculating the performance metric, we only use the threshold from the training data to calculate metrics that require a threshold for binary classification to evaluate whether restricting the age range would be an efficient way to improve the generalizability of the model in practice.

Assessing calibration

Ideally, the predicted probabilities of dengue should be close to the true probabilities in the test data, in which case the model predictions are said to be *calibrated*. We assessed calibration through calibration plots, in which the predicted probabilities for the test data are binned into deciles, and the true rate of dengue cases in each bin is plotted against the bin midpoint. When predictions are calibrated, the points should be close to the diagonal line with intercept 0 and slope 1; deviations from this diagonal indicate mis-calibrated predictions. Mis-calibrated predictions do not necessarily impact AUC (because the AUC is calculated over all possible thresholds, and any monotonic transformation of the predictions will result in the same AUC), but will lead to problems when applying a threshold to binarize predicted probabilities.

Using label shift to adjust classifier predictions

If model predictions fail to generalize, it is because the distribution of data has changed between training and test sets. It is generally impossible to correct the predictions and improve generalizability without making assumptions about the specific nature of this distributional change.

The two common assumptions made about distributional change are *covariate shift* and *label shift*. Covariate shift (32; 33) posits a change in the distribution of the explanatory variables, but no change in the relationship between the explanatory variables and the response. If covariate shift holds, then classifier predictions should generalize without correction, because the training and test sets have the same relationship between predictors and response. Conversely, label shift (34; 35; 36) assumes that the distribution of the response (i.e., dengue prevalence) changes between datasets, but the distributions of age, white blood cell count, and platelet count remain the same conditional on dengue status. Label shift has received substantial attention in the classification and machine learning literature, and previous work has suggested that label shift may be an appropriate assumption for modeling diseases, in which the disease rate can change but the symptoms remain the same (36). Label shift may also be suitable for the datasets considered here, as the proportion of study patients who were diagnosed with dengue ranges from 16% in Dataset 3 to 61% in Dataset 4.

Under label shift, predictions on test data can be calibrated if the prevalence of dengue in the test data is known, without having to re-fit the model on the test data. Let $\widehat{p}_i$ denote the predicted probability of dengue for the ith patient in the test set, using a model fit on a different training set. Let π_{train} denote the marginal prevalence of dengue in the training set, and π_{test} the marginal prevalence of dengue in the test set. If π_{test} is known, then the calibrated predictions $\widehat{p}_{i,\text{calibrated}}$ on the test set can be calculated by applying Bayes' rule:

$$\widehat{p}_{i,\text{calibrated}} = \frac{\frac{\pi_{\text{test}}}{\pi_{\text{train}}}\widehat{p}_i}{\frac{\pi_{\text{test}}}{\pi_{\text{train}}}\widehat{p}_i + \frac{1-\pi_{\text{test}}}{1-\pi_{\text{train}}}(1-\widehat{p}_i)}. \tag{1}$$

After restricting age ranges in the test set as described above, we applied Eq. (1) and assessed performance of the label shift-adjusted predictions with calibration plots. We also compared predictive performance by calculating deviance for the test data ($-2\times$ the binomial log-likelihood) using the uncorrected predictions and the label shift-adjusted predictions.

If the rate of dengue π_{test} in an area is unknown, it may be possible to estimate this rate by leveraging the label shift assumption. Lipton *et al.* (36) proposed a discretization method for estimating test prevalence under label shift, which involves inverting an estimated confusion matrix. An alternative is a fixed point method proposed by Saerens *et al.* (34), selecting the estimate of the test prevalence which most closely matches the mean of the corrected predictions from Eq. (1). We assessed performance of both methods, and compare to the simple estimate of test prevalence which averages the uncorrected predicted probabilities $\widehat{p}_i$ for the test observations.

Results

Feature distributions in each dataset

The distribution of each explanatory variable in each dataset is shown in Fig. 1, which shows that these features share certain similarities in their distributions as differences. Full summary statistics for each variable are provided in S1 Table. The white blood cell count and platelet count in each dataset were scaled to the same measuring unit,

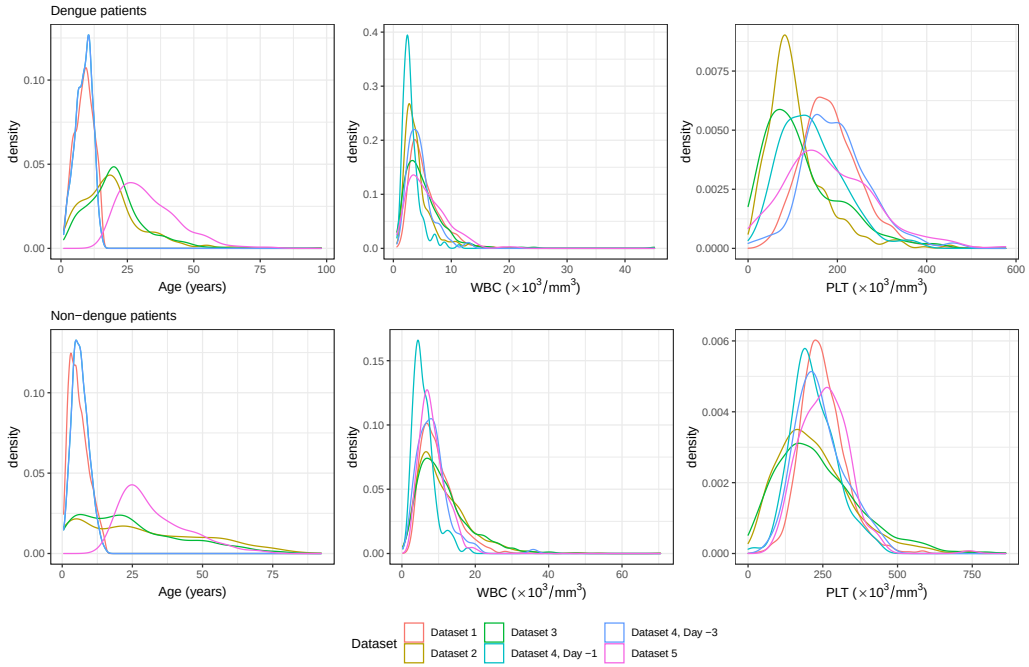

Fig 1. Distributions of explanatory variables across datasets. The distribution of age, white blood cell count (WBC), and platelet count (PLT) for patients in each dataset, shown separately for dengue and non-dengue patients.

$\times 10^3/\text{mm}^3$, for comparison. Despite differences in the range of age, the age distribution was unimodal and right-skewed in almost all datasets, except for Dataset 4, where the distribution of age was unimodal and roughly symmetric. White blood cell count had a right-skewed distribution that centered around $8 \times 10^3/\text{mm}^3$ among all datasets. The distribution of platelet count was also right-skewed, except in Dataset 1 and Dataset 5, in which the distributions of platelet count were unimodal and roughly symmetric. The center of the platelet count was around $200 \times 10^3/\text{mm}^3$ among the five datasets. Lastly, for Dataset 4, which collected the data of white blood cell count and platelet count from both two days and four days before the onset of fever, the distributions of white blood cell count and platelet count collected four days before the fever were more similar to the distributions we observed in other datasets. The mean of platelet count and white blood cell count from data collected two days before the fever are the lowest among the five datasets (supplementary table).

In-sample performance

The in-sample performance of each dataset is summarized in Table 1. The model on Dataset 5 had a sensitivity of 72%, a specificity of 64%, and an AUC of 0.74 (95% CI: (0.69, 0.79)), which is the lowest value among the five datasets and slightly lower than the AUC (0.82) reported by Ngim *et al.* (2). All other datasets had sensitivities and specificities around 75%, and AUCs above 0.82. The AUCs for Dataset 1 (0.83, 95% CI: (0.81, 0.84)) and Dataset 4, Day -3 (0.82, 95% CI: 0.77, 0.88)) were close to the AUCs reported by (8) and (23) (0.83 and 0.84, respectively), while the AUC for Dataset 4, Day -1 (0.86, 95% CI: (0.81, 0.90)) was slightly lower than the value of 0.93 reported by Park *et al.* (23). AUC values were not reported by Gasem *et al.* (25) and Saito *et al.* (21) for Datasets 2 and 3; we found respective AUC values of 0.89 (95% CI: (0.87, 0.91))

and 0.85 (95% CI: (0.83, 0.87)), demonstrating relatively good predictive performance.

Table 1. In-sample performance results for each dataset.

Dataset	Threshold	Sensitivity	Specificity	PPV	NPV	AUC
Dataset 1	0.33	0.74	0.76	0.56	0.87	0.83
Dataset 2	0.45	0.79	0.85	0.70	0.90	0.89
Dataset 3	0.21	0.78	0.78	0.41	0.95	0.85
Dataset 4, Day -3	0.59	0.82	0.70	0.81	0.72	0.82
Dataset 4, Day -1	0.64	0.82	0.77	0.85	0.74	0.86
Dataset 5	0.43	0.72	0.64	0.62	0.74	0.74

Generalizability

Model performance for each training/test pair is shown in Fig. 2, and summarized in S2 Table. As expected, AUC values tended to be slightly lower than in-sample performance when using different training/test sets. However, the differences were mostly small, which supports model generalizability. The exception was training on Datasets 1 and 4, and testing on Datasets 2 and 3; these combinations had substantially lower AUCs on the test data, likely because Datasets 1 and 4 included only pediatric patients, whereas the other datasets included adult patients. The lowest AUC value, 0.547, was observed when the model was trained on Dataset 4 fever day -3 and tested on Dataset 2. However, when we tested the model on Datasets 1 and 4 using Dataset 2 and Dataset 3 as the training data, all AUC values were greater than 0.75. Datasets 2 and 3 include both pediatric and adult patients, and so extrapolation is not required to predict on Datasets 1 and 4. However, models trained on Dataset 5 (age $>$ 16 years) still predicted well on Datasets 1 and 4. This may suggest that datasets with a wider age range can extrapolate to younger patients who are not in the age group more accurately.

There were also several training/test pairs with higher AUCs than the observed in-sample performance. The highest AUC value, 0.894, was observed when the model was trained on Dataset 3 and tested on Dataset 2; this AUC was higher than the in-sample performances of both Dataset 2 and Dataset 3 (Table 1). Similarly, training on Dataset 1 and testing on Dataset 4 fever day -1 also gave an AUC higher than the in-sample performances for both Dataset 1 and Dataset 4. The error bars in Fig. 2 suggest that these differences in AUC are likely due to chance.

While AUC is threshold-independent, sensitivity and specificity are not. Fig. 2 shows that using a pre-specified threshold resulted in substantial imbalance between sensitivity and specificity, suggesting that one probably needs to use an alternative way of choosing the threshold for the model to be more efficient in practice.

Restricting age ranges

Since Dataset 2 and Dataset 3 had wider age ranges than the other datasets, we restricted their age range and re-assessed the model's generalizability when testing on the restricted Dataset 2 and Dataset 3.

As shown in Fig. 3, restricting the age range of the test set to match the training set improved the AUC and often led to a more balanced combination of sensitivity and specificity. An exception was training on Dataset 4 fever day -1 with a threshold of 0.64, and testing on Dataset 3; the AUC improved, but the imbalance between the sensitivity and specificity was not alleviated. Since 0.64 was the largest threshold across the five datasets (and the training threshold for Dataset 3, 0.21, was the smallest), the imbalance after filtering the data may be caused by the great discrepancies between the threshold value and the proportion of dengue patients in Dataset 3.

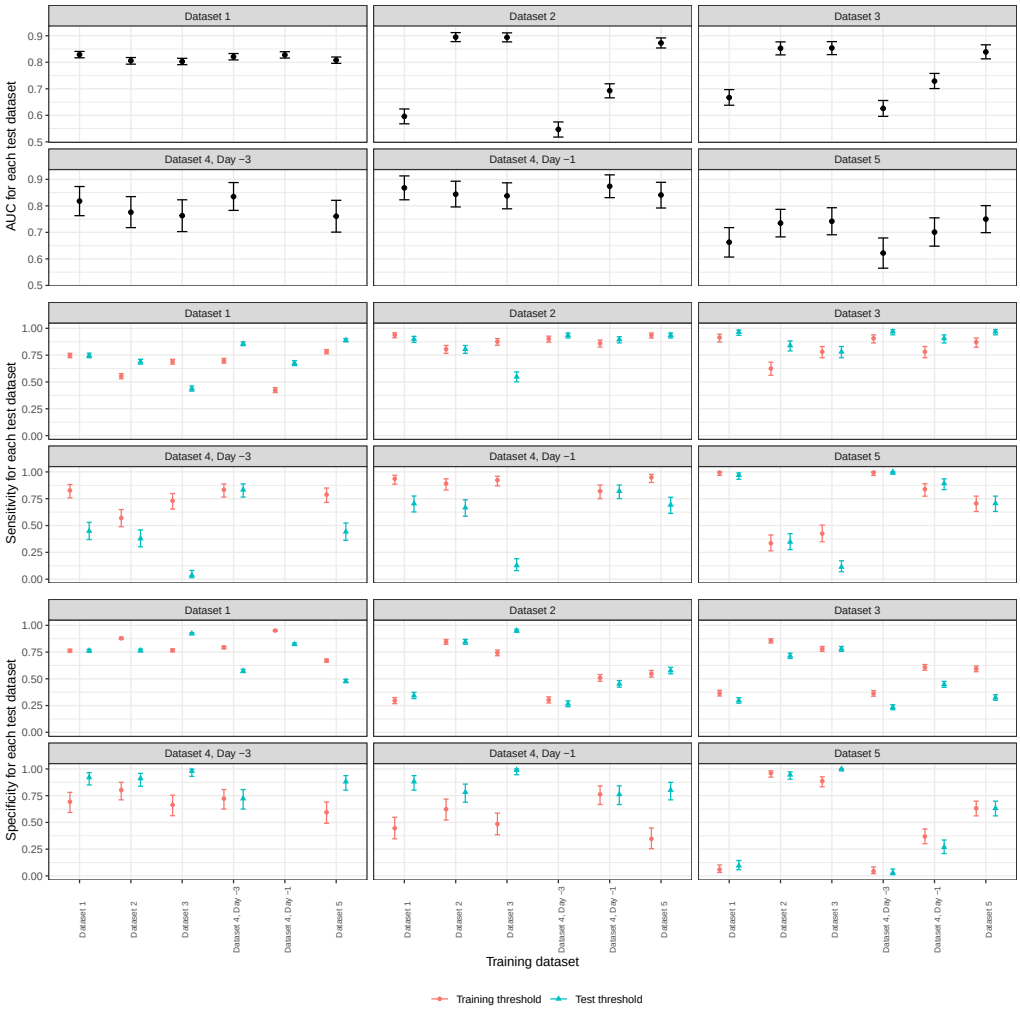

Fig 2. Model performance for each pair of training/test datasets. Each panel represents one test set, with predictive performance on that test set for each training set. Predictive performance is measured by AUC, sensitivity, and specificity. Sensitivity and specificity are both threshold dependent, and so we consider thresholds from both the training and test data (Table 1.)

The model trained on Dataset 5 and tested on restricted Dataset 2 and Dataset 3 showed a similar performance with improved AUC values but still imbalanced sensitivity and specificity, which may be due to the differences between the dengue proportion and the value of the chosen threshold as well.

Calibration and label shift

Several calibration plots for different training/test pairs are shown in Fig. 4 (calibration plots for all pairs are shown in S4 Fig and S5 Fig). The uncorrected predicted probabilities (applying the logistic regression model, fit on the training data, directly to the test data) tended to be poorly calibrated, particularly when the prevalence of dengue differed between training and test sets. Adjusting the predicted probabilities with the dengue prevalence in the test data (Eq. (1)) often improved calibration, as

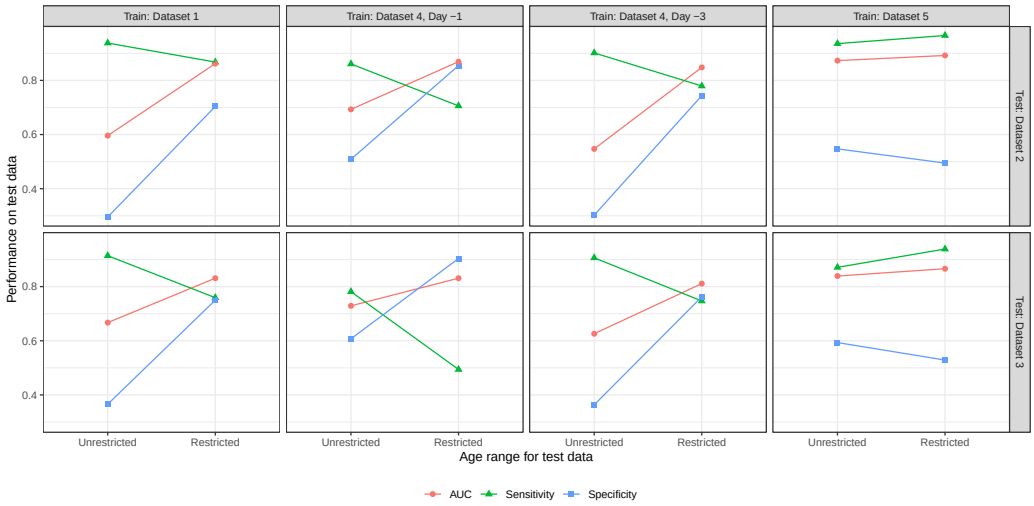

Fig 3. Model performance with restricted age ranges. Changes in AUC, sensitivity, and specificity are shown when restricting the age range of test data from Dataset 2 and Dataset 3. The Unrestricted performance metrics are calculated on the full test sets, while the Restricted performance metrics are calculated to match the age range in the training data (< 16 for Datasets 1 and 4, > 16 for Dataset 5). For sensitivity and specificity, the threshold for each training set (Table 1) was used.

shown in Fig. 4. The deviance for the test data decreased in 20 out of 28 training/test pairs when using these label shift-adjusted predictions (S3 Table).

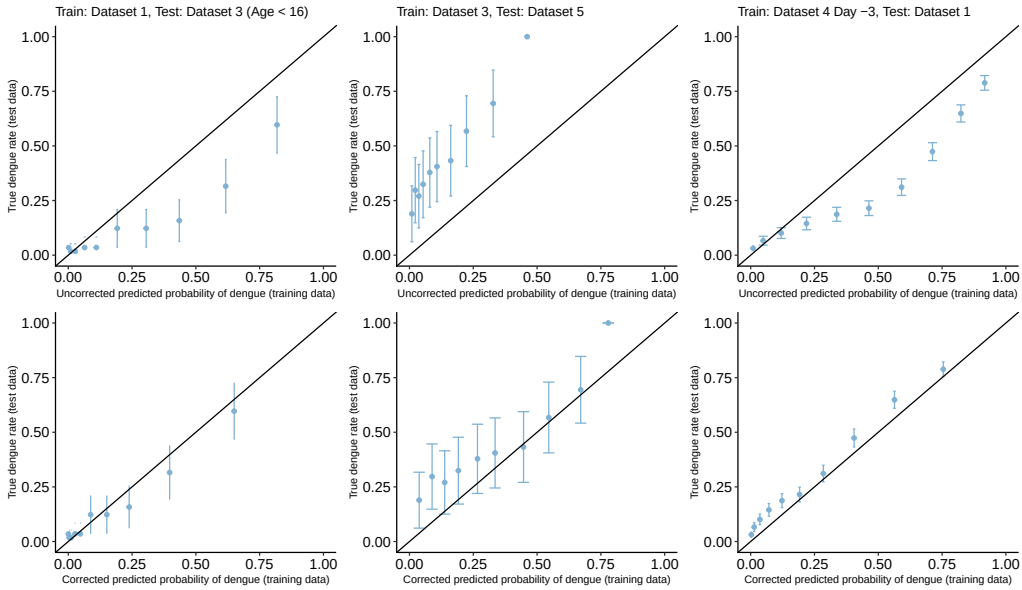

Fig 4. Example calibration plots for several training/test pairs. The logistic regression model is fit on the training set, then applied to the test set to calculate a predicted probability for each test observation. These predicted probabilities are binned into deciles, and the true rate of dengue cases in each bin is plotted against the bin midpoint. Plots are shown for both the raw predictions, and for the label shift-adjusted predictions using Eq. (1).

Estimating dengue prevalence in test data

The corrected calibration plots in Fig. 4 require the true dengue prevalence in each test set. Fig. 5 shows the results of three different estimates of dengue prevalence in the test data, for each training/test pair. The discretization method (36) performed worst, with several estimates outside the (0, 1) range, and a median log ratio (i.e., log(estimated prevalence/true prevalence)) of 0.244 compared to the true dengue prevalence. Performance of the fixed point method (34) was also poor, with a median log ratio of 0.184 compared to the true prevalence. Somewhat surprisingly, simply estimating prevalence with the mean of the uncorrected predicted probabilities performed best, with a median log ratio of 0.016. However, none of the three methods is consistently reliable, with particularly poor performance on test Datasets 4 and 5.

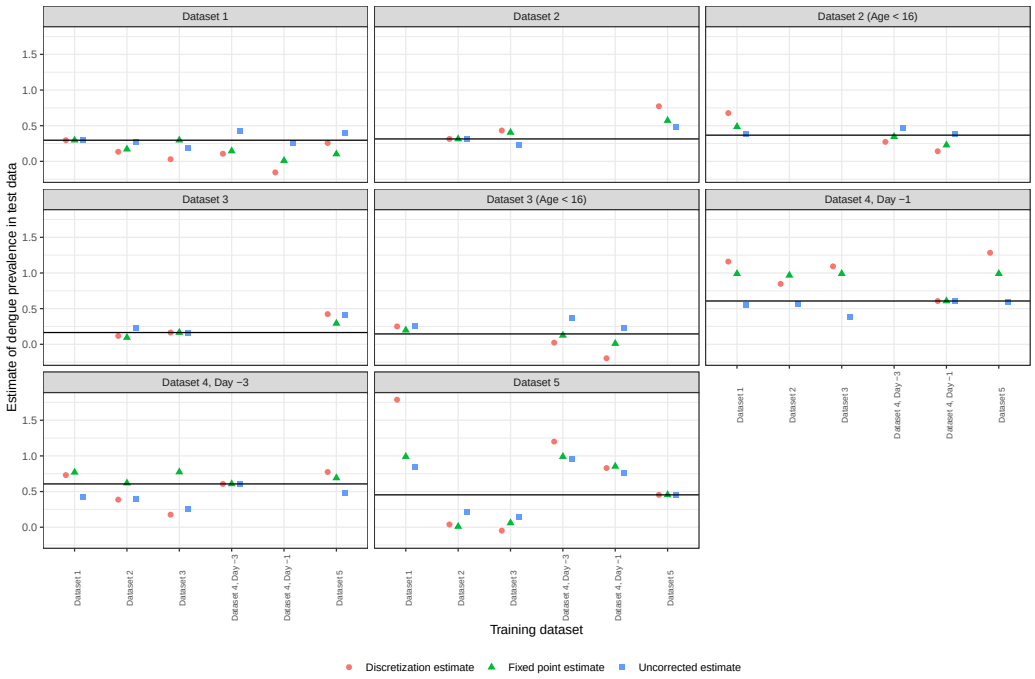

Fig 5. Estimating prevalence of dengue in test data. For each training/test pair, the prevalence of dengue in the test data is estimated by the discretization method, the fixed point method, and by simply averaging the predicted probabilities. The solid horizontal line in each panel represents the true prevalence of dengue in that test set.

Discussion

Many previous studies have investigated the use of statistical models to diagnose dengue in patients. When these models work well, they have the potential to quickly identify dengue fever without requiring potentially expensive or slow laboratory tests. However, each model has been developed for a different population, and collecting training data and fitting an appropriate model can be laborious and time-consuming tasks. If existing models can generalize to new populations, practitioners could make quick diagnoses without needing to create their own new model.

The goal of this paper is to assess the generalizability of a dengue classifier to new populations. We have focused on the Early Dengue Classifier proposed by Tuan *et al.* (8) as a good candidate, because it is easy for practitioners to use, it does not involve

too many explanatory variables (and Tuan *et al.* demonstrated similar performance to more complicated models), and the three covariates have been identified as important dengue predictors in multiple other studies. Our analysis explores whether the model performs well when trained on a new dataset (in-sample performance), and whether a pre-trained model can predict well on new test data from a different population (out-of-sample performance).

Our in-sample results in Table 1 show that, when re-trained on data from new settings, this logistic regression model performs well in a variety of settings with different age ranges and dengue prevalences. Indeed, the in-sample AUCs for each of the five datasets are above 0.75, and are between 0.8 and 0.9 for four of the five datasets. The similarities in AUCs across datasets suggest that the chosen logistic regression model is an effective choice for dengue classification in a wide variety of settings, and could be adapted to new environments by collecting new training sets. This performance is also similar to the AUCs reported in many previous studies using a range of classification techniques (8; 23; 26; 27), which supports that the model is competitive with other classifiers.

After confirming that the logistic regression model is a good choice across datasets, we next assessed whether the model could be applied to data from new settings *without* retraining. If the model were generalizable, we would expect its performance on a new dataset to be similar to the in-sample performance we observe. However, there are several potential challenges when applying an existing model to new data: differences in feature distributions (e.g., different age groups), different rates of dengue prevalence, and selecting an appropriate threshold for binary classification.

When we assessed the generalizability of the model with no restriction on the age range of the datasets, the AUC values for the training/test pairs that have a similar or the training dataset has a wider age range were all above 0.75, though most of the AUC values were slightly lower than the in-sample performances, as expected. For training/test pairs with training datasets that have a narrower age range, we observed substantial declines in the AUC values compared with their in-sample performances (S2 Table). The disparities in the generalizing performance demonstrate that the model trained on a dataset with a broader age range can extrapolate well when tested on a dataset with a narrower age range but not vice versa.

When we re-assessed the generalizability of the model with restrictions on the age range of the datasets, there were improvements shown in AUC values for all datasets (Fig. 3), which means that having similar age ranges in both training and test datasets can improve the generalizability of the model. Since the model has an AUC value above 0.75 and is similar to the in-sample performance for all training/testing pairs with similar age ranges, the logistic regression model we have studied in this paper can be applied to a new dataset with minimal reduction in its overall performance (as measured by AUC, which summarizes the overall ability to distinguish dengue and non-dengue cases) if the model is trained on a similar age range with the new dataset.

However, AUC does not capture whether predicted probabilities are calibrated, nor how to choose a threshold when making binary predictions. In the assessment and re-assessment of the generalizability, we observed an imbalance between sensitivity and specificity (S2 Table). When we evaluated the model with no restrictions on the age range, all training/test pairs have shown a great imbalance between sensitivity and specificity for both the thresholds from the training and testing datasets. The imbalance is somewhat alleviated when we re-evaluated the model with restricted age ranges using only the threshold from the training datasets. However, the discrepancy between sensitivity and specificity still existed for Dataset 4 Day -1 and Dataset 5 (Fig. 3). Thus, although restricting the age range can mitigate the issue of imbalanced sensitivity and specificity, it is only to an extent. Since having a model with imbalanced sensitivity

and specificity may not be helpful for practitioners making diagnoses, one would need to find an alternative way of choosing a threshold for the model to be effective in practice.

Additionally, although the AUC value of the model appears to be relatively good, the predictions were consistently higher or lower than the desired probabilities (Fig. 4) due to the discrepancies in dengue prevalence between the training and testing datasets, which is expected in practice. This suggests that an existing model might not be applicable to new data, without adjusting for differences in dengue prevalence.

The label shift assumption (34; 36) provides a natural method to account for changes in disease prevalence, and the datasets have roughly similar distributions of white blood cell count and platelet count, conditional on dengue status (Fig. 1). Furthermore, in many of the training/test pairs, the predicted probabilities could be calibrated using a label shift assumption (Fig. 4, and S4 Fig and S5 Fig). However, label shift adjustment requires knowing the dengue prevalence in the test set. If the population prevalence is unknown, or changes (e.g. due to seasonal trends), calibrating the model predictions would be impossible without collecting new data. Attempts to leverage the label shift assumption to estimate test prevalence were unsuccessful (Fig. 5).

In conclusion, we have shown that the logistic regression model from Tuan et al. can be applied to a new dataset without re-training, but its ability to help clinicians make diagnoses is limited. Specifically, for the model to generalize to a new dataset, the age range of the training and test datasets should be roughly the same, or the age range of the training dataset should be wider than the test dataset. Although having the age range restricted can ensure the model's general predictive ability, it may still not be very suitable to use the model in practice since it is hard to find an appropriate threshold value to generate a relatively balanced sensitivity and specificity. In future studies, to improve the model's efficiency in clinical settings, one could focus on investigating how to choose a threshold that can result in a balanced sensitivity and specificity. Additionally, since the predicted probabilities made by the model tend to be biased, and the technique required to calibrate it might be too complicated to exert in practice, one could also focus on improving the model's performance by adjusting the distribution of the explanatory variables before training or the model so that its predictive probabilities match with the desired probabilities without calibration.

Supporting information

S1 Table. Summary statistics for each dataset.

S2 Table. Full performance metrics for each combination of training/test datasets.

S3 Table. Label shift adjustment and estimation for each training/test pair.

S4 Fig. Calibration plots for all training/test pairs (uncorrected predictions). The logistic regression model is fit on the training set, and applied to the test set to calculate predicted probabilities. These raw predicted probabilities are compared against true instances of dengue.

S5 Fig. Calibration plots for all training/test pairs (label shift-adjusted predictions). The logistic regression model is fit on the training set, and applied to the test set to calculate predicted probabilities. The predicted probabilities are adjusted with a label shift correction to account for a different rate of dengue fever in the test dataset (see Eq. 1 in the manuscript).

References

1. WHO Regional Office for South-East Asia. Comprehensive guideline for prevention and control of dengue and dengue haemorrhagic fever. 2011;.
2. Ngim CF, Husain SMT, Hassan SS, Dhanoa A, Ahmad SAA, Mariapun J, et al. Rapid testing requires clinical evaluation for accurate diagnosis of dengue disease: A passive surveillance study in Southern Malaysia. PLOS Neglected Tropical Diseases. 2021;15(5):e0009445.
3. Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, et al. The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. The Lancet infectious diseases. 2016;16(6):712–723.
4. Sahak MN. Dengue fever as an emerging disease in Afghanistan: Epidemiology of the first reported cases. International Journal of Infectious Diseases. 2020;99:23–27.
5. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013;496(7446):504–507.
6. World Health Organization. Dengue and severe dengue;. `https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue`.
7. Wilder-Smith A, Byass P. The elusive global burden of dengue. The Lancet Infectious Diseases. 2016;16(6):629–631.
8. Tuan NM, Nhan HT, Chau NVV, Hung NT, Tuan HM, Tram TV, et al. Sensitivity and specificity of a novel classifier for the early diagnosis of dengue. PLOS Neglected Tropical Diseases. 2015;9(4):e0003638.
9. Hadinegoro SRS. The revised WHO dengue case classification: does the system need to be modified? Paediatrics and International Child Health. 2012;32(sup1):33–38.
10. Horstick O, Martinez E, Guzman MG, Martin JLS, Ranzinger SR. WHO dengue case classification 2009 and its usefulness in practice: an expert consensus in the Americas. Pathogens and Global Health. 2015;109(1):19–25.
11. Guzmán MG, Kourı G. Dengue diagnosis, advances and challenges. International journal of infectious diseases. 2004;8(2):69–80.
12. Casenghi M, Kosack C, Li R, Bastard M, Ford N, Group CID. NS1 antigen detecting assays for diagnosing acute dengue infection in people living in or returning from endemic countries. Cochrane Database of Systematic Reviews. 1996;2014(6).
13. Guzman MG, Harris E. Dengue. The Lancet. 2015;385(9966):453–465.
14. Organization WH, for Research SP, in Tropical Diseases T, of Control of Neglected Tropical Diseases WHOD, Epidemic WHO, Alert P. Dengue: guidelines for diagnosis, treatment, prevention and control. World Health Organization; 2009.
15. Centers for Disease Control and Prevention. Dengue Virus Antigen Detection; 2019. `https://www.cdc.gov/dengue/healthcare-providers/testing/antigen-detection.html`.

16. Luo R, Fongwen N, Kelly-Cirino C, Harris E, Wilder-Smith A, Peeling R. Rapid diagnostic tests for determining dengue serostatus: a systematic review and key informant interviews. Clinical Microbiology and Infection. 2019;25(6):659–666.

17. Mata VE, Andrade CAFd, Passos SRL, Hökerberg YHM, Fukuoka LVB, Silva SAd. Rapid immunochromatographic tests for the diagnosis of dengue: a systematic review and meta-analysis. Cadernos de Saúde Pública. 2020;36:e00225618.

18. Huhtamo E, Hasu E, Uzcátegui NY, Erra E, Nikkari S, Kantele A, et al. Early diagnosis of dengue in travelers: comparison of a novel real-time RT-PCR, NS1 antigen detection and serology. Journal of clinical virology. 2010;47(1):49–53.

19. Andries AC, Duong V, Ly S, Cappelle J, Kim KS, Lorn Try P, et al. Value of routine dengue diagnostic tests in urine and saliva specimens. PLOS Neglected Tropical Diseases. 2015;9(9):e0004100.

20. Utama IMS, Lukman N, Sukmawati DD, Alisjahbana B, Alam A, Murniati D, et al. Dengue viral infection in Indonesia: Epidemiology, diagnostic challenges, and mutations from an observational cohort study. PLOS Neglected Tropical Diseases. 2019;13(10):e0007785.

21. Saito N, Solante RM, Guzman FD, Telan EO, Umipig DV, Calayo JP, et al. A prospective observational study of community-acquired bacterial bloodstream infections in Metro Manila, the Philippines. PLOS Neglected Tropical Diseases. 2022;16(5):e0010414.

22. Tanner L, Schreiber M, Low JG, Ong A, Tolfvenstam T, Lai YL, et al. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PLOS Neglected Tropical Diseases. 2008;2(3):e196.

23. Park S, Srikiatkhachorn A, Kalayanarooj S, Macareo L, Green S, Friedman JF, et al. Use of structural equation models to predict dengue illness phenotype. PLOS Neglected Tropical Diseases. 2018;12(10):e0006799.

24. da Silva Neto SR, Tabosa Oliveira T, Teixeira IV, Aguiar de Oliveira SB, Souza Sampaio V, Lynn T, et al. Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: a systematic review. PLOS Neglected Tropical Diseases. 2022;16(1):e0010061.

25. Gasem MH, Kosasih H, Tjitra E, Alisjahbana B, Karyana M, Lokida D, et al. An observational prospective cohort study of the epidemiology of hospitalized patients with acute febrile illness in Indonesia. PLOS Neglected Tropical Diseases. 2020;14(1):e0007927.

26. Cavailler P, Tarantola A, Leo YS, Lover AA, Rachline A, Duch M, et al. Early diagnosis of dengue disease severity in a resource-limited Asian country. BMC Infectious Diseases. 2016;16(1):1–8.

27. Ho TS, Weng TC, Wang JD, Han HC, Cheng HC, Yang CC, et al. Comparing machine learning with case-control models to identify confirmed dengue cases. PLOS Neglected Tropical Diseases. 2020;14(11):e0008843.

28. Potts JA, Gibbons RV, Rothman AL, Srikiatkhachorn A, Thomas SJ, Supradish Po, et al. Prediction of dengue disease severity among pediatric Thai patients using early clinical laboratory indicators. PLOS neglected tropical diseases. 2010;4(8):e769.

29. R Core Team. R: A Language and Environment for Statistical Computing; 2022. Available from: `https://www.R-project.org/`.

30. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. In: ICML. vol. 97. Citeseer; 1997. p. 179.

31. Barandela R, Sánchez JS, Garcıa V, Rangel E. Strategies for learning in class imbalance problems. Pattern Recognition. 2003;36(3):849–851.

32. Bickel S, Brückner M, Scheffer T. Discriminative learning under covariate shift. Journal of Machine Learning Research. 2009;10(9).

33. Gretton A, Smola A, Huang J, Schmittfull M, Borgwardt K, Schölkopf B. Covariate shift by kernel mean matching. Dataset shift in machine learning. 2009;3(4):5.

34. Saerens M, Latinne P, Decaestecker C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural Computation. 2002;14(1):21–41.

35. Storkey A, et al. When training and test sets are different: characterizing learning transfer. Dataset Shift in Machine Learning. 2009;30:3–28.

36. Lipton Z, Wang YX, Smola A. Detecting and correcting for label shift with black box predictors. In: International Conference on Machine Learning. PMLR; 2018. p. 3122–3130.