Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: Patrick Corbett Date: 05/02/2021

Approval Sheet

Application and Interpretation of Dengue Fever Diagnostic and Prognostic Modeling in Yucatan, Mexico With Random Forest and Logistic Regression

By

Patrick Corbett

Master of Science in Public Health

Global Epidemiology Program

_____

Gonzalo [Committee Chair]

_____

Yamila Romer; MD, (Committee Member)

Abstract Cover Page

Application and Interpretation of Dengue Fever Diagnostic and Prognostic Modeling in Yucatan, Mexico With Random Forest and Logistic Regression

By

Patrick Corbett

Bachelor of Arts

University of Virginia

2016

Thesis Committee Chair: Dr. Gonzalo Vazques-Prokopec, PHD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirement for the degree of

Master of Science in Public Health

in the Global Epidemiology Program

2021

Abstract

Application and Interpretation of Dengue Fever Diagnostic and Prognostic Modeling in Yucatan, Mexico With Random Forest and Logistic Regression

By Patrick Corbett

Early identification of patients with dengue and patients at risk of progression into severe forms of disease is important in timely application of potentially life-saving therapies. The current WHO 2009 clinical classification is highly sensitive for detecting dengue but is not very specific which could lead to oversaturation of hospitals in outbreak scenarios. Machine learning methods such as random forest have the potential to supplement clinical classification systems for detecting dengue. In this paper, we apply both random forest and logistic regression for diagnostic and prognostic modeling of dengue disease in Yucatan, Mexico. We also describe the study population with a short descriptive analysis of the demographic and geospatial characteristics within Yucatan, Mexico. Our results indicate that both models perform relatively well when modeling severe dengue versus non-severe dengue as well as severe dengue versus all other febrile illnesses, but they do not perform well when modeling dengue versus other febrile illnesses. We found that logistic regression performed slightly better than random forest for the severe model groups, but results were mixed for the dengue versus non-dengue models. Furthermore, we found that when applying our severe vs non-severe model to novel years and to other Mexican states, model performance decreases thus challenging the applicability of the model in external populations. We conclude with a discussion on the potential applications and interpretability of random forest models in the clinical setting.

Cover Page

Application and Interpretation of Dengue Fever Diagnostic and Prognostic Modeling in Yucatan, Mexico With Random Forest and Logistic Regression

By

Patrick Corbett

Bachelor of Arts

University of Virginia

2012

Thesis Committee Chair: Dr. Gonzalo Vazques-Prokopec, PHD

A thesis submitted to the Faculty of the Rollins School of Public Health of Emory University in partial fulfillment of the requirement for the degree of Master of Science in Publish Health in the Global Epidemiology Program

2021

Table of Contents

Introduction

Background on the Disease

Dengue is an acute arthropod-borne disease caused by infection with Dengue virus (DENV), a positive-strand RNA virus with four different serotypes (DENV 1, DENV 2, DENV 3, DENV 4) that is part of the *Flavivirus* genus along with Yellow Fever, West Nile, and Zika (Fares, 2015; Simmons, 2012; Levi, 2015). Human infection results from the bite of an infected female *Aedes aegypti* or, less commonly, *Aedes albopictus*, mosquito with varied results from an asymptomatic infection to a flu-like febrile illness characterized by rapid onset of retro-orbital pain, myalgia, headache, and fatigue, to a more severe form with plasma leakage, shock, and hemorrhagic symptoms with a fatality rate of 4% (Fares, 2015; Dick, 2015). Two opposite albeit probably complementary theories explain dengue pathogenicity. One theory, based on the observance of relatively high severe case prevalence among populations with novel dengue exposure, states that differences in virus virulence are responsible for the broad spectrum of clinical outcomes. (Barnes, 1974; Gubler, 1978; Rosen, 1977,1986). The other theory explains severity as a result of sequential infection with a different serotype and an immunopathogenic role of the specific immune response to the first serotype which immune enhances the second infection (Halstead, 1970, 1988.)

There is currently no effective anti-viral therapy for dengue and existing vaccines are often not considered safe for populations with multiple dengue serotypes (Fares, 2015). However, dengue symptomatic treatment is highly effective in alleviating and preventing the progression to severe forms, which requires intensive care management and has a high mortality rate (Fares, 2015). As such, early detection of patients at risk for developing severe dengue is important for appropriate application of supportive therapies that curtail progression into severe disease (Deen, 2006; Cucunawangsih, 2015; Ho, 2020).

Epidemiology and Burden of Disease

In 2012 the WHO declared that dengue was the most prevalent and fastest spreading mosquito-transmitted disease in the world with annual incidence estimates ranging from 9 million (WHO, 2004),

50-100 million (WHO, 2012), and 96 million (Bhatt, 2013) across more than 100 endemic countries. One of the highest endemic regions for dengue is the Americas. The Pan American Health Organization (PAHO, 2020) reported that there were around 5.5 million cases from 2019-2020, 34,094 of which were severe and 2,785 of which resulted in death (PAHO, 2020). Of particular interest, Mexico was responsible for around 7% of these cases in 2019-2020, second only to Brazil in the Americas (PAHO, 2020). Moreover, Mexico accounted for around 13.5% of severe cases and 16% of deaths in the Americas in 2019-2020 (PAHO, 2020). Moreover, the economic burden in Mexico from 2010-2011 was around $170 million (95% CL: 151-292), equivalent to $1.56 per capita [95% CI 1.38, 2.68] (Undurraga, 2016). Table 1 illustrates the number and severity of cases, the number of hospitalizations, the number of deaths and serotype samples among confirmed dengue patients in Mexico from 2017-2019 based on data provided by PAHO's Plataforma de Información en Salud para las Américas -PLISA- (Health Information Platform for the Americas). The Mexican state of Yucatan is of specific interest in this study. PAHO's PLISA data indicates that confirmed incidence in Yucatan, Mexico increased by more than 1,500% from 41 in 2018 to 797 in 2019. The data also indicates a recent increase in severe cases compare to previous years.

Table 1: Cases, Hospitalizations, and Deaths Across Years in Mexico

Metrics	2017	2018	2019	2020	2021*
NSD	11571	8675	28408	19014	165
DWS	2562	3836	11116	4083	109
SD	386	919	3506	1055	25
Zika	125	192	318	74	2
Chikungunya	53	61	115	14	0
Mean Age	28.95	24.15	27.9	29.5	22.14
Hospitalizations	3261	5074	15227	5596	147
Deaths	217	77	396	142	1
Dengue Serotypes	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3,4

Data from overall Mexico dataset provided by PAHO's PLISA. NSD=non-severe dengue, DWS=dengue with warning signs, SD=severe dengue. Serotype data derived from small PCR samples in patient data. DENV1 was more prevalent in 2007. DENV2 was predominant in 2020-2021. DENV1 and DENV2 were about equally present in 2018-2019. * 2021 refers to January 1st to April 25th when the data was last retrieved from PAHO's site.

Based on results from published literature, Table 2 illustrates findings regarding the sensitivity, specificity, and other related metrics with respect to the two WHO classification systems in detecting dengue and severe forms of dengue. The 2009 WHO classification shows a higher sensitivity than the 1997 WHO classification in detecting patients at risk of severe outcome, though the specificity for both models was poor. Multiple researchers have expressed concern that the high sensitivity and overall lack of specificity within the revised WHO standard could and has led to a large influx of patients that can overwhelm healthcare system (Srikiatkhachorn, 2011; Tamibmaniam, 2016; Narvaez, 2011; Ho, 2020). While high sensitivity is mandatory to reduce morbidity and mortality rates, improving specificity would help prevent hospital overload during outbreaks, and would help preserve vital attention and resources (Srikiatkhachorn, 2011; Tamibmaniam, 2016; Narvaez; 2011; Ho, 2020).

Background on WHO and Modeling Classification

Given the high burden of disease and given how successful early application of supportive therapy is in preventing development of severe dengue in patients at risk, it is imperative to bolster our capacity for early detection of dengue disease (Ho, 2009). The dengue case classification proposed by the WHO in 1977 categorized dengue into dengue fever (DF), dengue hemorrhagic fever (DHF), and dengue shock syndrome (DSS) (WHO, 1997, 2009; Deen, 2006). The 1997 classification further defined four grades (I-IV) of DSS based on the severity of hemodynamic fail (WHO, 1997). However, this classification did not relate well with disease severity which led to many instances where patients with severe forms of dengue, even some that died from dengue, did not meet the strict 1997 WHO criteria for DHF (Sumarmo, 1983; Deen, 2006). To meet the demands for improved prognostic capabilities and early treatment, the WHO proposed a new classification system in 2009 which included dengue without

warning signs, dengue with warning signs, and severe dengue (WHO, 2009; Hadinegoro, 2012). Warning signs are used to triage patients at risk of severe outcomes before that event occurs (WHO, 2009).

Table 2: WHO Classification Validation Study Results

Comparison Metrics	**Dengue Versus Non-Dengue Illness**						**Severe Dengue Versus Non-Severe Dengue**	
Outcome	Dengue: Old Class[1]	Dengue: New Class[1]	Dengue: Old Class[2]	Dengue: New Class[2]	Dengue: New Class[3]	Dengue: Old Class[3]	Severity-Measure: Old Class[4]	Severity-Measure: New Class[4]
Sensitivity(%)	89.3 (86.2–91.9)	86.6 (83.2–89.5)	96.7 (95.1–97.9)	99.3 (99.2–100.0)	95.4 (90.9–98.2)	79.9 (72.7–85.9)	39 (31.8-46.6)	92.1 (87.1–95.6)
Specificity(%)	43.1 (41.3–44.9)	55.2 (53.4–57.0)	22.0 (18.2–26.1)	8.5 (6.0–11.5)	36.0 (29.4–43.1)	57.0 (49.8–64.0)	75.5 (70.7–79.8)	78.5 (73.9–82.6)
Positive Predictive Value (PPV) (%)	20.3 (18.6–22.1)	23.9 (21.9–26.0)	67.2 (64.3–70.1)	64.2 (61.3–67.0)	--	--	43.4 (35.6–51.5)	67.4 (61.1–73.2)
Negative Predictive Value (NPV) (%)	96.1 (94.9–97.1)	96.2 (95.2–97.1)	80.0 (71.7–86.7)	88.1 (74.4–96.0)	--	--	71.9 (67.2–76.4)	95.4 (92.3–97.4)

(1) Predicting dengue diagnosis for Children >=14 years old presenting to health centers within 6 days of onset for febrile illness or as suspected dengue cases using clinical and laboratory data (Gutiérrez, 2013)
(2) Similar approach to 1. for predicting dengue diagnosis, but for children in hospitals in a hospital setting with febrile illness or suspect of dengue using clinical and laboratory data (Gutiérrez, 2013)
(3) Predicting dengue diagnosis for adult (>=18 years old) patients presenting to clinics in Singapore within 72 hours of febrile illness using clinical data (Shera, 2011)
(4) Predicting dengue severity for Children >=14 years in Managua, Nicaragua in the hospital setting with febrile illness or suspected of dengue using clinical and laboratory data (Federico, 2011)

Researchers have sought to address the need for continued improvement over early dengue detection through applied modeling and machine learning techniques. Table 3 highlights results from a literature review on models regarding dengue versus other febrile illness (OFI) and severe dengue versus

non-severe dengue (Cotterman, 2015). From our literature review, it appears that logistic regression and decision tree are common modeling techniques for clinical diagnostic and prognostic classification. These kinds of models can be appealing due to the relatively simple and interpretable nature of the algorithms (Ho, 2020).

Table 3: WHO Classification Validation Study Results

	Dengue Versus Not-Dengue					Severe Versus Non-Severe Dengue			
Metrics	LR: Dengue Vs OFI (Prior of 0.388)[1]	LR: Dengue Vs OFI (Prior of 0.636)[1]	DT: DF Vs OFI[2]	DT: DF Vs OFI: Hospital A[3]	DT: DF Vs OFI: Hospital B[3]	DT: Severe Vs All Other: Hospital A[4]	DT: Severe Vs All Other: Hospital B[4]	DT: Severe Vs NSD[6]	DT: Severe Vs NSD[6]
Sensitivity(%)	90.1	66.3	71.2	81.6	89.2	83.3	83	78.2	81
Specificity(%)	63.6	80.5	90.1	90.9	88.4	76.7	82.3	80.2	54
Accuracy(%)	79.6	72	84.3	84.6	88.7	78.1	82.4	79.5	57
NPV(%)	81	61.3	--	70.4	90.1	94.4	96.6	--	96
PPV(%)	78.9	83.7	--	94.9	87.1	49.2	44.3	--	16
ROC-AUC	--	--	0.88	0.93	0.96	0.86	0.92	0.83	--

Terms: LR=Logistic Regression, DT=Decision Tree, DF=dengue fever, NSD=non-severe dengue, and OFI = other febrile illness. (1): Predicting DF Vs OFI for patients with dengue-like illnesses in an emergency room at a hospital in Taiwan with Logistic Regression with different probability thresholds: 0.388 and 0.636 (Ho, 2020). (2): Predicting DF Vs OFI for adult patients in Singapore that present within 72 hours of symptom onset using Decision Tree (C4.5) (Tanner, 2008). (3): Predicting DF Vs OFI for children in Thailand presenting with fever and no localizing symptoms at two different hospitals (A=KPPPH B=QSNICH, see reference) using Decision Tree (CART) (with mostly laboratory data (Potts, 2010). (4): Predicting Severe dengue versus all other illnesses (using independently created classification system) for children in Thailand in two different hospitals A=KPPPH B=QSNICH, see reference) using Decision Tree (CART) with mostly laboratory data (Potts, 2010). (5): Predicting form of severe dengue versus NSD for adults in Singapore with Decision Tree (C4.5) using laboratory values (Tanner, 2008). (6): Predicting severe dengue versus NSD for patients known to have dengue in a hospital in Malaysia with Decision Tree using personal history, clinical characteristics, and laboratory values (Tamibmaniam, 2016).

When comparing literature of WHO classification validation studies, logistic regression and decision tree modeling have equal or lower sensitivity but have higher specificity. Results from modeling

studies varied, even when performed the same by the same researchers in hospitals within the same country (Potts, 2010). Externally applying clinical models to other populations with different data collection processes can have varied outcomes (Pajouheshnia, 2018), often with worse results than when applied to the internal population that the model was built on (Siontis,2015; Ramspek, 2020). While there are multiple studies regarding machine learning clinical models for dengue in Southeast Asia and Brazil, not much has been done with regards to Mexico. Thus, there is a need for greater representation of dengue classification models among endemic regions such as Mexico to assess their internal application.

This paper will address this disparity by evaluating the potential use of prognostic and diagnostic models in the Mexican state of Yucatan. Specifically, we will conduct a descriptive analysis of the clinical and epidemiological characteristic of dengue cases in Yucatan over a 10-year period to help characterize the patient population. Then we will perform prognostic and diagnostic modeling across different years and patient groups within Yucatan, Mexico to evaluate the potential use of such models in clinical settings as well as to identify variables most important for prediction. Subsequently, we will validate our model against a large dataset with similar clinical variables for all other Mexican states to assess the generalizability of our model. Lastly, we will demonstrate how our models could potentially be used and interpreted in the clinical setting.

While complex machine learning models such as artificial neural network and support vector machine may be able to achieve better accuracy metrics than simpler models such as logistic regression and decision tree, they are not as easy to interpret (Ho, 2020). We believe that the random forest algorithm strikes a useful balance between comprehension and accuracy. See the Basic Review of Models in the Supplemental Section for more information on logistic regression, decision tree, and random forest. Some researchers such as Christodoulou et al. have found standard logistic regression to outperform random forest, but many other studies have found random forest to be more accurate than other considered classification models (Christodoulou, 2919; Couronné,2018). Additionally, while random forest is often considered a “black box” model, recent techniques such as the Local Interpretable Model-

Agnostic Explanations (LIME) method developed by researchers at the University of Washington may be able to help better explain the application of random forest in the clinical setting (Ribeiro, 2016).

In our analysis we seek to answer five related questions. What are the variables most important to correctly classifying dengue and severe dengue? What are the general odds ratio associations for having dengue or severe dengue among the most important predictors? How do our models perform and are they potentially useful in the clinical setting? Specifically, how does logistic regression compare to random forest? How does our severe versus non-severe model perform when validating it against new time periods and different populations? Lastly, how do localized model interpretations seem to corroborate or contradict variable importance analysis from the global models? We hypothesize that variables associated with warning signs and severe symptoms will be the most important variables for identifying dengue and severe dengue due to their importance in the 2009 WHO classification system. We also hypothesize, based on findings in other classification studies, that random forest will be more accurate than logistic regression across our different modeling groups. We also theorize that our models will best differentiate patients with and without severe disease due to the distinctive nature of severe illness in dengue patients.

Methods

Data Source

We utilized data from patients with febrile acute syndrome “suspected” (cases with fever of 2-7 days of duration, residing or proceeding from an area with dengue transmission) of having dengue that were notified to the Sistema Nacional de Vigilancia Epidemiológica -SINAVE- (National System of Epidemiologic Surveillance) from 2008-2019. Health services in Mexico are required to report suspected dengue cases to SINAVE (Murillo-Zamora, 2017). Clinical and epidemiological data were obtained from the case report format used by the SINAVE. Dengue diagnosis was stated as “probable” (based on clinical and epidemiological features evaluated by a physician at a health care service) and “final” (based on laboratory testing with IgG, IgM, Ns1, and PCR). Predictive model analysis was restricted to patients with a non-missing final diagnosis pertaining to dengue or “other” disease. Patients without a final diagnosis were not included in our models.

Dengue classification followed WHO 1997 guidelines for years 2008-2015 and WHO 2009 guidelines for 2016-2019. Common Variables among both datasets include demographic information such as age and sex, and ethnicity as well as clinical information such as time of symptom onset, patient outcomes, and hospitalization status. Clinical variables for the 2008-2019 and 2016-2019 datasets used in models are listed in Supplemental Table 2 and Supplemental Table 3, respectively. Data collected from the latter years was more detailed with a greater number of relevant variables.

We additionally obtained data from PAHO's PLISA Health Information Platform for the Americas regarding similar demographic, clinical, diagnosis, and serotype information for all Mexican states. Only final diagnoses for non-severe dengue, dengue with warning signs, severe dengue, Zika, and Chikungunya were provided. We used this dataset to compare descriptive information from our Yucatan-specific data to the rest of Mexico. We then filtered the dataset to only include patients with laboratory confirmed dengue in order to perform temporal and external model validation.

The 2010 population data was obtained from the Instituto Nacional De Estadística y Geografía -INEGI- (National Institute of Statistics and Geography).

Descriptive Statistics

The distribution of dengue prognosis, other febrile illnesses (OFI) diagnosis, deaths, hospitalizations, dengue serotype samples, mean age, and median time from symptom onset to receiving care across years and months for Yucatan, Mexico was calculated and illustrated in plots and a table. QGIS mapping software was used to map dengue per 100K population as well as the Poisson and Bernoulli clustering for each municipality in Yucatan, Mexico across biannual intervals. SATScan was used to calculate the significant clusters and the centroids form the municipalities were used for Bernoulli clustering. Furthermore, Tables were created for covariates of interest stratified by relevant diagnostic and prognostic outcomes.

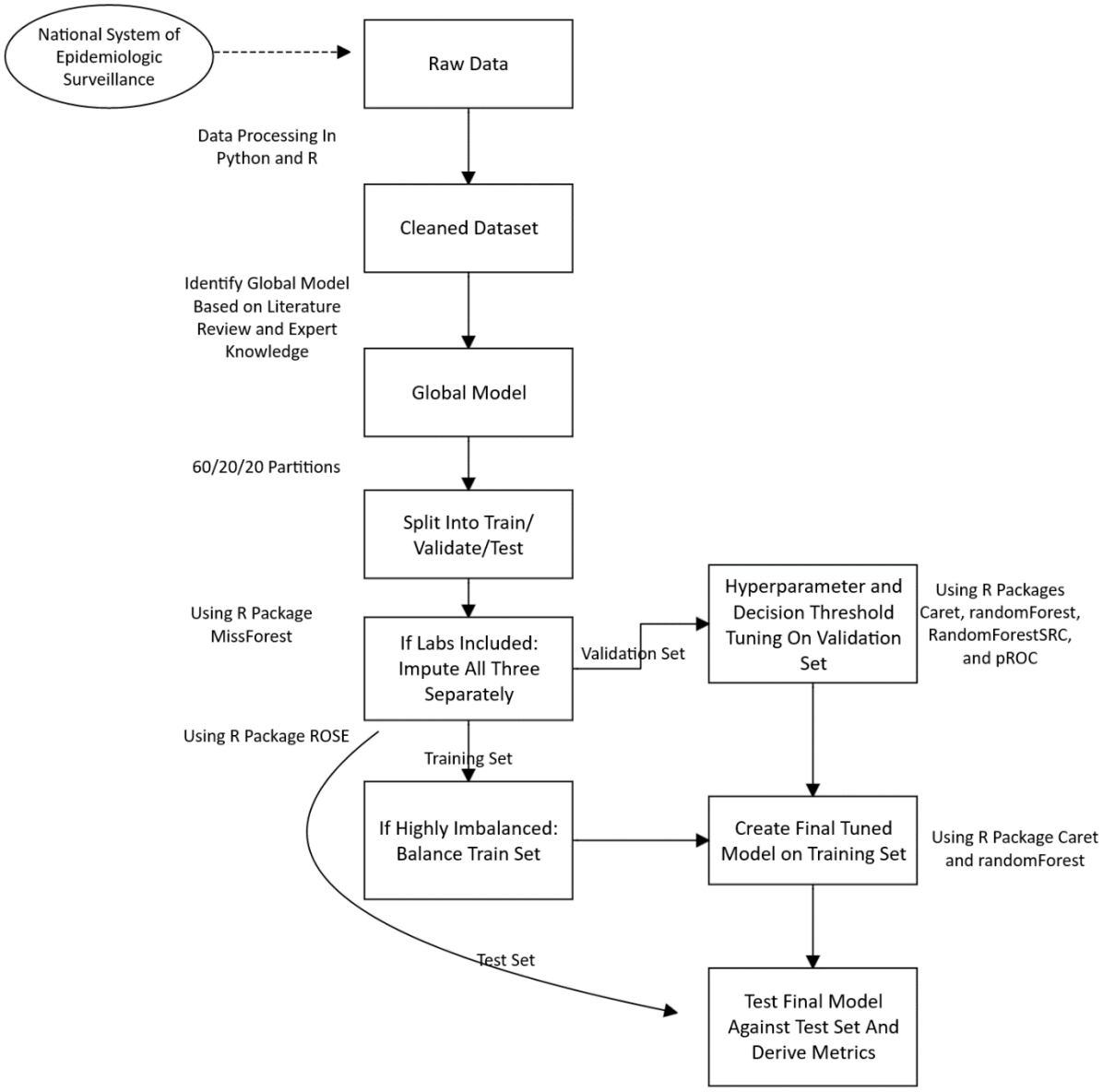
Data Preparation and Statistical Analysis

Figure 1 demonstrates the general processing and modeling steps. Two aggregated datasets were constructed for modeling purposes pertaining to the years 2008-2019 and 2016-2019 with a total of 24

and 69 variables considered, respectively. 37 of the 69 variables considered in the latter years were aggregated into 7 grouped variables (based on the underlying pathophysiological mechanisms), see footnotes in Table 5.

Specifically, patients were determined to have “bleeding” if they had bloody mucus, melena, ecchymosis, hematomas, epistaxis, gingivorrhagia, hematemesis, metrorrhagia, subarachnoid hemorrhage, and or mottled skin. Patients were noted as having “neurological” symptoms if they had an altered state of consciousness, photophobia, stupor, and or disorientation. Similarly, the group “hemodynamic” was assigned to those with either cold extremities, capillary filling, convergent pressure, arterial hypotension, and or lipothymia. “Atypical manifestations” was assigned to those with dyspnea, respiratory failure, and or myocarditis. “Plasma leakage” was determined by ascites, pleural effusion, plasma leakage, swelling, and or increased hematocrit. Patients with nausea, emesis, diarrhea, and or abdominal pain were reported as having “nausea plus”, and patients with splenomegaly and or hepatomegaly were noted as having “Visceromegaly”.

Figure 1: Data Processing and Modeling Structure



Both the 2008-2019 and 2016-2019 datasets were used to model confirmed dengue versus other febrile illness, but the dataset from 2016-2019 was also used to model severe dengue (confirmed "severe dengue" or "dengue with warning signs") versus non-severe dengue (confirmed dengue without warning signs) as well as model severe dengue (confirmed severe dengue and dengue with warning signs) versus all other kinds of febrile illness (confirmed dengue without warning signs and OFI).

Based on expert recommendation, we assumed that all positive comorbidity and symptomatic data were properly recorded and that all missing or 0 values were negative. We transformed all negative

values to 0 and all positive values to 1 for modeling purposes. Around 470 duplicated rows were identified by patient ID and date of first symptoms. One row for each unique ID and symptom onset date was selected based on completeness of information or was otherwise randomly picked. The datasets based on these assumptions and edits were used for all relevant tables, descriptive analyses, and models.

For each modeling group, we filtered to all rows with non-missing outcome values based on the recorded final diagnosis. We then shuffled the data to resolve any intrinsic ordering, and then randomly partitioned the data into 60/20/20 test, validate, and test splits around the outcome variable. Single random forest-based imputation using the R package "missForest", shown to be accurate when compared to other popular methods like knn, was then performed independently for each set if any lab variables were included (Stekhoven, 2012). We decided not to include any lab values with more than 60% missing rows and ultimately did not include models with labs in our paper as the lowest lab variable missing rates were still around 40%.

The severe vs all and the dengue versus OFI models experienced significant unbalancing in favor of the controls which is something known to affect random forest accuracy (More, 2020; Lunardon, 2014). To address the imbalance, we employed R package ROSE, which uses smoothed bootstrapping to generate new classes and balance the data, on the training set for these model groups (Lunardon, 2014). We did not balance the test or validation sets as we sought to overcome the random forest's weakness to imbalanced data but still predict on this imbalance since it most reflects reality.

The aggregated databases were created using software python, QGIS and SATScan were used for spatial analysis, and R statistical software was used for the descriptive and modeling analyses. A set random seed was used in all stochastic settings to maintain reproducibility.

Model Framework

Through clinical expertise and literature review, we created a proposed global model regarding variables that are relevant when differentiating between dengue, severe dengue, and other common febrile illnesses. We then used the training sets for each modeling group to subset the global model into a final model though variable selection with the R package Boruta to minimize number of predictor variables

and improve the signal to noise ratio efficiency of our model (Speiser, 2019; Sanchez-Pinto, 2018). Sanchez-Pinto, et al. found that the R package "Boruta", which employs progressive elimination random forest variable selection, performs especially well when compared to other forest and regression-based methods (Sanchez-Pinto; Kursa, 2010). Boruta estimates each variable's importance (mean decrease in accuracy of overall random forest model) compared to the expected importance achieved with randomized permutated copies (Kursa, 2010). For each model group the top 10 most important variables were chosen for analysis.

An initial random forest model was created on the training set for each group, hyperparameter tuning was performed on the validation set, and the final model was created on the training set with relevant tuned hyperparameters. Two different tuning grids for the random forest "mtry" hyperparameter, the number of samples randomly selected for each node, and the "nodesize" hyperparameter, the minimal final node size, were created based on the validation using the R package caret to maximize Kappa and randomforestSRC to minimize the out of bag error (OOBE) (Breiman, 2018). All random forest models were created with 1000 trees, caret hyperparameter tuning was performed with 2 repeated 5-fold cross validation, and final tuned models were created using 5 repeat 10-fold cross validation. The final tuned models were then used to predict diagnosis or prognosis classification against the test set and final evaluation metrics such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), Accuracy, AUC, kappa, and F_2 score were calculated. See Supplemental Table 1 for interpretations of these metrics.

Binomial logistic regression models were constructed from the training sets using 5 repeat 10-fold cross validation. Four decision thresholds were defined using the R package "PresenceAbscence" based on the model's performance against the validation set: the default of 0.5, Youden's J statistic which maximizes sensitivity and specificity, an index utilizing a cost function where false negative cost twice as much as false positives, and the threshold at which the validation set reached 90% sensitivity (Freeman, 2007). See Figure 6 for more information about these threshold indexes. These tuned thresholds were used as probability cut-off points for classifying the outcome when predicting against the test set. The

same evaluation metrics as with random forest were then evaluated for each threshold index based on the same underlying logistic regression model.

Furthermore, we used the overall Mexico data from PAHO pertaining to the years 2017-2021 to perform basic validation of our severe versus non-severe dengue model. We recreated our severe vs non-severe models on variables common both to the Yucatan-specific and overall Mexican datasets. We followed the same steps as above except that we chose the 7 most important variables instead of 10. Please refer to PAHO at https://www.paho.org/data/index.php/es/temas/indicadores-dengue/dengue-subnacional/521-mex-egi-cuadro-clinico-es.html to see variables shared by both datasets. Grouped variables such as "bleeding" and "neurological" were constructed in the same manner as the Yucatan-specific models, but some of the variables that made up these groups in the Yucatan data were absent from the PAHO Mexico data.

We then temporally validated our model by using it to predict severe versus non-severe dengue on the overall Mexican data filtered to just Yucatan for the years 2020-2021. Of note, the Mexico data was most recently obtained from PAHO on April 25th, 2021. Lastly, we used the same trained model to predict severe versus non-severe dengue for all states other than Yucatan using the Mexico data from 2017-2019.

Aside from predictive models, we also performed a 5 repeated 10-fold cross-validated logistic regression on the entire pre-split dataset filtered down to the 10 most important variables for each model group to investigate the general conditional odds of having the outcome. The conditional odds ratios and confidence intervals for the most important variables of each modeling group are provided down below.

Modeling Groups

We modeled severity in accordance with the 2009 WHO classification system such that patients with either the final diagnosis of severe dengue or dengue with warning signs were considered as being in the spectrum of severe forms of dengue. We did this because of our assumption that patients with warning signs will develop severe dengue without proper intervention (WHO, 2009). Furthermore, we modeled severity from two different perspectives: predicting severe versus non-severe dengue and predicting

severe versus all other febrile illnesses. We also modeled dengue versus other febrile illness (OFI) during both 2016-2019 and 2008-2019 as the former period has more variables and the latter has a much larger sample size.

Model Agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) is a widely used framework that seeks to provide local individual level explanations about the relative effect of features on the classification decision made by black box model (Ribeiro, 2016). LIME estimates the relative weights for relevant features for contributing to or against the outcome of interest for each individual (Ribeiro, 2016). The LIME framework works by simulating a dataset from the training data used to fit the model and transforming the simulation data and predicted case (Goode, 2021). The trained random forest model is then applied to the simulated data and a distance between the simulated data points and prediction of interest is calculated (Goode, 2021). An explainable model is subsequently employed with selected features from the simulated data serving as predictors and the prediction results from the simulated data serving as the outcome (Goode, 2021). Ribeiro created a python package for LIME which was adapted into a R package by Pederson and Betsy; both packages utilize ridge regression as the explainable model (Goode, 2021; Ribeiro, 2021). LIME analysis is accompanied by a deviance ratio, R^2, that explains the linear fit of the explainer model (Goode, 2021).

We utilized the R package implementation of LIME to estimate relative feature weights across 8 different features for a particular patient as part of a case study analysis. We then performed a similar analysis across 13 randomly selected patients in the target groups to see if estimate feature weights matched our previous global feature importance results. Using the LIME package options, we used a lasso model to choose 8 features, a kernel width of 3, a Manhattan distance function, and we created 5000 permutations for each explanation. By default, the continuous age variable is separated into 4 quantile bins for LIME analysis.

Results

Descriptive Statistics

Table 4 illustrates the distribution of final diagnoses, deaths, hospitalizations, median days between symptom onset and presenting to the clinic or hospital, and median duration of hospitalization across all years. 2011 and 2012 had relatively high number of dengue cases, hospitalizations, and deaths. There was a relatively low number of cases from 2014 through 2019, though with a high number of deaths and hospitalizations in 2019. Deaths were mostly recorded in patients with confirmed dengue except for the later years in 2016-2019 where most deaths occurred in patients with OFI or without a final diagnosis. The median days from symptomatic onset to receiving care was 4 from 2008-2012 and 3 from 2013-2019. The median days of hospitalization was 3 throughout the study. The mean age of febrile patients ranged from 23.3-29 years old across the years. Table 4 shows circulation of dengue serotypes in Yucatan by year of study. The number of samples with serotype detection was small, but DENV1 was more prevalent from 2012-2015 and DENV2 was more prevalent from 2009-2010.

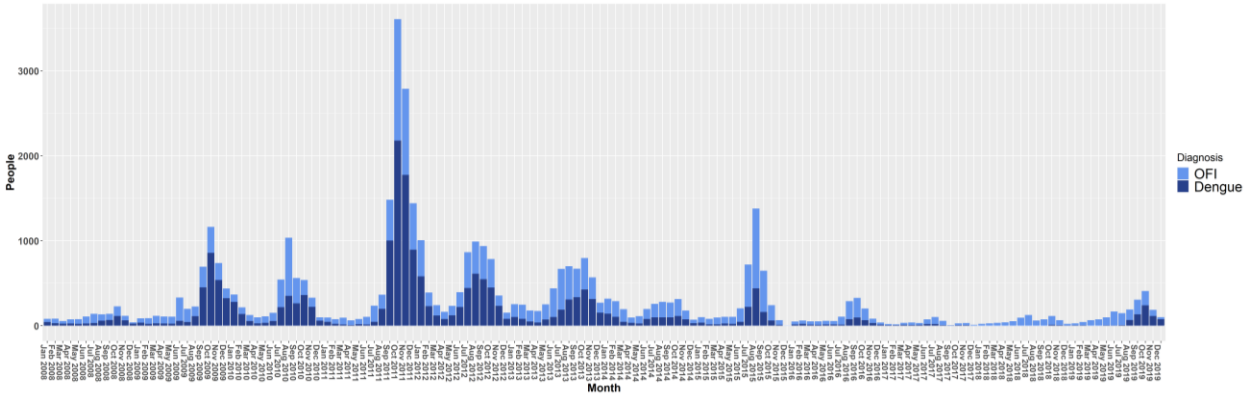
Figure 2 shows the occurrence of cases in Yucatan, Mexico throughout the entire study period. The incidence shows a seasonal increase that coincides with the mid and late summer, reaching its peak in the fall. The number of cases varies per year, with periods of high incidence (2009-2012) and others of more moderate incidence (2016-2017-2018). Figure 3 shows the distribution of cases based on the 2007 WHO classification for the period 2016-2019. From this plot, we can observe an intraepidemic increase in the occurrence of severe cases, with a higher proportion of severe forms as the outbreak progresses especially in 2016 and 2019.

Table 4: Cases, Hospitalizations, and Deaths Across Years in Yucatan, Mexico

Metrics	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
DF (N)	425	1662	1394	4108	1874	1399	561	804	--	--	--	--
DHF (N)	110	879	651	2108	1838	771	390	254	--	--	--	--
NSD (N)	--	--	--	--	--	--	--	--	264	74	13	264
DWS (N)	--	--	--	--	--	--	--	--	103	31	7	357
SD (N)	--	--	--	--	--	--	--	--	17	3	0	34
OFI (N)	745	1782	2105	4216	2801	3044	1651	2686	995	343	715	1164
Mean Age	25.86	26.00	25.72	25.94	27	24.86	25.07	29.47	26.83	26.36	26.54	23.29
Hospitalizations (N)	286	1229	1173	4409	3236	1529	897	1206	374	249	190	1063
Dengue Pts Deaths (N)	1	2	2	24	15	9	0	0	9	4	1	15
OFI Pts Deaths (N)	1	0	0	0	0	0	0	2	12	13	14	24
Deaths: No Final Diag. (N)	0	0	0	1	0	0	0	0	1	6	1	13
Days From Sxs to Care: Median (Q1,Q3)	4 (2,5)	4 (3,5)	4 (2,5)	4 (3,5)	4 (3,5)	3 (2,5)	3 (2,5)	2 (1,4)	3 (1,5)	3 (1,5)	3 (2,4)	3 (2,4)
Days Hospitalized: Median (Q1,Q3)	3 (1,4)	3 (2,5)	3 (2,5)	3 (2,5)	3 (2,5)	3 (2,4)	3 (2,5)	3 (2,5)	3 (2,5)	3 (2,6)	3 (2,5)	3 (2,5)
Dengue Serotypes	--	1,2	1,2	1,2	1,2,4	1,2,4	1,2,4	1,2,4	--	--	--	--

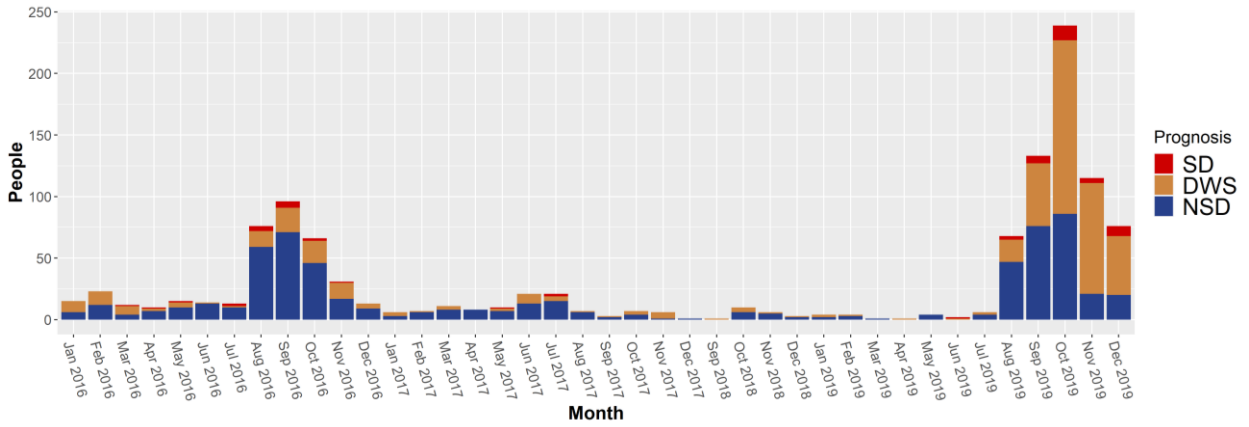
Terms: DF=Dengue Fever, DHF=Dengue Hemorrhagic Fever, NSD=Non-Severe Dengue, DWS= Dengue with Warning Signs, SD=Severe Dengue, OFI=Other Febrile Illness. Data comes from SINAVE Yucatan Dataset. Dengue diagnosis based on laboratory confirmation and classification based on the 1997 WHO Classification (DF, DHF) and 2009 WHO Classification (NSD, DWS, SD). Serotype data derived from small PCR samples in patient data. DENV1 was more prevalent in years 2012-2015. DENV2 was predominant in 2011. DENV1 and DENV2 were about equally present in 2009-2010.

Figure 2: Confirmed Dengue Diagnosis Across All Months (2008-2019)



Terms: OFI= other febrile illness. Plot depicts the total number of confirmed dengue and non-dengue cases for each month from 2008-2019. Incidence peaks appear to occur form Summer-Fall. Especially large peaks occurred in 2009, 2011, and 2013. Cases without a final diagnosis were not included.

Figure 3: Confirmed Dengue From 2016-2019 Disease According to 2009 WHO Classification

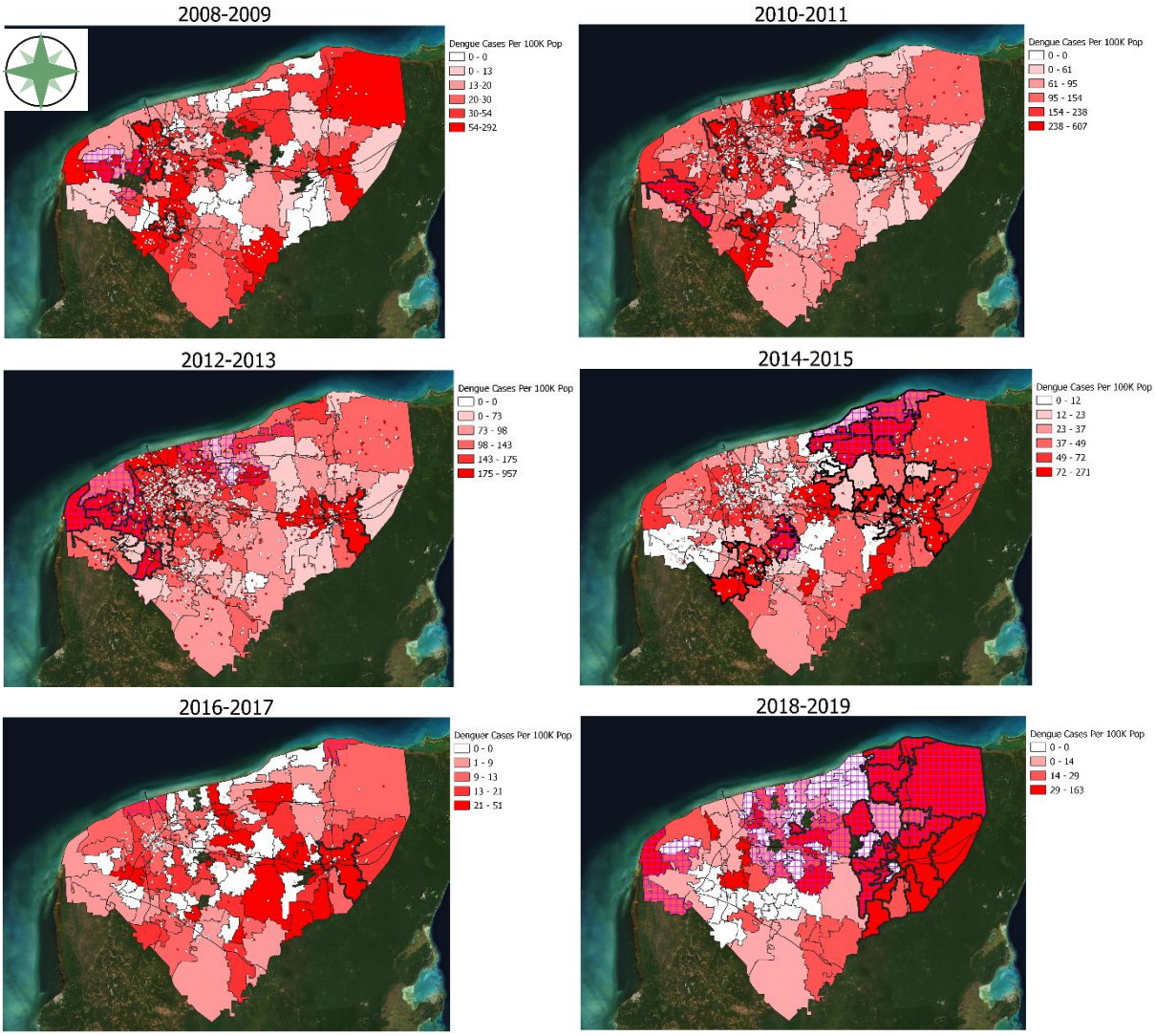


Terms: SD= severe dengue, DWS = dengue with warning signs, and NSD= non-severe dengue. Plot depicts the total number of patients with confirmed dengue separated by their final prognosis according to the 2009 WHO classification system. There is a large increase in dengue incidence in 2019, especially for patients with severe dengue and dengue with warning signs.

Figure 4 additionally reveals the spatial distribution and clustering of dengue cases in Yucatan. Kuldorff's Poisson clustering indicates high concentration of dengue given municipality population while Bernoulli clustering indicates high aggregation of dengue cases compared to other febrile illnesses. Significantly high Poisson clustering is predominantly centered in Mérida and Oxkutzcab municipalities from 2008-2013 whereas this clustering appears to gradually shift to the East around Valladolid from

2014-2019. Bernoulli clustering of dengue compared to non-dengue cases occurred West of Mérida circa 2008-2010, in North-central Yucatan circa 2012-2013, East of Oxkutzcab and in North-East Yucatan circa 2014-2015, around Mérida and in the North-East circa 2016-2017, and across most of the West circa 2018-2019.

Figure 4: Space-Time Spatial Distribution and Clustering of Dengue in Yucatan Mexico (2008-2019)



LEGEND:

Highways

Point Cases

OFI

Dengue

High Bernoulli Cluster

High Poisson Cluster

Dengue/OFI (other febrile illness) cases are randomized using NRandK method with GeoPrivacy in QGIS to preserve spatial patterns and protect privacy. Population based on 2010 national census. Clustering determined though Kuldrof Test with SatScan. Bernoulli clustering indicates high clusters of dengue cases compared to OFI. Poisson clusters indicate high clustering of dengue given municipality population. Point locations were not recorded for years 2018-2019.

Regarding the 2016-2019 dataset (see Supplemental Table 3), the dengue versus OFI univariate risk ratios were non-significant for all comorbidities and common acute findings, significantly protective for 1 common respiratory symptom (cough) and non-significant for the rest, and significantly harmful for

8 general symptoms (positive-torniquet, arthritis, petechia, exanthema, headache, pruritus, polyarthralgia, and conjunctivitis), 4 warning signs (intense abdominal pain, persistent vomiting, epistaxis, gingivorrhagia), and 6 severe symptoms (cold extremities, abdominal pain, tachycardias, hematemesis, arterial hypotension, and metrorrhagia). Univariate relative risk for severe dengue (SD or DWS) vs non-severe dengue was not significant for all common acute findings and was significantly harmful for 4 comorbidities (hypertension, bleeding disorder, diabetes, and pregnant), 4 general symptoms (positive torniquet, diaphoresis, chills, and petechia), 2 common respiratory symptoms (dyspnea and cough), all but one warning sign (see Supplemental Table 3), and all but 4 severe symptoms.

Severe Versus Non-Severe Dengue

All 1,167, patients that met the criteria for confirmed dengue from 2016-2019 were included in the analysis. As seen in Figure 7, variable importance was assessed in two different ways. The left figures illustrate the importance, estimated mean decrease in accuracy upon removal, for each variable selected by Boruta's progressive elimination random forest algorithm. The right figures depict the variable importance and mean decrease in gini entropy upon removal for selected variables based in the final tuned random forest models. In both the Boruta plot and final model plot, intense abdominal pain appears to be the most important predictor to accurately predicting severity followed by bleeding, plasma leakage, petechiae, and persistent vomiting.

Based on the pre-split data logistic regression on the 10 most important variables, we found that intense abdominal pain, bleeding, plasma leakage, and petechiae had large odds ratios and wide confidence intervals of 71.25 (25.2,201.43), 48.41 (20.29, 115.52), 36.75, (13.99, 96.54), 4.86 (2.74, 8.61), respectively. The high variances of our estimates are likely due to problems with separability since very few of these predictor variables were evident in patients without severe dengue. While we addressed multicollinearity when choosing our final variables, some of the predictors are still highly correlated with one another which can affect the magnitude of the odds ratios.

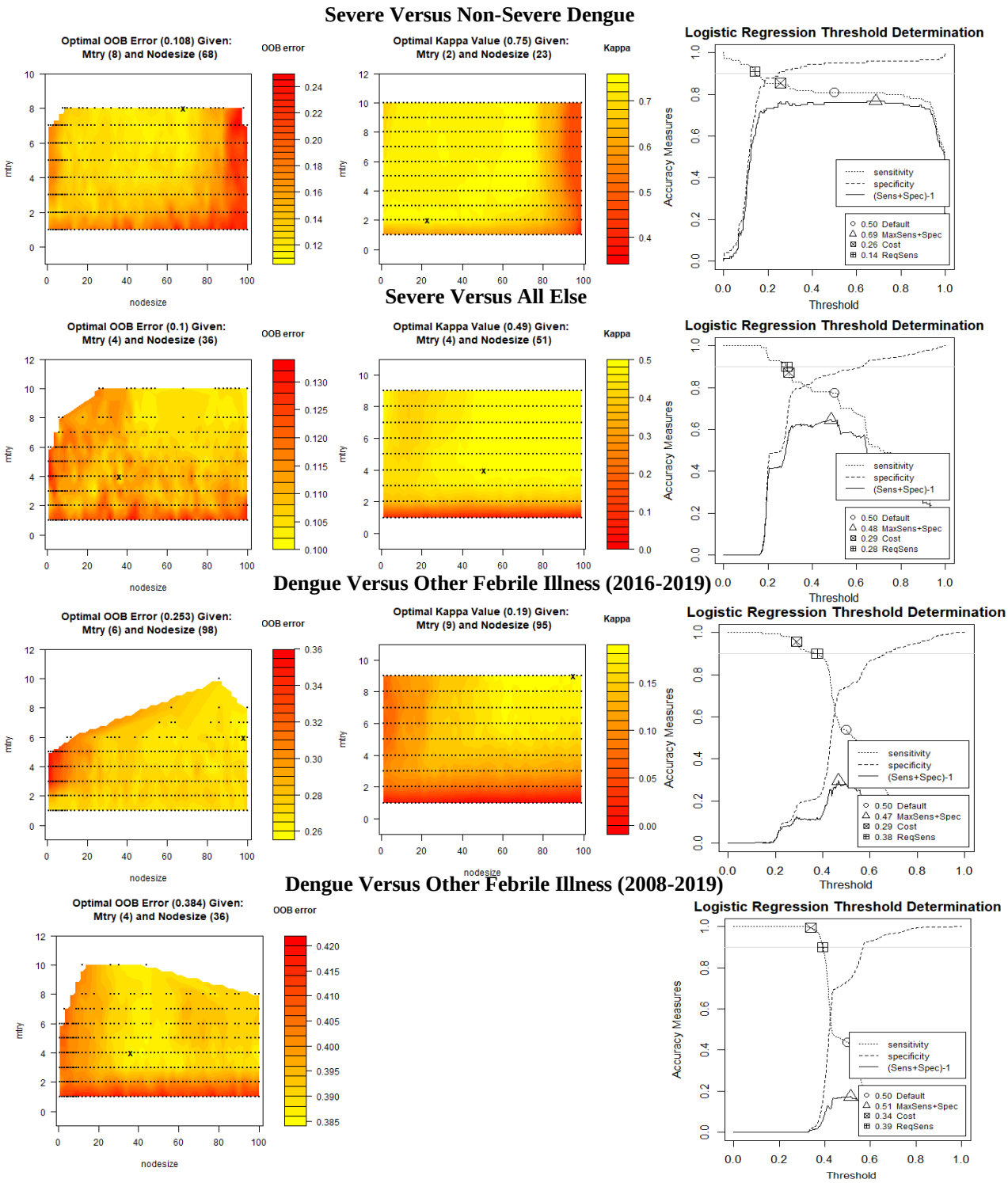
Table 4 depicts the results from our predictive logistic regression model and random forest models pertaining to the selected variables for all model groups. The logistic regression model results are

presented from the same model using different decision thresholds tuned on the validation set. The random forest results are from three different models, one with default setting and two that were tuned on the validation. When comparing the default versions for both logistic regression and random forest, the latter has slightly higher sensitivity, NPV, accuracy, Cohen's Kappa, and F-2 Score. The default logistic regression meanwhile has higher specificity, PPV, and AUC. The differences between the two default models, however, is relatively minute. Using the cost-sensitive threshold increases logistic regression sensitivity to 85.5% (78.8-92.0) with a slight cost to the specificity. Likewise, F-2 Score, Choen's Kappa, accuracy, and negative predictive value all increase using the cost function threshold. In contrast, hyperparameter tuning "mtry" and "nodesize" for the the random forest model does not seem to make much of a difference. Figure 6 depicts the resulting Kappa and OOBE values for random forest trained on the validation set across the different mtry-nodesize combinations. Figure 6 also illustrates the different sensitivity and specificity values obtained using the 4 pre-determined decision thresholds when testing the trained logistic regression against the validation set.

Severe Dengue Versus All Else

4,384 patients that had a final diagnosis from 2016-2019 were included in the analysis. The 10 most important variables were highly similar, and the 5 most important variables were the exact same as in the previous model, albeit with slightly different ordering. Pre-split data odds ratio estimates for intense abdominal pain, plasma leakage, bleeding, persistent vomiting, and petechia were significant at 4.50 (3.42, 5.91), 3.69 (2.75,4.95), 3.25 (2.45,4.31), 2.66 (1.84,3.46), 2.40 (1.84,3.15), respectively. Odds ratio estimates were much stable, likely due to a larger sample size reducing the likelihood of non-separability. When comparing the default models, the logistic regression slightly outperformed the random forest in all metrics that we considered. The threshold derived using Youden's Index was similar to that of the default at 0.48 while the cost function threshold was much lower at 0.295. The latter resulted in a noticeable increase in sensitivity, 86.4% (80.0,92.8), but a large reduction in specificity and accuracy. Both random forest-tuned models showed slightly higher sensitivity, 79.1% (71.5,86.7) and 80.9% (73.6,88.3) and marginally lower accuracy and positive predictive value when compared to the default model.

Figure 6: Decision Threshold and Hyperparameter Tuning Against Validation Sets



Left: Hyperparameter tuning grid of mtry and nodesize values that achieves the highest lowest OOBE (out of bag error) on the validation set for "alternate tuned random forest" model using R package randomforestSRC. **Middle**: Hyperparameter tuning grid of mtry and nodesize to maximize Kappa for the "tuned random forest" models with R package caret. This tuning was not performed on the Dengue Vs Other (2008-2019) model group due to computational constraints. **Right**: Decision threshold cut offs based on the validation set according to predefined criteria for logistic regression model

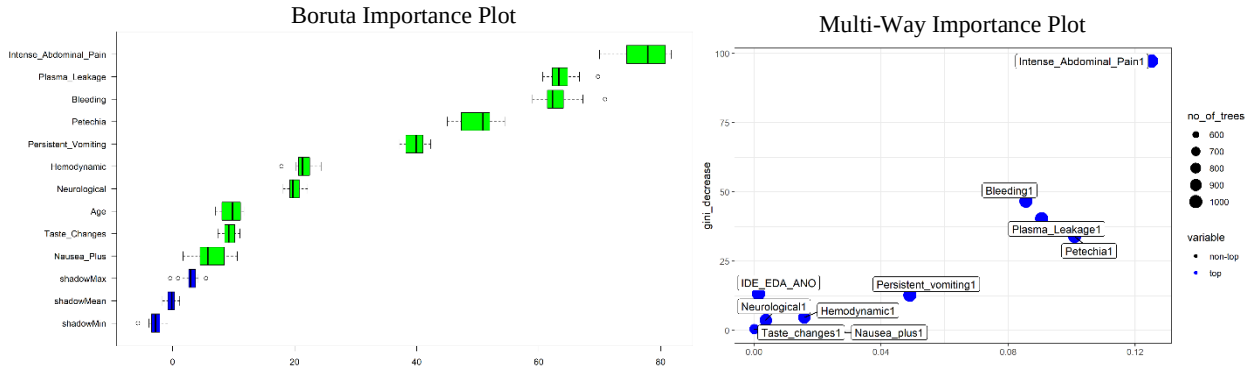
Dengue Versus OFI (2016-2019)

4,384 patients with a final diagnosis that were included in the model. As shown in Figure 7, intense abdominal pain, age, persistent vomiting, petechia, and plasma leakage bleeding were the five most important variables according to the Boruta analysis. However, the final random forest results contradict that of the Boruta selection, indicating that age was the most important predictor, not persistent vomiting, and that exanthema was the third most important. The presplit data logistic regression resulted in significant odds ratios for intense abdominal pain, petechia, exanthema, and persistent vomiting as 1.95 (1.53, 2.49), 1.44 (1.15, 1.81), 1.59 (1.35, 1.87), and 1.81 (1.30, 2.52), respectively. Chills appeared protective with an odds ratio of 0.61 (0.5,0.73). When categorizing age, odds ratios for 15-28 years, 29-42 years, 43-57 years, and >58 years compared to ages <14 years old were 1.03 (0.87-1.23), 0.65 (0.52-0.81), 0.49 (0.37-0.66), and 0.8 (0.57-1.13), respectively.

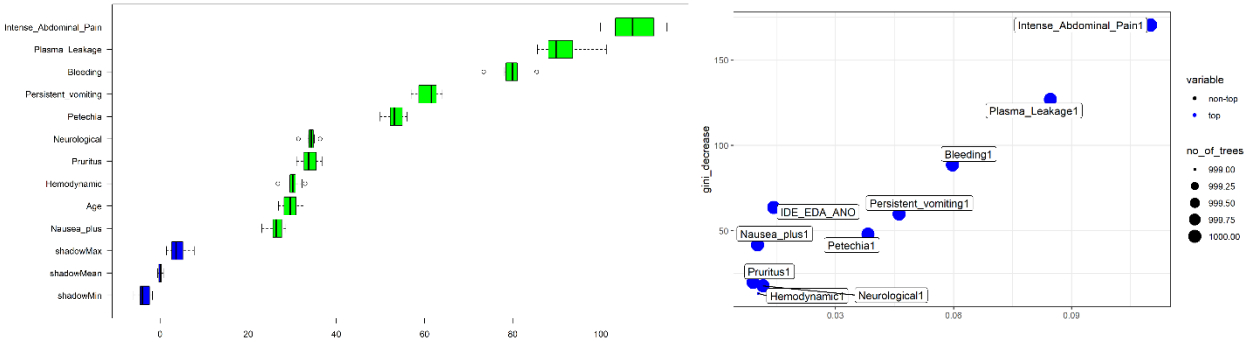
When comparing predictive model results, the logistic regression with the default threshold slightly outperformed random forest in all considered metrics aside from specificity. In both cases sensitivity is only slightly larger than 50% while accuracy, Cohen's Kappa, and AUC are considerably low. The alternate decision thresholds improve the logistic regression's sensitivity though at increasing cost to accuracy, Kappa, and specificity. Likewise, the tuned random forest models improve sensitivity while decreasing specificity and accuracy.

Figure 7: Feature Importance Using Boruta Analysis and Final Tuned Random Forest

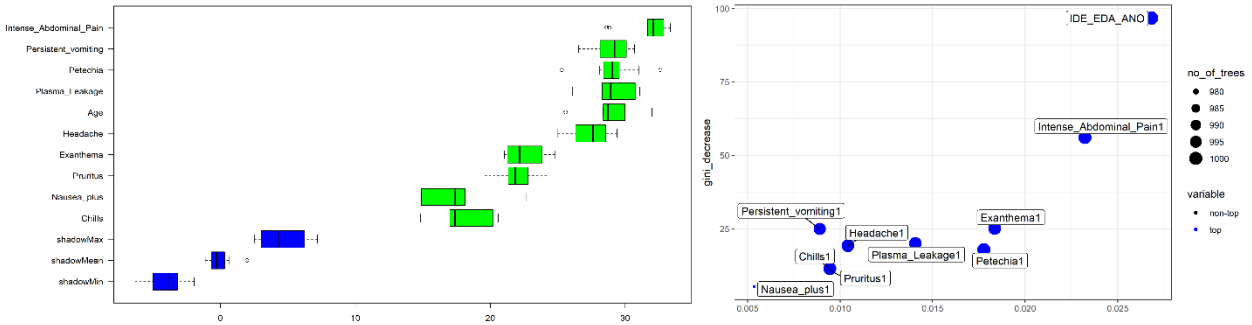
Severe Versus Non-Severe Dengue



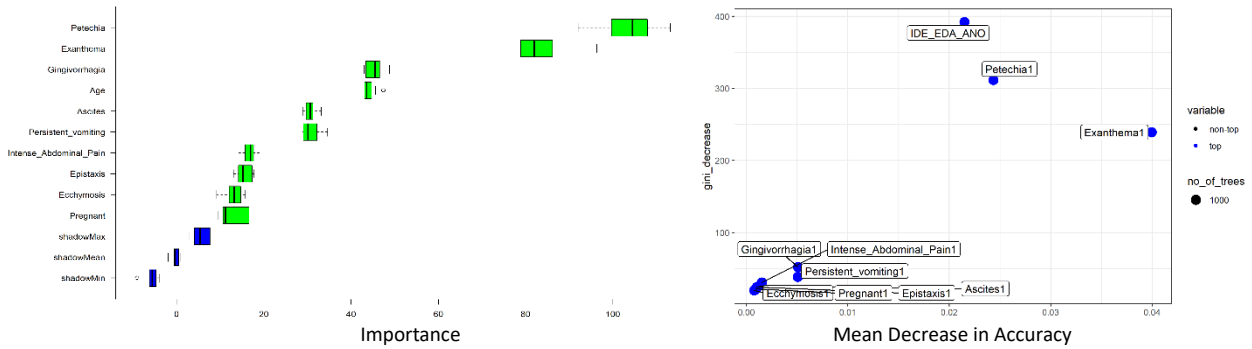
Severe Versus All Else



Dengue Versus Other Febrile Illness (2016-2019)



Dengue Versus Other Febrile Illness (2008-2019)



Left: Boruta importance output on final variables selected to go into model. Blue boxplots indicate the minimum, mean, and max importance expected by Boruta. **Right:** Final importance metrics derived from the final trained random forest model including decrease in Gini impurity and mean decrease in accuracy. Size of indicators relates to the number of times the particular variables was present in a tree out of the 1,000 trees used to create the random forest model. "IDE_EDA_ANO" stands for patient's age.

Dengue Versus OFI (2008-2019)

From the 39 variables that were measured across all years for 42,624 patients with final diagnoses, the most significant factors for modeling dengue appeared to be petechia, exanthema, gingivorrhagia, age, and ascites. These variables had pre-split data conditional odds ratios of 2.6 (2.41,2.80) for petechia, 1.66 (1.59,1.73) for exanthema, 1.59 (1.42,1.79) for gingivorrhagia, and 2.95 (2.16-4.03) for ascites when considering the other 6 most important variables. With respect to those < 14 years old, the odds ratios for 15-28 year old, 29-42 year old, 43-57 year old, and > 58 year old patients was 1.37 (1.30-1.44), 0.94 (0.88,0.99), 0.95 (0.89,1.02), and 0.79 (0.72,0.87) when controlling for the other 9 important variables.

The default random forest slightly outperformed the base logistic regression in all metrics except sensitivity for which they were both equally low at 44.3%. The threshold derived from Youden's Index was approximately the same as the default. The cost function index led to a dramatic increase in sensitivity to around 99% but decreased specificity down to just 1.6%. In contrast, there is essentially no change between the default and tuned random forest models.

Table 4: Final Assessment Metrics Among Model Groups

Model Group	Metric	Logistic Regression	Logistic Regression: Youden's Index	Logistic Regression: Cost Function	Minimum Sensitivity Threshold (0.9)	Random Forest	Tuned Random Forest	Alternate Tune Random Forest
Severe (SD or DWS)	Threshold and Hyperparameters	DT= 0.5	DT= 0.69	DT= 0.256	DT= 0.14	Mtry= 3.16, NS=1	Mtry=2, NS=23	Mtry=8, NS=68

Versus NSD	Sensitivity	79.1 (71.5,86.7)	74.5 (66.4,82.7)	85.5 (78.9,92)	91.8 (86.7,96.9)	81.8 (74.6,89)	81.8 (74.6,89)	81.8 (74.6,89)
	Specificity	95.9 (92.4,99.4)	98.4 (96.1,100)	91.9 (87,96.7)	71.5 (63.6,79.5)	94.3 (90.2,98.4)	94.3 (90.2,98.4)	94.3 (90.2,98.4)
	PPV	94.6 (89.9, 99.2)	97.6 (94.4, 100.9)	90.4 (84.7, 96.1)	74.3 (66.9, 81.6)	92.8 (87.6, 97.9)	92.8 (87.6, 97.9)	92.8 (87.6, 97.9)
	NPV	83.7 (77.6, 89.8)	81.2 (74.9, 87.5)	87.6 (81.9, 93.3)	90.7 (84.9, 96.5)	85.3 (79.3, 91.2)	85.3 (79.3, 91.2)	85.3 (79.3, 91.2)
	Accuracy	88 (83.1, 91.9)	87.1 (82.1, 91.1)	88.8 (84.1, 92.6)	81.1 (75.5, 85.9)	88.4 (83.6, 92.2)	88.4 (83.6, 92.2)	88.4 (83.6, 92.2)
	AUC	0.93	0.93	0.93	0.93	0.92	0.92	0.91
	Kappa	0.757	0.738	0.775	0.626	0.766	0.766	0.766
	F-2 Score	0.82	0.78	0.86	0.88	0.84	0.84	0.84
Severe (SD or DWS) Versus All Else (NSD or OFI)	Threshold and Hyperpara meters	DT= 0.5	DT= 0.48	DT= 0.295	DT= 0.285	Mtry= 3.16 , NS=1	Mtry=4, NS=51	Mtry=4, NS=36
	Sensitivity	80 (72.5, 87.5)	80.9 (73.6, 88.3)	86.4 (80,92.8)	90 (84.4, 95.6)	76.4 (68.4, 84.3)	79.1 (71.5, 86.7)	80.9 (73.6, 88.3)
	Specificity	88.8 (86.5, 91)	88.4 (86.1, 90.7)	76.4 (73.4, 79.4)	67.1 (63.8, 70.4)	87.9 (85.5, 90.2)	86.7 (84.3, 89.1)	85.9 (83.4, 88.4)
	PPV	50.6 (43.1, 58)	50 (42.7, 57.3)	34.4 (28.8, 40)	28.2 (23.5, 32.9)	47.5 (40.1, 54.8)	46 (38.9, 53.1)	45.2 (38.2, 52.1)
	NPV	96.9 (95.6, 98.2)	97 (95.7, 98.3)	97.5 (96.3, 98.7)	97.9 (96.7, 99.1)	96.3 (94.9, 97.7)	96.7 (95.3, 98)	96.9 (95.6, 98.2)
	Accuracy	87.7 (85.3, 89.8)	87.4 (85.1, 89.6)	77.6 (74.7, 80.3)	70 (66.8, 73)	86.4 (84, 88.6)	85.7 (83.2, 88)	85.3 (82.8, 87.6)
	AUC	0.89	0.89	0.89	0.89	0.88	0.88	0.89
	Kappa	0.551	0.548	0.381	0.295	0.509	0.503	0.499
	F-2 Score	0.72	0.72	0.66	0.63	0.68	0.69	0.7

Dengue Versus OFI (2016-2019)	Threshold and Hyperparameters	DT= 0.5	DT= 0.47	DT= 0.289	DT= 0.377	Mtry= 3.16 , NS=1	Mtry=9, NS=95	Mtry=6, NS=98
	Sensitivy	54.1 (47.7 ,60.5)	57.5 (51.2 ,63.9)	95.3 (92.6 ,98)	90.1 (86.3 ,94)	51.1 (44.7 ,57.5)	60.1 (53.8 ,66.4)	60.1 (53.8 ,66.4)
	Specificity	73.1 (69.7 ,76.5)	69.4 (65.8 ,72.9)	18.5 (15.5 ,21.5)	25.8 (22.4 ,29.2)	73.9 (70.5 ,77.3)	65.3 (61.6 ,69)	67 (63.4 ,70.7)
	PPV	42.1 (36.5 ,47.7)	40.5 (35.2 ,45.8)	29.8 (26.5 ,33)	30.6 (27.1 ,34)	41.5 (35.8 ,47.2)	38.6 (33.6 ,43.6)	39.8 (34.7 ,44.9)
	NPV	81.5 (78.3 ,84.6)	81.8 (78.6 ,85.1)	91.5 (86.8 ,96.3)	87.8 (83.2 ,92.5)	80.6 (77.5 ,83.8)	81.9 (78.5 ,85.2)	82.3 (79 ,85.5)
	Accuracy	68 (64.8, 71.1)	66.2 (63, 69.3)	38.9 (35.7, 42.2)	42.9 (39.6, 46.3)	67.8 (64.6, 70.9)	63.9 (60.6, 67.1)	65.2 (61.9, 68.3)
	AUC	0.68	0.68	0.68	0.68	0.67	0.67	0.67
	Kappa	0.249	0.237	0.081	0.098	0.232	0.216	0.233
	F-2 Score	0.51	0.53	0.66	0.65	0.49	0.54	0.55
Dengue Versus OFI (2008-2019)	Threshold and Hyperparameters	DT= 0.5	DT= 0.51	DT= 0.339	DT= 0.391	Mtry= 3.16 , NS=1	--	Mtry=4, NS=36
	Sensitivity	44.3 (42.7 ,45.8)	42.6 (41.1 ,44.2)	99.1 (98.8 ,99.4)	89.7 (88.8 ,90.6)	44.3 (42.7 ,45.8)	--	44.4 (42.9 ,45.9)
	Specificity	73.6 (72.3 ,74.9)	74.7 (73.4 ,75.9)	1.6 (1.2 ,2)	17 (15.9 ,18.1)	75.3 (74 ,76.5)	--	75.2 (73.9 ,76.5)
	PPV	60.6 (58.8 ,62.3)	60.7 (58.9 ,62.5)	48 (46.9 ,49.1)	49.8 (48.6 ,50.9)	62.1 (60.4 ,63.9)	--	62.1 (60.4 ,63.9)
	NPV	59 (57.7 ,60.3)	58.7 (57.4 ,60)	65.1 (56.2 ,74.1)	64.3 (61.5 ,67)	59.6 (58.3 ,60.8)	--	59.6 (58.3 ,60.9)
	Accuracy	59.6 (58.5, 60.6)	59.3 (58.3, 60.4)	48.2 (47.2, 49.3)	51.8 (50.7, 52.8)	60.4 (59.4, 61.5)	--	60.5 (59.4, 61.5)

	AUC	0.62	0.62	0.62	0.62	0.64	--	0.64
	Kappa	0.18	0.175	0.006	0.065	0.198	--	0.198
	F-2 Score	0.47	0.45	0.82	0.77	0.47	--	0.47

Severe Versus Non-Severe Validation

The severe versus non-severe model was reconstructed using variables common to the PAHO Mexico and the SINAVE 2016-2019 Yucatan datasets. The seven most important variables in order were plasma leakage, bleeding, petechia, persistent vomiting, neurological, hemodynamic, and age. This is the same order of importance as the original severe versus not-severe dengue model save for "intense abdominal pain" which was not included in the overall Mexico dataset. This reconstructed model is referred to as the "Internal" model group in Table 5. As depicted in Table 5, the internal model had slightly lower specificity, sensitivity, AUC, and accuracy than the original severe vs non-severe model. Furthermore, when performing temporal validation by predicting against the PAHO Mexico data that was filtered to just Yucatan for the years 2020-2021, the model performance decreased such that the base logistic regression and random forest both had accuracies of 76.6 and as AUC of 0.85 and 0.84, respectively. When performing external validation by using the internal model to predict against all other Mexican states from 2017-2019, the accuracy decreased slightly less compared to the temporal validation, but the AUC score decreased further to 0.83 for both random forest and logistic regression. Both validation steps led to a decrease in specificity and sensitivity, but the external validation to all other states had the lowest sensitivities across models.

Table 5: Assessment Metrics for Severe Vs Non-Severe Temporal and External Validation

Group	Metric	Logistic Regression	Logistic Regression: Youden's Index	Logistic Regression: Cost Function	Random Forest	Alternate Tune Random Forest
Internal Model	Threshold and Hyperparameters	DT=0.5	DT=0.38	DT=0.202	Mtry= 2.83 NS=1	Mtry= 2 NS=9
	Sensitivy	73.6 (65.4, 81.9)	80.9 (73.6, 88.3)	86.4 (80, 92.8)	77.3 (69.4, 85.1)	76.4 (68.4, 84.3)
	Specificity	94.3 (90.2, 98.4)	89.4 (84, 94.9)	74.8 (67.1, 82.5)	91.9 (87, 96.7)	93.5 (89.1, 97.9)
	PPV	92 (86.4, 97.7)	87.3 (80.8, 93.7)	75.4 (67.9, 82.9)	89.5 (83.3, 95.6)	91.3 (85.5, 97.1)
	NPV	80 (73.5, 86.5)	84 (77.7, 90.3)	86 (79.4, 92.6)	81.9 (75.5, 88.3)	81.6 (75.2, 88)
	Accuracy	84.5 (79.3, 88.9)	85.4 (80.2, 89.7)	80.3 (74.6, 85.2)	85 (79.7, 89.3)	85.4 (80.2, 89.7)
	AUC	0.91	0.91	0.91	0.88	0.88
	Kappa	0.687	0.706	0.607	0.696	0.705
	F-2 Score	0.77	0.82	0.84	0.79	0.79
Temporal Validation	Sensitivy	70.1 (61, 79.2)	71.1 (62.1, 80.2)	81.4 (73.7, 89.2)	70.1 (61, 79.2)	70.1 (61, 79.2)

	Specificity	89.6 (80.9, 98.2)	89.6 (80.9, 98.2)	64.6 (51.1, 78.1)	89.6 (80.9, 98.2)	89.6 (80.9, 98.2)
	PPV	93.2 (87.4, 98.9)	93.2 (87.5, 99.0)	82.3 (74.7, 89.9)	93.2 (87.4 ,98.9)	93.2 (87.4 ,98.9)
	NPV	59.7 (48.4, 71.1)	60.6 (49.2, 71.9)	63.3 (49.8, 76.8)	59.7 (48.4, 71.1)	59.7 (48.4, 71.1)
	Accuracy	76.6 (68.8, 83.2)	77.2 (69.5, 83.8)	75.9 (68.1, 82.6)	76.6 (68.8, 83.2)	76.6 (68.8, 83.2)
	AUC	0.85	0.85	0.85	0.84	0.85
	Kappa	0.53	0.542	0.458	0.53	0.53
	F-2 Score	0.74	0.75	0.82	0.74	0.74
External Validation	Sensitivy	65.5 (64.9, 66.1)	69.2 (68.6, 69.9)	76.4 (75.8, 76.9)	66.3 (65.7, 67.0)	66.1 (65.4, 66.7)
	Specificity	90.7 (90.5, 91)	89.6 (89.3, 89.9)	79.2 (78.8, 79.5)	90.5 (90.2, 90.7)	90.5 (90.3, 90.8)
	PPV	76.1 (75.5, 76.7)	75 (74.4, 75.6)	62.3 (61.7, 62.9)	75.8 (75.2, 76.4)	75.8 (75.2, 76.5)
	NPV	85.4 (85.1, 85.7)	86.6 (86.3, 86.9)	88.1 (87.8, 88.4)	85.6 (85.3, 85.9)	85.5 (85.2, 85.8)
	Accuracy	82.9 (82.6, 83.2)	83.3 (83, 83.6)	78.3 (78.0, 78.6)	83 (82.7, 83.2)	82.9 (82.6, 83.2)

	AUC	0.83	0.83	0.83	0.83	0.83
	Kappa	0.585	0.601	0.523	0.588	0.587
	F-2 Score	0.67	0.7	0.73	0.68	0.68

Internal: results based on the train/validate/test split from the 2016-2019 SINAVE Yucatan dataset using the top 7 most important variables that are also in the overall Mexico dataset. **Temporal:** using the same trained model from the internal model to predict on Yucatan patients from the PAHO Mexico dataset for the years 2020-2021. **External:** using the same trained model from the internal model to predict patients across all Mexican states other than Yucatan from the PAHO dataset in 2017-2019.

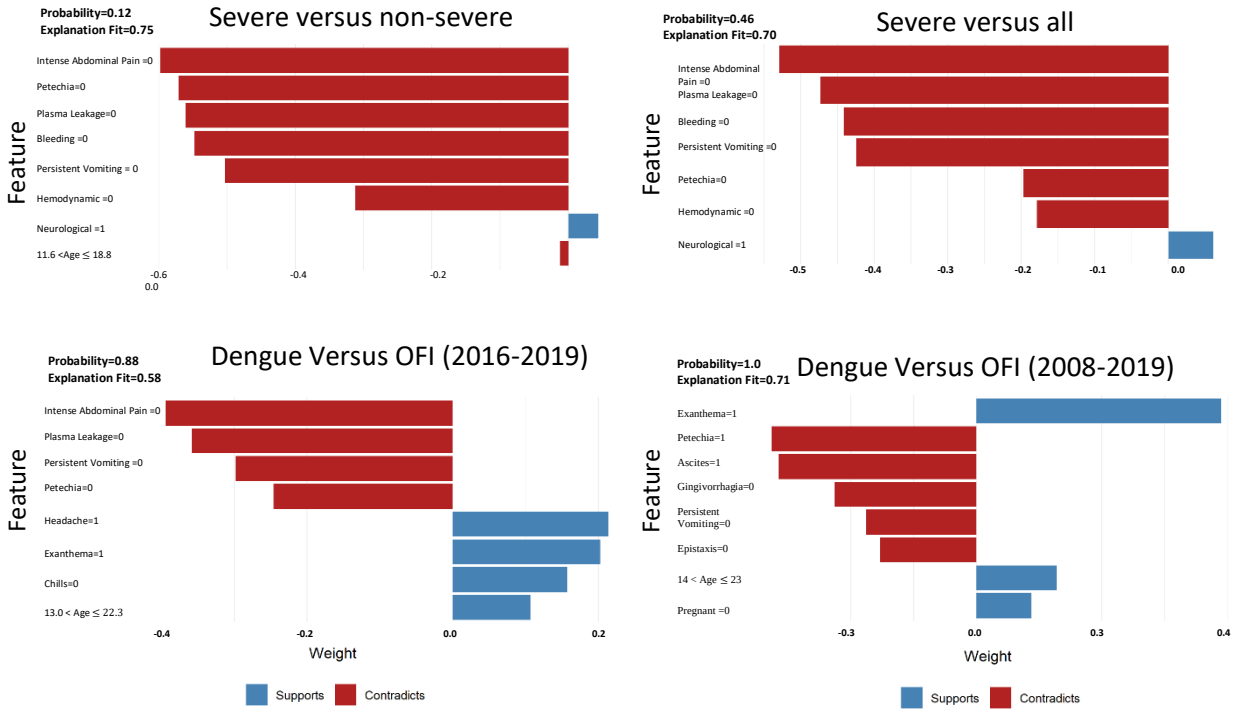
LIME Analysis

At the end of the discussion section, we provide a case study example for "Patient A" to show how random forest operates at the individual level. Figure 8 illustrates the estimated feature weights from LIME analysis for this patient across all four model groups. The probability indicates proportion of the 1000 trees that voted for the particular outcome in each model and the explanation fit is the R^2 for the explainable ridge regression model. The severe versus non-severe model incorrectly classifies "Patient A" as not having severe dengue with the only feature with a weight towards having the outcome being neurology=1. The severe versus all model has similar results. The patient is correctly classified as having dengue in both dengue models. Having a headache, having exanthema, not having chills, and being between 13-22.3 years old were features with positive weights for the outcome in the 2016-2019 dengue model. Being 14-23 years old, not being pregnant, and especially having exanthema are features estimated to contribute to the dengue diagnosis of the 2008-2019 dengue model. The explanation fits are around 0.7 for most of the LIME models but is only 0.58 for the 2016-2019 dengue model.

Figure 9 depicts the relative weights of features for having or not having the outcome across 13 randomly selected patients from the test sets of the Severe vs non-severe and dengue vs OFI (2016-2019) models. Feature values that seem to indicate having severe disease across the 13 patients are bleeding (=1), intense abdominal pain (=1), plasma leakage (=1), and persistent vomiting (=1). Feature values that

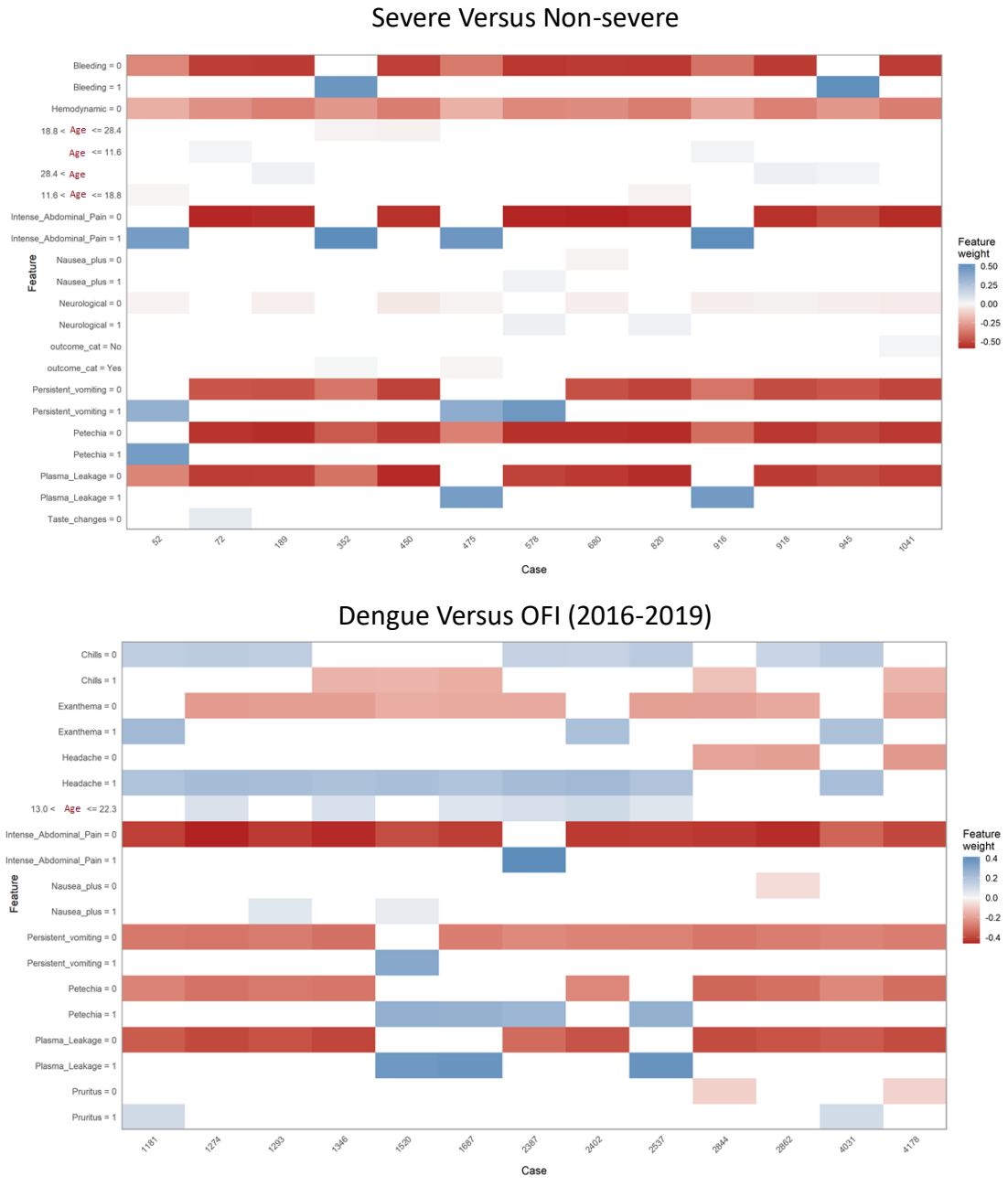
appear to indicate having dengue are chills (=0), Exanthema (=1), Headache (=1), Intense abdominal pain (=1), persistent vomiting (=1), petechia (=1), age (13-22.3 years old), and plasma leakage (=1).

Figure 8: LIME Estimation of Variable Feature Weights for Prediction of "Patient A"



Red (negative) feature weights are estimated to influence the model to vote against the outcome for "Patient A" whereas the blue (positive) feature weights support classification of the outcome. Probability stands for the proportion of votes among the 1000 trees that classified "Patient A" as having the outcome. Explanation Fit refers the R^2 value of the explainable ridge regression in its assessments of feature weights—a higher number indicates a better fit.

Figure 9: LIME Feature Weights Across 13 randomly Selected Patients in the Target Population



Estimated weights of features calculated by the R package LIME across random samples of 13 different patients from the test sets in the severe versus non-severe (top) and dengue versus OFI 2016-2019 (bottom) modeling groups. Red is an indicator for a feature that contributes towards a negative outcome whereas blue indicates contribution towards a positive outcome. Darkness of color reflects the magnitude of the relative weight.

Discussion

Epidemiolocal and clinical features

The occurrence of dengue outbreaks shows a seasonal pattern, related to the rainy seasons that begin in summer. There is an important variation in the interannual incidence with peaks of occurrence maintained between the years 2009 to 2015, followed by an interval of low occurrence, which increases again towards 2019. Many factors influence the occurrence of outbreaks, and this pattern of intervals with peaks of different magnitude has been described in other endemic regions. It is worth mentioning the introduction of Zika in 2016 has been linked by some authors with the lowest occurrence of dengue cases (Mugabe, 2021) Furthermore, when evaluating the biannual space-time Poisson clustering of dengue cases, there is an evident shift from Western Yucatan from 2008-2013 to Eastern Yucatan from 2014-2019. Further analysis regarding potential environmental factors behind this spatial shift in cases rate is needed to help explain the phenomenon. Additionally, an increase in the proportion of severe forms within an outbreak may be observed towards the end of the epidemic period. This has been reported by other authors who postulate the temporary selection of viral variants of greater virulence that could explain the increase in clinical severity (Rodriguez-Roche, 2016).

When compared to data pertaining to all of Mexico, our Yucatan dataset demonstrates similar trends with respect to serotype, sample prevalence, and severity of outcome for 2018-2019. However, the difference in mean age of those presenting to clinic with suspected dengue across the datasets indicates that there are some differences between the overall Mexico population and that specific to Yucatan, Mexico.

Most of the variables found to be important in predicting dengue and severe dengue, aside from age, petechia, and exanthema, are warning signs according to the 2009 WHO classification. We expected this result given that the WHO's list of warning signs and severe symptoms were formulated to help clinicians identify patients with severe dengue or at risk of developing severe dengue. The age categories more affected by dengue and also by severe forms of the disease in our study population were children

and younger. This finding is in agreement with reports from other countries in America and Southeast Asia (Teixeira, 2008; Halstead, 2006). Diabetes and hypertension were the most frequent comorbidities, and associated with severe forms of dengue, as described in other studies (Pang, 2012). Symptoms secondary to plasma extravasation, such as severe abdominal pain and persistent vomiting, were the most frequent warning signs, followed by hemorrhagic manifestations and hepatomegaly. The most frequent manifestations in seriously ill patients were hemodynamic instability and digestive bleeding.

Modeling Results

Logistic regression proved to generally outperform random forest with respect to prognostic modeling and had mixed results for diagnostic modeling. For both severe model groups, logistic regression using a cost index or Youden's index had higher accuracy and AUC, and had higher sensitivity. When predicting dengue versus OFI, the random forest models had higher AUC, accuracy, and Kappa scores for the years 2008-2019, but the default logistic regression performed better for the years 2016-2019. Both of the latter two models had low sensitivity and specificity. We hypothesized that random forest would perform better than logistic regression based on literature review, but there are possible explanations as to why this did not occur. See limitations for further discussion.

Random forest models with tuned mtry and nodesize hyperparameters were largely similar to the default model. The logistic regression results were relatively similar when using the default or Youden's index as a decision threshold. Using a cost index and minimum sensitivity, based on the validation dataset, led to higher sensitivities at the cost of specificity. The choice of which decision threshold to use depends on the particular question. When higher sensitivity at the cost of specificity is desired, the cost index may be of interest. When balancing specificity and sensitivity, the Youden's index may a better choice.

Compared to the new WHO classification validation studies in Table 1, our prognostic and diagnostic model results generally had lower sensitivity and negative predictive values with higher specificities and positive predictive values. When comparing our diagnostic models against those found in literature review in Table 2, our diagnostic models underperform in all areas whereas our prognostic

models appeared to outperform in all areas. However, comparing models from different studies is difficult and should be taken lightly without proper validation due to potentially significant differences in patient population, hospital protocol, and variables considered.

Moreover, only our severe modeling groups had results where both sensitivity and specificity were greater than 80%. It would thus appear that all of our models are potentially useful for modeling severe dengue, especially the logistic regression models. However, our diagnostic models performed poorly and would not be of much use in the clinical setting.

Model performance for predicting severe versus non-severe dengue among those with confirmed dengue further decreased when performing temporal and external population validation. While logistic regression with Youden's decision threshold led to an accuracy around 80% for the temporal and external validation steps, respectively, the sensitivity for both was below 80%. A decrease in performance may be expected given that we are testing against populations external to that of our main study. Additionally, the internal model was based on the SIANVE Yucatan dataset which had far more variables for consideration than PAHO's Mexico validation data. Notably, the overall Mexico dataset did not have a variable for intense abdominal pain which was our most important variables for predicting severity in the Yucatan dataset. Rebuilding the severe vs non-severe model to match the variables in the PAHO Mexico dataset may have affected the ability to apply of our model to predict external populations. Obtaining data with the same structure and variety as our SINAVE Yucatan dataset for other regions within Mexico would help further assess the ability to generalize our severity models as well as dengue vs OFI models to other populations within Mexico.

Apart from central fit measurements such as AUC and accuracy, it is important to consider the relative cost of false negatives and false positives. In clinical medicine, falsely identifying a case as a control can have drastic consequences for the patient. In this context of our study, diagnosing a patient as having the flu when they in fact have dengue with warning signs hinders the clinician's capacity to administer potentially lifesaving preventative treatment. As such, clinical models are most beneficial when emphasizing higher sensitivity at the cost of sensitivity. We were able to address this cost-sensitive

problem by tuning decision thresholds in our logistic regressions according to a cost function that doubles the impact of false negatives in relation to false positive. While there are some studies investigating the potential of calibrating random forest probabilities and, or performing threshold tuning, it is not appropriate in regular random forest models as the "probability" output is actually the proportion of all trees that ended up with the outcome (Dankowski, 2016; Kull, 2017; Boström, 2008; Reis, 2018). There are other ways to address this problem by employing modified cost sensitive or weighted random forest (Yang, 2009; Devi, 2019; Gajowniczek, 2020). Applying one of these methods to our above random forest models might help improve sensitivity to a more acceptable level, but likely at the cost of specificity.

Interpretation of Random Forest in The Clinical Setting

Regardless of the performance we obtained in this study, it is important for clinicians and policy makers to better understand and interpret random forest models in the context of the individual. To that end, we also seek to end our discussion by explaining the application and potential interpretation of applying random forest models on the individual level. The following details pertains to a single individual, "Patient A", isolated from the test set of the severe versus non-severe model group. Specifically, we examined each row in the test set in order of appearance until we found a row that was unique and traceable to the original data frame. Some attributes have been altered for privacy reasons. In 2019, a 13-year-old male in Yucatan, Mexico, presented the sudden onset of fever (39ºC). The illness escalated and the patient was admitted to a hospital three days post-onset as a dengue suspect. "Patient A" was found to be experiencing headaches, exanthema, neurological symptoms, and nausea related symptoms at some point throughout his follow up and he was found to be negative for all other important binary predictors. The diagnosis was confirmed with PCR 7 days after initial onset and the patient was classified as having dengue with warning signs.

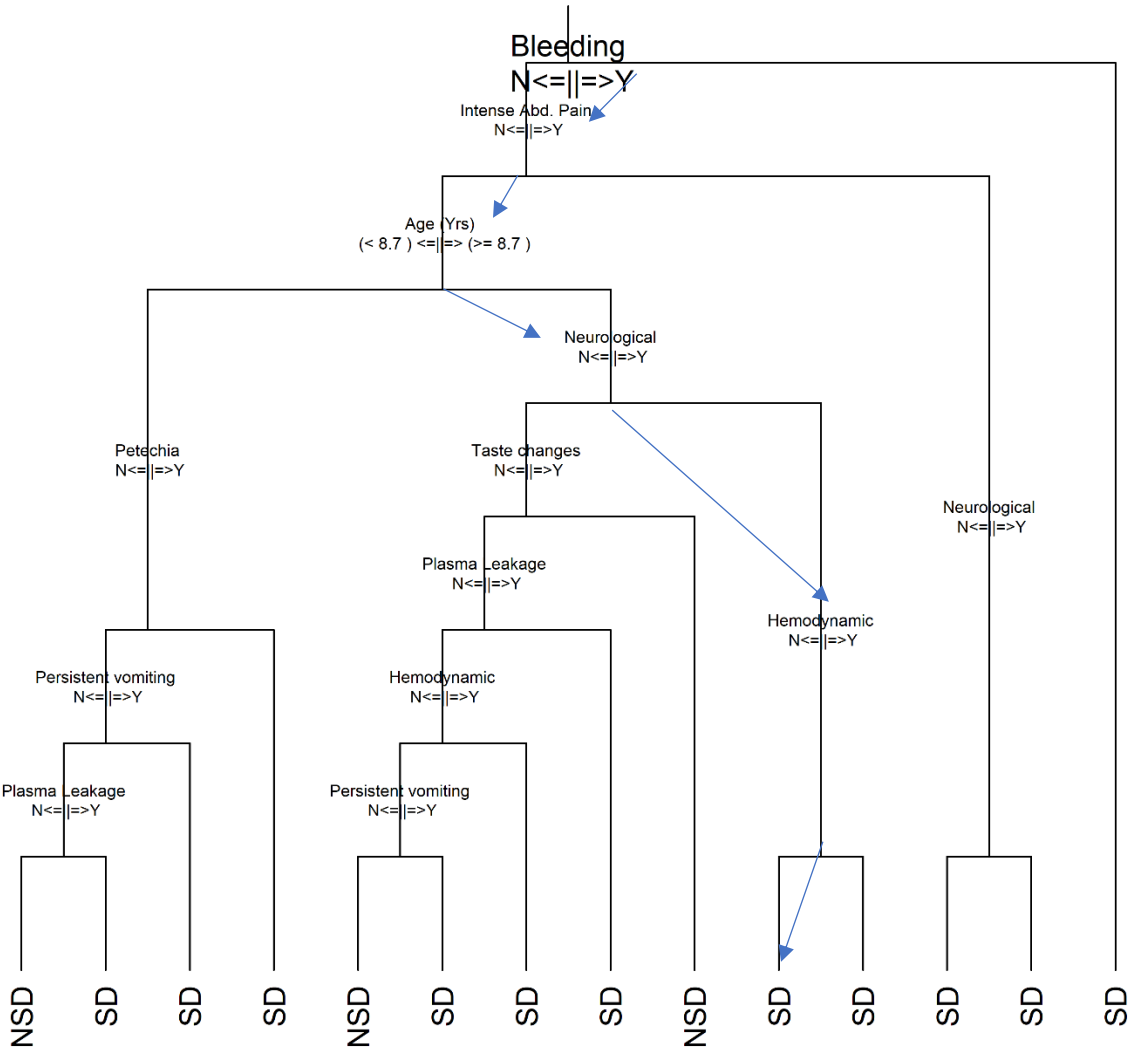
As random forest's final verdict is based on majority vote, the patient was incorrectly classified as not having severe dengue with only 12% and 46% of the 1000 trees voting for the outcome in the severe versus non-severe and severe vs all else models, respectively. The patient was correctly identified as

having dengue with 88.3% and 99.8% of the trees voting for dengue in the 2016-2019 and 2008-2019 dengue models, respectively. Figure 10 is an example of one such tree for the severe versus non-severe model group. We can follow the dendrogram in the following order: the patient does not have bleeding, is greater than 8.7 years old, does have neurological symptoms, and does not have hemodynamic changes. Hence, this tree is 1 out of only 120 trees that correctly labeled the patient as having severe dengue. Note that the performance of the models for singular individuals does not provide any meaningful information about the overall accuracy and reliability of the models.

LIME analysis is a way of potentially gaining further insight with regards to what features within the patient data led to the final verdict of the model. LIME has been used in previous clinical research to help provide individual interpretation. Tajgardoon et al. found that a panel of physicians had high levels of agreement when LIME was used in real patients but had poor agreement when LIME was used on fake data (Visani , Tajgardoon, Katuwal). However, other studies have found LIME lacks consistency and does not always meet Ribeiro et al.'s claims of faithfulness, linearity, and interpretability (Goode, 2021). Thus, LIME results should not be taken as complete truth, but may still be useful as supplemental information for clinicians and health experts.

The feature weights in Figure 8 pertaining to the LIME analysis of "Patient A" are non-surprising. All features considered important in Boruta analysis with harmful odds ratios on pre-split logistic regression contribute towards having the outcome when present in "Patient A" such as Neurological, headache, and exanthema. The feature weights for the patient's age contributing to having dengue also mirrors the relatively harmful odds ratios for patients in younger age categories. The explanation fit of the dengue versus non-dengue 2016-2019 is poor, but the fitness for the other modeling groups is relatively higher with R^2 around 0.7-0.75. When applying LIME analysis across 13 randomly selected test patients, we see that the general trend of what features contribute to and against the relevant outcome reflects what we would expect based on the global feature importance and odds ratios.

Figure 10: Individual Severe Versus Non-Severe Random Forest Tree For "Patient A"



Structure of one tree from the "Alternate Tuned" random forest model regarding severe (SD) versus non-severe dengue (SD). Left side is either negative (N) for binary variable or is less than continuous variable. Arrows show the course of "Patient A" on this particular tree (number 346 out of 1000) leading to the vote for severe dengue. The OOB Error, the error of the out of bag training set from this tree across all iterations, was 0.1.

Conclusion

Timely clinical management is critical to ensure the health of patients progressing into severe forms of dengue disease. Thus, there is a need for early detection of patients with dengue and early prognosis of those likely to progress into severe dengue. After performing diagnostic and prognostic modeling with both logistic regression and random forest, we conclude that both random forest and logistic regression may be useful for predicting severe dengue (confirmed severe dengue and or dengue

with warning signs) versus non-severe dengue. However, our models were not effective in detecting dengue versus other febrile illnesses in our patient population. The performance of our severe versus non-severe models lowered when validating against novel years and locations, thus potentially questioning its ability to be utilized in external populations.

We additionally evaluated model performance using hyperparameters and threshold values. We found that ransom forest hyperparameter led to fairly similar results as the default settings. We also found that using a cost function that penalized false negatives with logistic regression penalized les to an increase in the sensitivity at the expense of specificity, which may be more relevant in the clinical setting.

Furthermore, we identified a set of variables that were most important in the models such as bleeding and intense abdominal pain. We then estimated the odds ratio of exposure among the diseased to approximate the global association between the features. These global associations seemed to corroborate our model agnostic explanation analysis at the individual level, but further study is needed to properly assess the actual reliability of LIME for our models.

Limitations

We made several educated assumptions throughout this study that have the potential to impact our results. Based on input from a physician with dengue expertise in the region, we assumed that all positive clinical symptoms, comorbidities, and signs were correctly recorded and that all other values were indicative as negative for that particular covariate. This assumption affected the data fed into the models as well as all relevant tables and figures. We also performed ROSE balancing on the training set for the severe versus all and dengue versus OFI (2016-2019) modeling groups due to drastically imbalanced dataset. Though the stratified distributions of the training datasets before and after balancing were similar, ROSE balancing led to some nonsensical negative age values in the training set which may have had an effect on modeling results.

One particular shortcoming in the data that may help explain the performance of our models is the lack of laboratory testing that was available to include in the model. Clinical labs are fundamental to the diagnosis of dengue and were used as central predictor variables in most of the models we encountered

during literature review. We used hematocrit and leucocyte lab values in some of our models with noticeable improvement, especially in the random forest models. However, with missing rates around 40-50% for each lab variable, it would be difficult to defend using these models in real life scenarios and thus we did not include them in this paper.

Additionally, the relative superior performance of the logistic regression models in modeling severity may have been due to small sample sizes. While many studies have documented better performance of random forest models for classification, some of these studies have also noted the importance of using enough variables and having a large events per variable (EPV) ratio when performing random forest modeling (Christodoulou, 2019; Heinze, 2018; Sanchez-Pinto, 2018). Notably, van der Ploeg et al found that logistic regression AUC stabilized when the EPV is around 20-50 whereas random forest models demonstrate instability even when EPV is greater than 200 (van der Ploeg, 2014). Considering that the training data contains only 60% of the observations, the dengue model for years 2008-2019 is the only one with an EPV significantly higher than 200. With this in mind, we plan on investigating the potential impact that restricting our analysis to fewer variables in order of importance. While this might lose predictive information from variables dropped, it would increase our EPV and may produce more statistically efficient results.

The LIME package in R has many different options. Goode et al. found that these hyperparameters can affect the results for feature weights (Goode, 2021). Furthermore, the instability of LIME poses some problems for use in the clinical setting (Goode, 2021; Visani, 2020). Further steps to improve the local explanations include tuning the hyperparameter options or employing alternative versions of LIME such as Visani et al.'s OptiLIM (Visani, 2020).

Supplemental Section

Basic Review of Models

Logistic regression is performed on the training set where the Beta coefficients in the following equation below are estimated according to the maximum likelihood method (Peng, 2002). Then, new data, x predictor values, are provided from the test set and the probability that Y=1 given the predictors is estimated. A decision threshold, set by default to be 50%, is used to determine if the final classification for each observation of the test set is the outcome of interest or not based on the derived probability.

Supplemental Equation One

$$Log\left(\frac{P}{1-P}\right) = B_0 + B_{x1} + B_{x2} + \cdots B_n$$

x's=predictors. n= total number of predictors .

P=probability(Y=outcome | $B_0 + B_{x1} + B_{x2} + \cdots B_n$)=$\frac{e^{B_0+B_{x1}+B_{x2}+\cdots B_n}}{1+e^{B_0+B_{x1}+B_{x2}+\cdots B_n}}$

Decision trees generally operate by using the entire train dataset to construct a tree that is used to determine the classification of each observation in the test set (Patel, 2018). The variables that are assigned at each split location, called a node, is determined by an assigned splitting criterion such as information gain, gain ratio, or gini index (Patel, 2018). The root node, the first one in the tree, is determined by the variable found to have the lowest gin impurity, or highest information gain (Kingsford, 2008). Nodes after the root node are called the branch nodes and the nodes with the final classification are called the leaf nodes (Patel, 2018). Pruning is often performed to limit the growth of the tree by some minimal difference in the splitting criteria to prevent overfitting (Kingsford, 2008). Each leaf node contains a probability distribution of the possible outcomes; this probability is then used to determine the final classification for each observation (Kingsford, 2008). There are numerous types of decision tree algorithms such as C4.5 and CART that operate in slightly different ways (Patel, 2018).

Ensemble tree methods, such as random forest, are often more accurate than decision tree because they employ multiple trees (Kingsford, 2008). Random forest utilizes bagging, a process in which the training data is sampled with replacement to derive an in-bag training dataset (Breiman, 2001). The in-bag training data is used to construct the tree and the out of bag dataset, roughly 1/3 of the original training data, is used to calculate the out of bag error (OOBE) (Breiman, 2001). This process is repeated for as many trees created by the algorithm (Breiman, 2001). Unlike decision tree, only a certain number of features are randomly selected for consideration at each node using the determined selection criteria. Additionally, the trees are not pruned, but the minimum node size determines the smallest size of a leaf node allowed in the tree (Breiman, 2001). Once all the trees are constructed, the aggregated OOBE gives an estimate of the accuracy based on the out of bag datasets (Braiman, 2001). Then, each observation from the test set is run through all the trees created by the trained random forest and the outcome with the majority vote, the highest proportion of trees, is the final classified outcome for that observation (Breiman, 2001).

Supplemental Tables

Supplemental Table 1: Model Comparison Metrics

Metrics	Formula	Description
Sensitivity	$\frac{tp}{tp+fn}$	Proportion of positive outcome correctly identified (also known as recall)
Specificity	$\frac{tn}{tn+fp}$	Proportion of negative outcome correctly identified
Positive Predictive Value (PPV)	$\frac{tp}{tp+fp}$	The number of true positives divided by the number of positive calls (also known as precision)
Negative Predictive Value (NPV)	$\frac{tn}{tn+fn}$	The number of true negatives divided by the number of negative calls.

Accuracy	$\frac{tp+tn}{tp+fp+tn+fn}$	Proportion of total calls correctly classified
ROC-AUC	--	Area under the ROC curve
Kappa	$\frac{P_0 - P_C}{1-P_C}$ P_0=observed agreement (accuracy) P_c=chance agreement.	Value indicating strength of agreement. Landis and Koch proposed that a kappa coefficient ≤0 is poor, between 0.1-0.2 is slight, between 0.21-0.4 is fair, between 0.41-0.60 is moderate, between 0.61-0.80 is substantial, and between 0.81-1 is almost perfect (Landis, 1977; Sim, 2019). However, the interpretation is arbitrary and can vary.
F-2	F-measure= $\frac{(B^2+1)*Precision*Recall}{B^2*Precision+Recall}$ F-2 = F-Measure	The F-score is a balance between the recall (sensitivity) and precision (PPV). The F-2 Score puts greater weight on sensitivity which is typically used when

Formulas and descriptions obtained from Sokolova, 2006 and Sim, 1977

Supplemental Table 2: Distribution and Univariate Risk for Dengue Across Variables From 2008-2019

Category	Variable	Dengue N (%)	Other N (%)	Risk Ratio Dengue Vs OFI: RR (CI)	Fischer Exact: P-value
Demographics	Age <14	5454 (26.7%)	6220 (28.0%)	1	--
	Age 15-28	8023 (39.3%)	7049 (31.7%)	1.14 (1.11, 1.17) **	<0.001
	Age 29-42	3609 (17.7%)	4579 (20.6%)	0.94 (0.92, 0.97) **	<0.001
	Age 43-57	2285 (11.2%)	2850 (12.8%)	0.95 (0.92, 0.99) **	0.008
	Age >58	1024 (5.0%)	1549 (7.0%)	0.85 (0.81, 0.90) **	<0.001
	Sex (Male)	9948 (48.8%)	10497 (47.2%)	1.03 (1.01, 1.05) **	0.001
Comorbidities	Hypertension	183 (0.9%)	160 (0.7%)	1.12 (1.01, 1.23) **	0.045

	Peptic Ulcer	26 (0.1%)	25 (0.1%)	1.07 (0.81, 1.40)	0.676
	Diabetes	171 (0.8%)	225 (1.0%)	0.90 (0.81, 1.01)	0.069
	Bleeding Disorder	27 (0.1%)	37 (0.2%)	0.88 (0.66, 1.18)	0.383
	Pregnant	365 (1.8%)	611 (2.7%)	0.78 (0.72, 0.84) **	<0.001
	Immunosuppression	26 (0.1%)	79 (0.4%)	0.52 (0.37, 0.72) **	<0.001
	Liver cirrhosis	18 (0.1%)	56 (0.3%)	0.51 (0.34, 0.76) **	<0.001
Common Acute Findings	Fever	20393 (100.0%)	22244 (100.0%)	1.20 (0.41, 3.50)	1
General Symptoms	Positive tourniquet	2028 (9.9%)	862 (3.9%)	1.52 (1.48, 1.56) **	<0.001
	Retroorbital pain	13674 (67.0%)	14191 (63.8%)	1.08 (1.06, 1.10) **	<0.001
	Arthralgia	17279 (84.7%)	18453 (82.9%)	1.07 (1.04, 1.10) **	<0.001
	Myalgia	18569 (91.0%)	20158 (90.6%)	1.03 (0.99, 1.06)	0.122
Common Respiratory Symptoms	Petechia	3208 (15.7%)	1201 (5.4%)	1.62 (1.58, 1.65) **	<0.001
	Exanthema	6852 (33.6%)	4755 (21.4%)	1.35 (1.33, 1.38) **	<0.001
	Headache	19153 (93.9%)	20523 (92.3%)	1.15 (1.10, 1.20) **	<0.001
Warning Signs	Ascites	230 (1.1%)	53 (0.2%)	1.71 (1.61, 1.81) **	<0.001
	Pleural effusion	140 (0.7%)	56 (0.3%)	1.50 (1.37, 1.64) **	<0.001

	Ecchymosis	483 (2.4%)	220 (1.0%)	1.45 (1.38, 1.52) **	<0.001
	Gingivorrhagia	1030 (5.1%)	480 (2.2%)	1.45 (1.40, 1.50) **	<0.001
	Persistent vomiting	1657 (8.1%)	1022 (4.6%)	1.32 (1.28, 1.36) **	<0.001
	Hematomas	206 (1.0%)	124 (0.6%)	1.31 (1.20, 1.42) **	<0.001
	Epistaxis	674 (3.3%)	430 (1.9%)	1.29 (1.23, 1.35) **	<0.001
	Intense Abdominal Pain	2050 (10.1%)	1380 (6.2%)	1.28 (1.24, 1.32) **	<0.001
Severe Symptoms	Hematemesis	220 (1.1%)	118 (0.5%)	1.36 (1.26, 1.48) **	<0.001
	Melena	140 (0.7%)	118 (0.5%)	1.14 (1.01, 1.27) **	0.039

** Indicates significant univariate risk ratio compared to reference. Risk ratio references for sex and age were female patients and those $\leq$ 14 years old, respectively.

Supplemental Table 3: Distribution of Dengue Across Dengue Prognosis Outcomes From 2016-2019

Category	Variable	Non-Severe Dengue N (%)	Dengue with Warning Signs N (%)	Severe Dengue N (%)	Other Febrile Illness N (%)	Risk Ratio Dengue Vs OFI: RR (CI)	Risk Ratio SD/DSW Vs NSD RR (CI)
Demographics	Age <14	158 (25.7%)	194 (39%)	20 (37%)	837 (26%)	1	1
	Age 15-28	282 (45.9%)	207 (41.6%)	11 (20.4%)	1149 (35.7%)	0.98 (0.88, 1.10)	0.76 (0.66, 0.86) **
	Age 29-42	95 (15.4%)	55 (11%)	14 (25.9%)	661 (20.5%)	0.65 (0.55, 0.76) **	0.73 (0.60, 0.89) **

	Age 43-57	54 (8.8%)	20 (4%)	2 (3.7%)	410 (12.7%)	0.51 (0.41, 0.64) **	0.50 (0.35, 0.72) **
	Age 58-100	26 (4.2%)	22 (4.4%)	7 (13%)	160 (5%)	0.83 (0.65, 1.06)	0.92 (0.70, 1.19)
	Sex	283 (46%)	256 (51.4%)	26 (48.1%)	1477 (45.9%)	1.08 (0.98, 1.19)	1.11 (0.99, 1.26)
	Age: Mean (SD)	24.53 (15.38)	20.55 (15.05)	26.81 (20.24)	26.3 (16.81)	--	--
Comorbidities	Hypertension	1 (0.2%)	6 (1.2%)	3 (5.6%)	31 (1%)	0.92 (0.53, 1.57)	1.92 (1.55, 2.38) **
	Bleeding Disorder	2 (0.3%)	7 (1.4%)	2 (3.7%)	31 (1%)	0.98 (0.59, 1.64)	1.74 (1.31, 2.32) **
	Diabetes	5 (0.8%)	13 (2.6%)	6 (11.1%)	63 (2%)	1.04 (0.74, 1.46)	1.70 (1.37, 2.10) **
	Pregnant	67 (10.9%)	24 (4.8%)	3 (5.6%)	235 (7.3%)	1.08 (0.90, 1.29)	0.59 (0.42, 0.81) **
	Immunosuppression	13 (2.1%)	2 (0.4%)	0 (0%)	38 (1.2%)	1.06 (0.69, 1.64)	0.28 (0.08, 1.02)
	Liver cirrhosis	1 (0.2%)	0 (0%)	0 (0%)	14 (0.4%)	0.25 (0.04, 1.66)	0
	Peptic Ulcer	1 (0.2%)	0 (0%)	0 (0%)	4 (0.1%)	0.75 (0.13, 4.34)	0
Common Acute Findings	Fever	614 (99.8%)	497 (99.8%)	54 (100%)	3214 (99.9%)	0.67 (0.23, 1.95)	0.95 (0.24, 3.79)
	Nasal Congestion	20 (3.3%)	12 (2.4%)	2 (3.7%)	116 (3.6%)	0.85 (0.63, 1.14)	0.87 (0.58, 1.30)

	Taste changes	33 (5.4%)	13 (2.6%)	5 (9.3%)	185 (5.8%)	0.80 (0.63, 1.03)	0.74 (0.51, 1.07)
General Symptoms	Positive tourniquet	11 (1.8%)	128 (25.7%)	12 (22.2%)	173 (5.4%)	1.86 (1.64, 2.12) **	2.29 (2.10, 2.49) **
	AST ALT	1 (0.2%)	0 (0%)	3 (5.6%)	13 (0.4%)	0.88 (0.37, 2.08)	1.59 (0.90, 2.81)
	Diaphoresis	31 (5%)	44 (8.8%)	4 (7.4%)	223 (6.9%)	0.98 (0.81, 1.19)	1.31 (1.09, 1.58) **
	Chills	114 (18.5%)	120 (24.1%)	11 (20.4%)	709 (22%)	0.96 (0.85, 1.08)	1.17 (1.02, 1.34) **
	Arthritis	27 (4.4%)	25 (5%)	6 (11.1%)	116 (3.6%)	1.27 (1.02, 1.57) **	1.14 (0.89, 1.46)
	Back ache	152 (24.7%)	142 (28.5%)	12 (22.2%)	757 (23.5%)	1.11 (0.99, 1.24)	1.09 (0.95, 1.24)
	Arthralgia	505 (82.1%)	421 (84.5%)	43 (79.6%)	2620 (81.4%)	1.08 (0.95, 1.24)	1.08 (0.91, 1.28)
	Myalgia	582 (94.6%)	474 (95.2%)	50 (92.6%)	2994 (93.1%)	1.26 (1.00, 1.58)	1.03 (0.78, 1.36)
	Petechia	28 (4.6%)	152 (30.5%)	25 (46.3%)	253 (7.9%)	1.83 (1.63, 2.05) **	2.21 (2.01, 2.44) **
	Exanthema	224 (36.4%)	201 (40.4%)	19 (35.2%)	755 (23.5%)	1.63 (1.48, 1.80) **	1.08 (0.95, 1.22)
	Headache	577 (93.8%)	471 (94.6%)	51 (94.4%)	2771 (86.1%)	2.15 (1.71, 2.69) **	1.08 (0.82, 1.42)

	Pruritus	102 (16.6%)	94 (18.9%)	10 (18.5%)	312 (9.7%)	1.60 (1.42, 1.80) **	1.08 (0.93, 1.26)
	Polyarthralgia	29 (4.7%)	17 (3.4%)	6 (11.1%)	99 (3.1%)	1.31 (1.04, 1.64) **	0.93 (0.68, 1.27)
	Conjunctivitis	49 (8%)	26 (5.2%)	6 (11.1%)	137 (4.3%)	1.43 (1.19, 1.71) **	0.83 (0.63, 1.09)
	Retroorbital pain	444 (72.2%)	357 (71.7%)	31 (57.4%)	2214 (68.8%)	1.09 (0.98, 1.22)	0.95 (0.84, 1.09)
Common Respiratory Symptoms	Dyspnea[7]	1 (0.2%)	4 (0.8%)	3 (5.6%)	27 (0.8%)	0.86 (0.47, 1.58)	1.86 (1.42, 2.43) **
	Cough	21 (3.4%)	29 (5.8%)	4 (7.4%)	211 (6.6%)	0.75 (0.59, 0.96) **	1.31 (1.05, 1.64) **
	Pharyngitis	25 (4.1%)	23 (4.6%)	3 (5.6%)	169 (5.3%)	0.86 (0.68, 1.11)	1.08 (0.82, 1.43)
	Adenomegaly	7 (1.1%)	5 (1%)	1 (1.9%)	45 (1.4%)	0.84 (0.52, 1.36)	0.98 (0.54, 1.76)
Warning Signs	Intense Abdominal Pain	4 (0.7%)	222 (44.6%)	21 (38.9%)	216 (6.7%)	2.27 (2.05, 2.52) **	2.93 (2.67, 3.21) **
	Persistent vomiting	1 (0.2%)	108 (21.7%)	9 (16.7%)	89 (2.8%)	2.27 (1.99, 2.58) **	2.39 (2.22, 2.57) **
	Epistaxis[1]	0 (0%)	25 (5%)	8 (14.8%)	51 (1.6%)	1.49 (1.14, 1.95) **	2.18 (2.05, 2.33) **
	Gingivorrhagia[1]	1 (0.2%)	30 (6%)	10 (18.5%)	73 (2.3%)	1.36 (1.06, 1.75) **	2.15 (1.98, 2.32) **

	Ecchymosis[1]	0 (0%)	6 (1.2%)	4 (7.4%)	19 (0.6%)	1.30 (0.78, 2.15)	2.13 (2.01, 2.27) **
	Ascites[4]	0 (0%)	2 (0.4%)	1 (1.9%)	11 (0.3%)	0.80 (0.29, 2.20)	2.12 (2.00, 2.25) **
	Hematomas[1]	0 (0%)	2 (0.4%)	0 (0%)	13 (0.4%)	0.50 (0.14, 1.82)	2.12 (1.99, 2.25) **
	Pleural effusion[4]	0 (0%)	1 (0.2%)	1 (1.9%)	13 (0.4%)	0.50 (0.14, 1.82)	2.12 (1.99, 2.25) **
	Hepatomegaly[6]	1 (0.2%)	24 (4.8%)	4 (7.4%)	53 (1.6%)	1.34 (0.99, 1.80)	2.10 (1.91, 2.30) **
	Plasma leakage[4]	1 (0.2%)	2 (0.4%)	1 (1.9%)	8 (0.2%)	1.25 (0.56, 2.79)	1.59 (0.90, 2.81)
	Photophobia[2]	40 (6.5%)	51 (10.2%)	5 (9.3%)	279 (8.7%)	0.96 (0.80, 1.15)	1.26 (1.05, 1.51) **
Severe Symptoms	Disorientation[2]	0 (0%)	2 (0.4%)	5 (9.3%)	13 (0.4%)	1.32 (0.72, 2.40)	2.13 (2.00, 2.26) **
	Jaundice	0 (0%)	1 (0.2%)	1 (1.9%)	23 (0.7%)	0.30 (0.08, 1.13)	2.12 (1.99, 2.25) **
	Shock[3]	0 (0%)	0 (0%)	2 (3.7%)	15 (0.5%)	0.44 (0.12, 1.62)	2.12 (1.99, 2.25) **
	Stupor[2]	0 (0%)	0 (0%)	2 (3.7%)	11 (0.3%)	0.58 (0.16, 2.07)	2.12 (1.99, 2.25) **
	Cold Extremities[3]	1 (0.2%)	1 (0.2%)	13 (24.1%)	20 (0.6%)	1.62 (1.10, 2.38) **	2.00 (1.72, 2.32) **
	Mottled skin[1]	1 (0.2%)	7 (1.4%)	4 (7.4%)	20 (0.6%)	1.41 (0.90, 2.22)	1.96 (1.63, 2.35) **

	Abdominal pain[5]	38 (6.2%)	135 (27.1%)	15 (27.8%)	350 (10.9%)	1.37 (1.21, 1.56) **	1.94 (1.75, 2.16) **
	Capillary filling[3]	1 (0.2%)	3 (0.6%)	4 (7.4%)	17 (0.5%)	1.20 (0.68, 2.14)	1.86 (1.42, 2.43) **
	Tachycardia	3 (0.5%)	6 (1.2%)	13 (24.1%)	37 (1.2%)	1.41 (1.01, 1.97) **	1.86 (1.55, 2.22) **
	Melena[1]	1 (0.2%)	4 (0.8%)	2 (3.7%)	13 (0.4%)	1.32 (0.72, 2.40)	1.82 (1.34, 2.48) **
	Respiratory failure[7]	1 (0.2%)	1 (0.2%)	5 (9.3%)	14 (0.4%)	1.25 (0.68, 2.30)	1.82 (1.34, 2.48) **
	Hematemesis[1]	2 (0.3%)	2 (0.4%)	7 (13%)	13 (0.4%)	1.73 (1.12, 2.68) **	1.74 (1.31, 2.32) **
	Altered State of Consciousness[2]	1 (0.2%)	0 (0%)	4 (7.4%)	19 (0.6%)	0.78 (0.36, 1.71)	1.70 (1.09, 2.64) **
	Arterial hypotension[3]	1 (0.2%)	1 (0.2%)	3 (5.6%)	5 (0.2%)	1.88 (1.01, 3.50) **	1.70 (1.09, 2.64) **
	Metrorrhagia[1]	1 (0.2%)	2 (0.4%)	2 (3.7%)	4 (0.1%)	2.09 (1.16, 3.76) **	1.70 (1.09, 2.64) **
	Lipothymia[3]	1 (0.2%)	2 (0.4%)	1 (1.9%)	25 (0.8%)	0.52 (0.21, 1.28)	1.59 (0.90, 2.81)
	Convergent pressure[3]	1 (0.2%)	0 (0%)	1 (1.9%)	5 (0.2%)	1.07 (0.33, 3.47)	1.06 (0.26, 4.23)

	Myocarditis[7]	1 (0.2%)	0 (0%)	1 (1.9%)	2 (0.1%)	1.88 (0.70, 5.01)	1.06 (0.26, 4.23)
	Subarachnoid hemorrhage[1]	1 (0.2%)	0 (0%)	0 (0%)	2 (0.1%)	1.25 (0.25, 6.21)	0

Risk Ratio NSD Vs OFI in terms of dengue compared to Other. Risk Ratio SD/DSW Vs NSD in terms of severe (severe dengue or dengue with warning sings) compared to non-severe dengue. ** Indicates significant univariate risk ratio compared to reference. Risk ratio references for sex and age were female patients and those $\leq$ 14 years old, respectively. Superscripts (1-7) indicate groups made for model processing: 1=Bleeding, 2=Neurological, 3=Hemodynamic, 4=Plasma Leakage, 5=Nausea plus, 6=Visceromegaly, 7=Atypical Manifestation.

Citations

Barnes, W J, and L Rosen. “Fatal hemorrhagic disease and shock associated with primary dengue infection on a Pacific island.” *The American journal of tropical medicine and hygiene* vol. 23,3 (1974): 495-506. doi:10.4269/ajtmh.1974.23.495

Bhatt, Samir et al. “The global distribution and burden of dengue.” *Nature* vol. 496,7446 (2013): 504-7. doi:10.1038/nature12060

Boström, Henrik. (2008). Estimating class probabilities in random forests. 211-216. 10.1109/ICMLA.2007.64.

Brady, Oliver J et al. “Refining the global spatial limits of dengue virus transmission by evidence-based consensus.” *PLoS neglected tropical diseases* vol. 6,8 (2012): e1760. doi:10.1371/journal.pntd.0001760

Brathwaite Dick, Olivia et al. “The history of dengue outbreaks in the Americas.” *The American journal of tropical medicine and hygiene* vol. 87,4 (2012): 584-93. doi:10.4269/ajtmh.2012.11-0770

Breiman, Leo. (2001). Random Forests. *Machine Learning*, Volume 45, 5-32 (2001): 10.1023/A:1010933404324.

Breiman, L., Cutler, A. Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-14 (2018).

Cavalcanti, L. P., Mota, L. A., Lustosa, G. P., Fortes, M. C., Mota, D. A., Lima, A. A., Coelho, I. C., & Mourão, M. P. (2014). Evaluation of the WHO classification of dengue disease severity during an epidemic in 2011 in the state of Ceará, Brazil. *Memorias do Instituto Oswaldo Cruz*, *109*(1), 93–98. https://doi.org/10.1590/0074-0276140384

Chaterji, Shera et al. “Evaluation of the NS1 rapid test and the WHO dengue classification schemes for use as bedside diagnosis of acute dengue fever in adults.” *The American journal of tropical medicine and hygiene* vol. 84,2 (2011): 224-8. doi:10.4269/ajtmh.2011.10-0316

Christodoulou, Evangelia et al. “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.” *Journal of clinical epidemiology* vol. 110 (2019): 12-22. doi:10.1016/j.jclinepi.2019.02.004

Cotterman, Carolyn. *Statistical Methods for Predicting Dengue Diagnosis using Clinical and LC-MS Data*. 2015. University of California, Berkeley, PhD dissertation.

Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018 Jul 17;19(1):270. doi: 10.1186/s12859-018-2264-5. PMID: 30016950; PMCID: PMC6050737.

Cucunawangsih et al. “Scoring Model to Predict Dengue Infection in the Early Phase of Illness in Primary Health Care Centre.” *Archives of Clinical Microbiology* 6 (2015): n. pag.

Dankowski, Theresa, and Andreas Ziegler. “Calibrating random forests for probability estimation.” *Statistics in medicine* vol. 35,22 (2016): 3949-60. doi:10.1002/sim.6959

Deen, Jacqueline & Harris, Eva & Wills, Bridget & Balmaseda, Angel & Hammond, Samantha & Rocha, Crisanta & Dung, Nguyen & Hung, Nguyen & Tinh Tran, Hien & Farrar, Jeremy. (2006). The WHO dengue classification and case definitions: time for a reassessment. Lancet. 368. 170-3. 10.1016/S0140-6736(06)69006-5.

Devi, Debashree, et al. “A Cost-Sensitive Weighted Random Forest Technique for Credit Card Fraud Detection.” *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, doi:10.1109/icccnt45670.2019.8944885.

Flaxman, Abraham D, and Theo Vos. “Machine learning in population health: Opportunities and threats.” *PLoS medicine* vol. 15,11 e1002702. 27 Nov. 2018, doi:10.1371/journal.pmed.1002702

Freeman, Elizabeth & Gretchen, Moisen. (2007). PresenceAbsence: An R Package for Presence Absence Analysis. Journal of Statistical Software. 23. 10.18637/jss.v023.i11.

Gajowniczek, Krzysztof et al. “Weighted Random Forests to Improve Arrhythmia Classification.” *Electronics* vol. 9,1 (2020): 10.3390/electronics9010099. doi:10.3390/electronics9010099

GBD 2017 DALYs and HALE Collaborators (2018). Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet (London, England)*, *392*(10159), 1859–1922. https://doi.org/10.1016/S0140-6736(18)32335-3

Goode, Ka. And Hofmann, H. Visual diagnostics of an explainer model: Tools for the assessment of LIME explanations. Statistical Analysis and Data Mining: The ASA Data Science Journal. 14, (2021). 10.1002/sam.11500.

Gubler, D J et al. “Epidemiologic, clinical, and virologic observations on dengue in the Kingdom of Tonga.” *The American journal of tropical medicine and hygiene* vol. 27,3 (1978): 581-9. doi:10.4269/ajtmh.1978.27.581

Gubler, D J. “Dengue and dengue hemorrhagic fever.” *Clinical microbiology reviews* vol. 11,3 (1998): 480-96.

Gutiérrez, G., Gresh, L., Pérez, M. Á., Elizondo, D., Avilés, W., Kuan, G., Balmaseda, A., & Harris, E. (2013). Evaluation of the diagnostic utility of the traditional and revised WHO dengue case

definitions. *PLoS neglected tropical diseases*, *7*(8), e2385. https://doi.org/10.1371/journal.pntd.0002385

Hadinegoro S. R. (2012). The revised WHO dengue case classification: does the system need to be modified?. *Paediatrics and international child health*, *32 Suppl 1*(s1), 33–38. https://doi.org/10.1179/2046904712Z.00000000052

Hadinegoro, Sri Rezeki S. "The revised WHO dengue case classification: does the system need to be modified?." *Paediatrics and international child health* vol. 32 Suppl 1,s1 (2012): 33-8. doi:10.1179/2046904712Z.00000000052

Halstead, S B. "Observations related to pathogensis of dengue hemorrhagic fever. VI. Hypotheses and discussion." *The Yale journal of biology and medicine* vol. 42,5 (1970): 350-62.

Halstead, S B. "Pathogenesis of dengue: challenges to molecular biology." *Science (New York, N.Y.)* vol. 239,4839 (1988): 476-81. doi:10.1126/science.3277268

Halstead, S. "Dengue in the Americas and Southeast Asia: do they differ?" *Revista panamericana de salud publica = Pan American journal of public health* 20 6 (2006): 407-15 .

Heinze, Georg et al. "Variable selection - A review and recommendations for the practicing statistician." *Biometrical journal. Biometrische Zeitschrift* vol. 60,3 (2018): 431-449. doi:10.1002/bimj.201700067

Ho, Tzong-Shiann et al. "Comparing machine learning with case-control models to identify confirmed dengue cases." *PLoS neglected tropical diseases* vol. 14,11 e0008843. 10 Nov. 2020, doi:10.1371/journal.pntd.0008843

Hoogland, Jeroen et al. "Handling missing predictor values when validating and applying a prediction model to new patients." *Statistics in medicine* vol. 39,25 (2020): 3591-3607. doi:10.1002/sim.8682

Institute for Health Metrics and Evaluation (IHME). Findings from the Global Burden of Disease Study 2017. Seattle, WA: IHME, 2018.

Kalayanarooj, Siripen. (1999). Standardized Clinical Management: Evidence of Reduction of Dengue Haemorrhagic Fever Case-Fatality Rate in Thailand.. WHO Regional Office for South-East Asia. https://apps.who.int/iris/handle/10665/148661

Katuwal, Gajendra J. and R. Chen. "Machine Learning Model Interpretability for Precision Medicine." *arXiv: Quantitative Methods* (2016): n. pag.

Kingsford, Carl, and Steven L Salzberg. "What are decision trees?." *Nature biotechnology* vol. 26,9 (2008): 1011-3. doi:10.1038/nbt0908-1011

Kull, Meelis et al. "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration." *Electronic Journal of Statistics* 11 (2017): 5052-5080.

Kursa, Miron & Rudnicki, Witold. (2010). Feature Selection with Boruta Package. Journal of Statistical Software. 36. 1-13. 10.18637/jss.v036.i11.

Landis, J R, and G G Koch. "The measurement of observer agreement for categorical data." *Biometrics* vol. 33,1 (1977): 159-74.

Levi J. E., Nishiya A., Félix A. C., et al. Real-time symptomatic case of transfusion-transmitted dengue. *Transfusion*. 2015;55(5):961–964.

Lunardon, N. et al. "ROSE: a Package for Binary Imbalanced Learning." *R J.* 6 (2014): 79.

More, A. S., and D. P. Rana. *An Experimental Assessment of Random Forest Classification Performance Improvisation with Sampling and Stage Wise Success Rate Calculation*, vol. 167, 2020, doi:10.1016/j.procs.2020.03.381.

Mugabe VA, Borja LS, Cardoso CW, Weaver SC, Reis MG, Kitron U, Ribeiro GS. Changes in the dynamics of dengue incidence in South and Central America are possibly due to cross-

population immunity after Zika virus epidemics. Trop Med Int Health. 2021 Mar;26(3):272-280. doi: 10.1111/tmi.13526. Epub 2020 Nov 30. PMID: 33190402.

Murillo-Zamora, Efrén et al. "Clinical markers associated with acute laboratory-confirmed Dengue infection: results of a national epidemiological surveillance system." *Revista de salud publica (Bogota, Colombia)* vol. 19,6 (2017): 744-748. doi:10.15446/rsap.V19n6.61823

Narvaez, Federico et al. "Evaluation of the traditional and revised WHO classifications of Dengue disease severity." *PLoS neglected tropical diseases* vol. 5,11 (2011): e1397. doi:10.1371/journal.pntd.0001397

Pajouheshnia, Romin & van Smeden, Maarten & Peelen, L.M. & Groenwold, R.H.H.. (2018). How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. Journal of Clinical Epidemiology. 105. 10.1016/j.jclinepi.2018.09.001.

Pan American Health Organization (PAHO). *Reported Cases of Dengue Fever in the Americas by Country or Territory Cumulative Cases*. PLISA Health Information Platform for the Americas, 2020.

Pang, Junxiong et al. "Diabetes with hypertension as risk factors for adult dengue hemorrhagic fever in a predominantly dengue serotype 2 epidemic: a case control study." *PLoS neglected tropical diseases* vol. 6,5 (2012): e1641. doi:10.1371/journal.pntd.0001641

Patel, H. & Prajapati, P. *Study and Analysis of Decision Tree Based Classification Algorithms*. International Journal of Computer Sciences and Engineering. 6. 74-78 (2018): 10.26438/ijcse/v6i10.7478.

Peng, C. et al. "An Introduction to Logistic Regression Analysis and Reporting." *The Journal of Educational Research* 96 (2002): 14 - 3.

Potts, James A et al. “Classification of dengue illness based on readily available laboratory data.” *The American journal of tropical medicine and hygiene* vol. 83,4 (2010): 781-8. doi:10.4269/ajtmh.2010.10-0135

Ramspek, Chava L et al. “External validation of prognostic models: what, why, how, when and where?.” *Clinical kidney journal* vol. 14,1 49-58. 24 Nov. 2020, doi:10.1093/ckj/sfaa188

Random forest vs logistic. RF better with more sampls

Reis, I. et al. “Probabilistic Random Forest: A machine learning algorithm for noisy datasets.” *ArXiv* abs/1811.05994 (2018): n. pag.

Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. 97-101. 10.18653/v1/N16-3020.

Ribeiro, Marco Tulio et al. “Model-Agnostic Interpretability of Machine Learning.” *ArXiv* abs/1606.05386 (2016): n. pag.

Rodriguez-Roche R, Blanc H, Bordería AV, Díaz G, Henningsson R, Gonzalez D, Santana E , Alvarez M, Castro O, Fontes M, Vignuzzi M, Guzman MG. Increasing Clinical Severity during a Dengue Virus Type 3 Cuban Epidemic: Deep Sequencing of Evolving Viral Populations. J Virol. 2016 Apr 14; 90 (9): 4320-4333 . doi: 10.1128 / JVI.02647-15. PMID: 26889031; PMCID: PMC4836355.)

Rosen, L. “Dengue in Greece in 1927 and 1928 and the pathogenesis of dengue hemorrhagic fever: new data and a different conclusion.” *The American journal of tropical medicine and hygiene* vol. 35,3 (1986): 642-53. doi:10.4269/ajtmh.1986.35.642

Rosen, L. “The Emperor's New Clothes revisited, or reflections on the pathogenesis of dengue hemorrhagic fever.” *The American journal of tropical medicine and hygiene* vol. 26,3 (1977): 337-43. doi:10.4269/ajtmh.1977.26.337

Sanchez-Pinto, L. N., et al. "Comparison of Variable Selection Methods for Clinical Predictive Modeling." *International Journal of Medical Informatics*, vol. 116, 2018, pp. 10-17. doi:10.1016/j.ijmedinf.2018.05.006

Shepard, D. S., Undurraga, E. A., Halasa, Y. A., & Stanaway, J. D. (2016). The global economic burden of dengue: a systematic analysis. *The Lancet. Infectious diseases*, *16*(8), 935–941. https://doi.org/10.1016/S1473-3099(16)00146-8

Sim, Julius, and Chris C Wright. "The kappa statistic in reliability studies: use, interpretation, and sample size requirements." *Physical therapy* vol. 85,3 (2005): 257-68.

Simmons, Farrar JJ, Nguyen vV, Wills B. Dengue. N Engl J Med. 2012 Apr 12;366(15):1423-32. doi: 10.1056/NEJMra1110265. PMID: 22494122.

Siontis, George C M et al. "External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination." *Journal of clinical epidemiology* vol. 68,1 (2015): 25-34. doi:10.1016/j.jclinepi.2014.09.007

Sokolova, Marina et al. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation." *Australian Conference on Artificial Intelligence* (2006).

Speiser, Jaime Lynn et al. "A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling." *Expert systems with applications* vol. 134 (2019): 93-101. doi:10.1016/j.eswa.2019.05.028

Srikiatkhachorn A, Rothman AL, Gibbons RV, Sittisombut N, Malasit P, Ennis FA, Nimmannitya S, Kalayanorooj S. Dengue - how best to classify it. Clin Infect Dis. 2011;53:563–567. [PMC free article] [PubMed] [Google Scholar]

Stekhoven, Daniel J, and Peter Bühlmann. "MissForest--non-parametric missing value imputation for mixed-type data." *Bioinformatics (Oxford, England)* vol. 28,1 (2012): 112-8. doi:10.1093/bioinformatics/btr597

Sumarmo et al. "Clinical observations on virologically confirmed fatal dengue infections in Jakarta, Indonesia." *Bulletin of the World Health Organization* vol. 61,4 (1983): 693-701.

Sumarmo, Wulur, H., Jahja, E., Gubler, D. J., Suharyono, W., & Sorensen, K. (1983). Clinical observations on virologically confirmed fatal dengue infections in Jakarta, Indonesia. *Bulletin of the World Health Organization*, *61*(4), 693–701.

Sumarmo, Wulur, H., Jahja, E., Gubler, D. J., Suharyono, W., & Sorensen, K. (1983). Clinical observations on virologically confirmed fatal dengue infections in Jakarta, Indonesia. *Bulletin of the World Health Organization*, *61*(4), 693–701.

Tajgardoon, Mohammadamin et al. "Patient-Specific Explanations for Predictions of Clinical Outcomes." *Applied Clinical Informatics (ACI) Journal* 03(02), (2019): e88-e97. DOI: 10.1055/s-0039-1697907.

Tamibmaniam, J., Hussin, N., Cheah, W. K., Ng, K. S., & Muninathan, P. (2016). Proposal of a Clinical Decision Tree Algorithm Using Factors Associated with Severe Dengue Infection. *PloS one*, *11*(8), e0161696. https://doi.org/10.1371/journal.pone.0161696

Tanner, Lukas et al. "Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness." *PLoS neglected tropical diseases* vol. 2,3 e196. 12 Mar. 2008, doi:10.1371/journal.pntd.0000196

Teixeira, Maria Gloria et al. "Recent shift in age pattern of dengue hemorrhagic fever, Brazil." Emerging infectious diseases vol. 14,10 (2008): 1663. doi:10.3201/eid1410.071164

Tuan, Nguyen Minh et al. “Sensitivity and specificity of a novel classifier for the early diagnosis of dengue.” *PLoS neglected tropical diseases* vol. 9,4 e0003638. 2 Apr. 2015, doi:10.1371/journal.pntd.0003638

Valuations, E.. “A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS.” (2015).

van der Ploeg, Tjeerd et al. “Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints.” *BMC medical research methodology* vol. 14 137. 22 Dec. 2014, doi:10.1186/1471-2288-14-137

Visani, Giorgio et al. “OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms.” *ArXiv* abs/2006.05714 (2020): n. pag.

WHO, Kochar, A. “Dengue and Severe Dengue.” Fact Sheet. Word Health Organization. June 2020. D.C n.d. Web.

WHO. *Dengue guidelines for diagnosis, treatment, prevention and control : new edition*. Geneva: World Health Organization, (2009). x, 147 p. p.

WHO. *Dengue haemorrhagic fever* : *diagnosis, treatment, prevention* and *control, 2nd ed*. *World Health Organization, 1997*.

WHO, Department of Control of Neglected Tropical Diseases. Global strategy for dengue prevention and control 2012-2020. (August 2012). Retrieved from https://www.who.int/denguecontrol/9789241504034/en/

World Health Organization (WHO). (*1997*). *Dengue haemorrhagic fever* : *diagnosis, treatment, prevention* and *control, 2nd ed*. *World Health Organization*.

World Health Organization (WHO). (2009). Dengue guidelines for diagnosis, treatment, prevention and control : new edition. Chapter Two. World Health Organization. https://apps.who.int/iris/handle/10665/44188

World Health Organization. The Global Burden of Disease: 2004 Update. doi:https://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf?ua=1

Yang, Fan et al. “Using random forest for reliable classification and cost-sensitive learning for medical diagnosis.” *BMC bioinformatics* vol. 10 Suppl 1,Suppl 1 S22. 30 Jan. 2009, doi:10.1186/1471-2105-10-S1-S22

Zubieta-Zavala, Adriana et al. “Economic impact of dengue in Mexico considering reported cases for 2012 to 2016.” *PLoS neglected tropical diseases* vol. 12,12 e0006938. 14 Dec. 2018, doi:10.1371/journal.pntd.0006938