Heliyon 9 (2023) e16634

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

A machine learning model to assess potential misdiagnosed dengue hospitalization

Claudia Yang Santos [a], Suely Tuboi [a], Ariane de Jesus Lopes de Abreu [b], Denise Alves Abud [a], Abner Augusto Lobao Neto [a], Ramon Pereira [b], Joao Bosco Siqueira Jr. [c,*]

[a] Takeda Pharmaceuticals Brazil, Av. das Nações Unidas 14401, São Paulo, SP, Brazil
[b] IQVIA Brazil, Rua Verbo Divino 2001, São Paulo, SP, Brazil
[c] Federal University of Goias, Av. Esperança, Goiania, GO, Brazil

ARTICLE INFO

Keywords:
Dengue
Machine learning
Disease burden
Probability learning
Brazil

ABSTRACT

Dengue, like other arboviruses with broad clinical spectra, can easily be misdiagnosed as other infectious diseases due to the overlap of signs and symptoms. During large outbreaks, severe dengue cases have the potential to overwhelm the health care system and understanding the burden of dengue hospitalizations is therefore important to better allocate medical care and public health resources. A machine learning model that used data from the Brazilian public healthcare system database and the National Institute of Meteorology (INMET) was developed to estimate potential misdiagnosed dengue hospitalizations in Brazil. The data was modeled into a hospitalization level linked dataset. Then, Random Forest, Logistic Regression and Support Vector Machine algorithms were assessed. The algorithms were trained by dividing the dataset in training/test set and performing a cross validation to select the best hyperparameters in each algorithm tested. The evaluation was done based on accuracy, precision, recall, F1 score, sensitivity, and specificity.

The best model developed was Random Forest with an accuracy of 85% on the final reviewed test. This model shows that 3.4% (13,608) of all hospitalizations in the public healthcare system from 2014 to 2020 could have been dengue misdiagnosed as other diseases. The model was helpful in finding potentially misdiagnosed dengue and might be a useful tool to help public health decision makers in planning resource allocation.

1. Introduction

Dengue is a mosquito-borne viral illness caused by 4 serotypes that is clinically characterized by a wide spectrum of signs and symptoms ranging from a mild acute febrile illness to potentially fatal severe dengue [1]. Classic dengue fever is a self-limited acute febrile illness with sudden onset of fever that is accompanied by non-specific signs and symptoms [2]. However, approximately 1 in 20 patients with dengue may progress to a severe, life-threatening disease owing to a variety of risk factors, including secondary exposure to a different serotype [3,4]. There is no specific treatment for dengue and case management relies mostly on supportive measures. Since confirmatory tests are not mandatory, clinical-epidemiological criteria can be used to guide treatment [5]. As a result, these tests

* Corresponding author. Federal University of Goias, Av. Esperança, Goiania, GO 74690-900, Brazil.
E-mail address: joao.siqueira@ufg.br (J.B. Siqueira).

https://doi.org/10.1016/j.heliyon.2023.e16634
Received 20 September 2022; Received in revised form 23 May 2023; Accepted 23 May 2023
Available online 30 May 2023

are frequently not promptly available, and a considerable proportion of hospitalized cases are not laboratory confirmed. According to the Pan American Health Organization (PAHO), only 43% of hospitalized cases of dengue had a confirmatory test in 2019 [6]. In addition, the emergence of other circulating febrile diseases with similar clinical manifestations in Brazil, such as chikungunya (CHIKV) and Zika (ZIKV), especially in the last decade, has become a potential source of misdiagnosis of dengue [7,8]. Misdiagnosis not only hinders epidemiological understanding of the burden of affected diseases but also may have a profound impact on patients' outcomes by delaying dengue-specific supportive treatment and might increase lethality [4,9].

In Brazil, the majority of hospitalized dengue cases are treated by the Public Healthcare System (SUS), which provides care to 75% of the population [10]. As with any notifiable disease, dengue hospitalization should be reported in two databases: the National Notifiable Diseases Information System (SINAN) and the Hospital Information System (SIH). Both are managed by and stored at a large data system, the Department of Information Technology (DATASUS) and are publicly available at (https://datasus.saude.gov.br/). Although reporting is mandatory, both databases are known to lack sensitivity and completeness [7,11,12]. In addition, it is also likely that a significant proportion of hospitalizations in SIH may have been misdiagnosed as other diseases, especially because there is a considerable overlap of dengue clinical signs and symptoms with other infectious diseases [11].

The use of secondary data in health outcomes research is attractive as they represent a quick and less expensive way to conduct exploratory analyses of large volumes of information with population and geographic variability. As a result of increasing availability of data sources, the literature on the use of regression methods to evaluate multiple predictors in disease modeling [13,14] is mounting, and so is the search for approaches that incorporate new technologies. Machine learning (ML) is one such example, given its ability to analyze highly complex data without the constraints of the traditional modeling requirements. Most ML studies in dengue are done either in diagnostic, intervention or epidemic forecast [15]. Some studies focus on the early diagnosis of dengue using clinical, laboratory and demographic data [16], or the differentiation of dengue from other arboviruses [17]. For intervention, only few models have been developed; they focus on biological control [18], vaccination [19] and fumigation [20]. A literature review conducted in Latin America for dengue prediction using ML determined that many factors can act as predictors for dengue outbreak [21]. The most explored factor was climate data in different countries: Argentina [22], Brazil [15] and Colombia [23]. To our knowledge, ML has never been used to evaluate potential dengue hospitalizations misclassified in the public system. In this study, the aim was to use a ML model to predict the potential number of hospitalized cases reported in the SIH that could have been misdiagnosed cases of dengue. For this purpose, surveillance, hospitalization and climate data from large publicly available datasets were used. The model can be a useful tool to identify hospitalized patients with misdiagnosed dengue and help health care decision makers to better allocate resources.

2. Methods

2.1. Data sources and feature selection

Data from three publicly available datasets were used to develop this ML model. SIH [24], SINAN and Meteorological Database from the National Institute of Meteorology (INMET) [25]. SIH is an administrative database for reimbursement purposes. Administrative claims data are presented as procedure codes from billing records and include demographic information, procedures performed (type and number), costs of procedures, and other information. Procedure codes and reference names are available in SIGTAP, which is

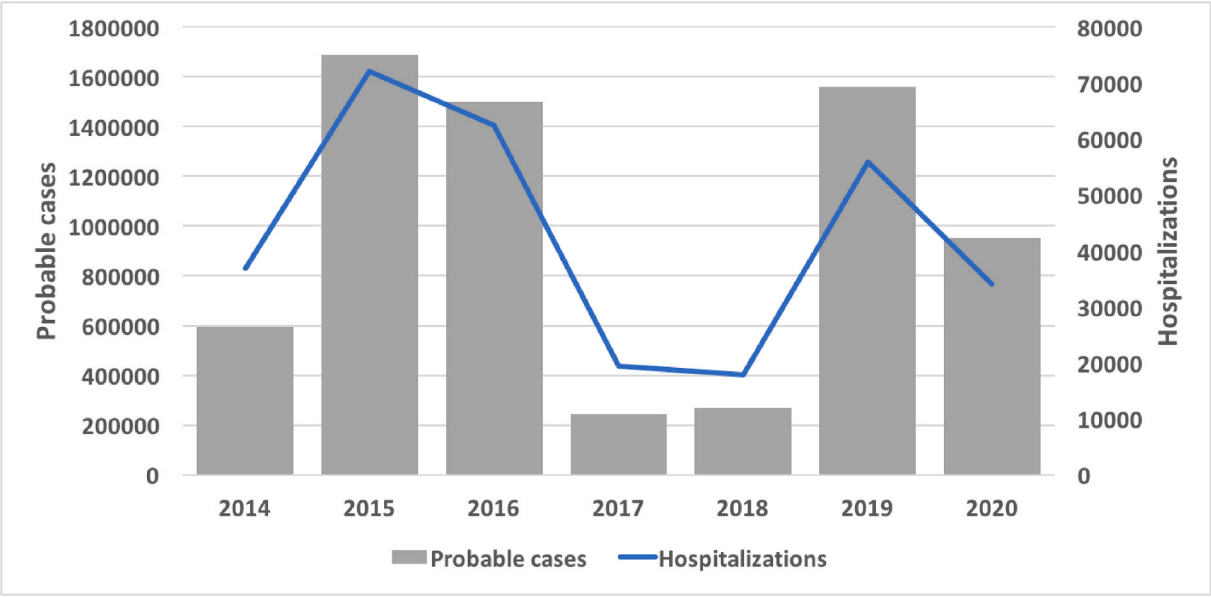


Fig. 1. Dengue probable cases (SINAN) and hospitalizations (SIH), Brazil, 2014–2020.

the database of standardized procedures approved within SUS [26]. Due to its administrative nature, SIH does not contain clinical data (e.g., signs and symptoms). Thus, cause of admission (as per International Classification of Diseases (ICD) code) and procedures performed during the hospitalization were used as predictor variables. Additionally, data related to patient's age, municipality of hospitalization, hospitalization date, diagnosis at entry (ICD based), final diagnosis (ICD based), procedures prescribed and performed, and length of stay (days) were also extracted (Table S1).

To build the model, a list of ICD-10 codes was created considering symptoms linked to dengue, as well as a list of procedures used for the management of dengue fever, based on the literature [8,17,27–31]. These lists were then validated by a group of dengue experts and are presented at Supplementary Material (Tables S1–S3).

For dengue incidence and hospitalization events, SINAN and SIH databases were used. Finally, since dengue is a seasonal disease, with greater transmission in seasons with high temperatures and precipitation, INMET was assessed for information on the average precipitation from the entire country. The model considered the weekly average of rainfall and temperature around hospitalization dates. The final dataset generated by the model was also compared to that in SIH.

2.2. Cohort generation

Hospitalizations registered in DATASUS with the selected ICD-10 codes from January 2014 to December 2020 (based on hospital admission date) were included in the cohort. The study period was defined to cover the more recent dengue epidemics in Brazil (Fig. 1), as well as to cover the emergence of other febrile diseases circulating in Brazil in recent years. The data from the three different sources were merged in a consolidated dataset, grouping with counting, and summing of procedures and ICD codes (Fig. 2). No statistical approach was taken to handle missing data. The categorical variables were converted with One Hot Encoder and a missing data was represented by zero and no records were excluded from the analysis.

2.3. Machine learning model

All machine learning analyses were performed in Python (using numpy, pandas, pyspark, sklearn 0.24.2, matplotlib, datetime and imblearn 0.8). Following recent publications to predict hospitalizations for a specific disease [32,33], three different supervised algorithms (logistic regression, support vector machine, and random forest) were used, so the one with the best performance could be chosen. The first step of the ML approach was to separate a subset of 500 hospitalizations for further manual review and a "real world test". Then, training was performed on the training input dataset with a train-to-test split ratio of 80:20 based on the Pareto principle and following guidance from the literature [34], using the aforementioned cohort with stratified approach. Each hospitalization was labeled as dengue (ICD-10 A90.0 or A91.0) or not dengue (any other selected ICD-10 code). A cross validation technique was performed using grid search with fivefold [35] and SMOTE oversampling [36] on minority only on the training set of the fold, to select the best hyperparameters and generate a tuned model for each algorithm tested (Step 2) (Fig. 3). The tuned models were trained on the training set (80%) and tested in the 20% holdout sample. The evaluation was done based on accuracy, precision, recall, F1 score, sensitivity, and specificity (Fig. 4).

The random forest model was the algorithm with the best performance (as evaluated by the F1 score and by the trade of between sensitivity and specificity) and was carried forward for further final testing. The final model test was performed using the subset of all data separated at step 1 (test dataset with 0.1% of all hospitalizations in the study period) that was manually reviewed and labeled by two independent researchers who checked the hospitalization classifications according to dengue's guidelines.

An important point on this job was the use of the model to proceed with the analysis. The result of the classification of the model was used to create a new class defined as "Dengue-like" to be statistically evaluated. It represents those hospitalizations not diagnosed as dengue (neither A90 nor A91) but classified as dengue by the model and which also had at least one record of dengue treatment OR at least one mandatory test AND a complementary test (Table S4). In addition, to define dengue-like hospitalization.

The year of hospitalization (years known to be epidemic for dengue: 2015, 2016, 2019) and the incidence of dengue for each analyzed year were observed, according to the epidemiological reports of the Ministry of Health (Fig. 1).

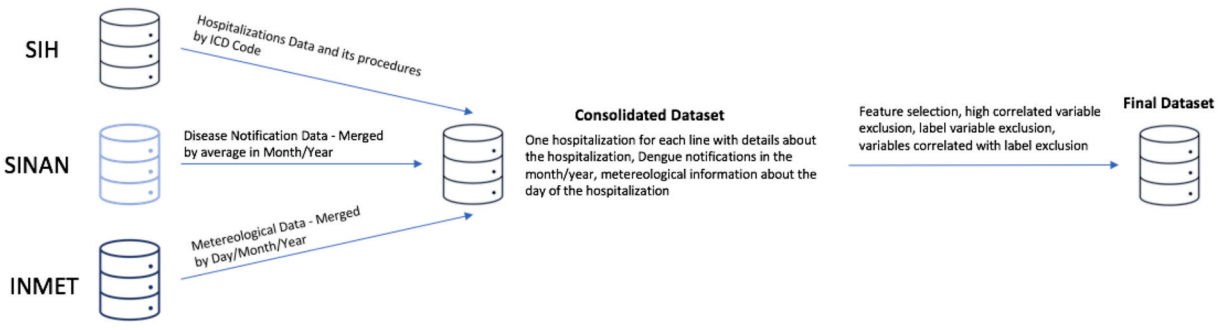


Fig. 2. Flowchart of data acquisition for consolidated dataset and final dataset.

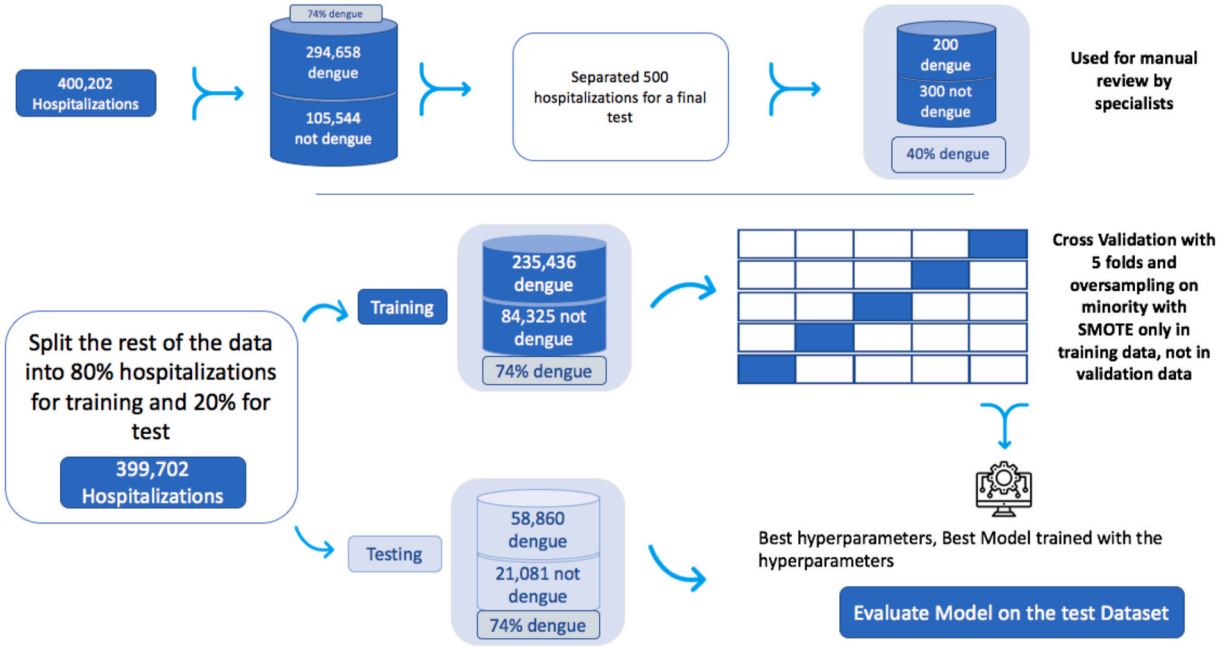
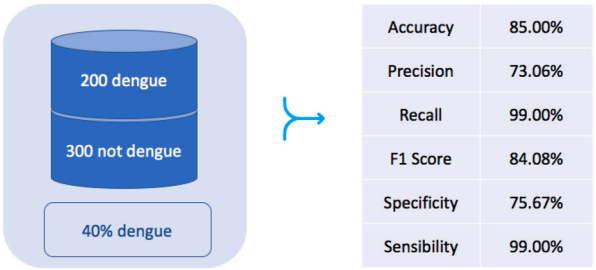


Fig. 3. Step 1 and 2 for the machine learning approach.



Accuracy	85.00%
Precision	73.06%
Recall	99.00%
F1 Score	84.08%
Specificity	75.67%
Sensibility	99.00%

Fig. 4. Results from test model and results from final test validated model.

2.4. Data analysis

The data analyses were performed after classification of the hospitalizations by the ML model using Python version 3.6.9. Categorical variables (number and proportion of dengue-like hospitalizations) were described by simple and crossed contingency tabulation with frequencies and absolute percentages and stratified by year and age group. Dengue-like hospitalization rates were calculated by the number of hospitalizations classified as dengue-like per year divided per the total Brazilian population per year and multiplied by 100,000. The proportional rates of the model defined dengue-like cases to dengue cases and age-group distribution were compared by dividing events of dengue and dengue like cases by their respective totals to better describe how the model performed.

3. Results and discussion

In this study, we propose a predictive model based on machine learning techniques in identifying dengue-like cases from retrospective data. Our approach provides a novel and efficient way to analyze large datasets, particularly in the context of infectious diseases where timely and accurate diagnosis is crucial. By leveraging a comprehensive set of features, including clinical and demographic data, we were able to build a predictive model. Our study used a ML model to predict the amount of potential misdiagnosed cases of dengue in the SIH database from 2014 to 2020 and found that as many as 3.4% of all hospitalizations reported could have been due to dengue. The final validated ML model showed a good performance in classifying dengue-like hospitalizations, predicting 99% of dengue hospitalizations and around 76% of not dengue hospitalizations with an accuracy of 85% (Fig. 4). To our knowledge, this is the first time that a ML model is used to assess potential dengue misdiagnosis in the Brazilian hospital setting.

A total of 400,202 hospitalizations were predicted by the model in the SIH database for the selected ICD-10 codes from 2014 to 2020 and are summarized in Table 1 (A). Of this total, the model classified 13,608 (3.4%) as dengue-like hospitalizations, 294,658

(73.6%) as dengue and 91,936 (23.0%) as not dengue hospitalizations.

The model classification was also compared against the SIH data in Table 1 (B). From 2014 to 2020, 13,608 dengue-like hospitalizations were identified by the model. Of these, 10,508 records were not in SIH, and 3,100 dengue hospitalizations in SIH were not classified as dengue by the model. The years with most cases missed by the model were 2015 and 2019 with 1,013 and 1,326 respectively. In 2020, however, the model predicted 408 potential dengue cases in addition to the already predicted hospitalizations registered in SIH.

Regarding diagnoses reported as the dengue-like hospitalizations, the most frequent ICD-10 codes found in the dengue-like hospitalizations were those related to febrile diseases, accounting for more than 80% of hospitalizations (Table S5). This finding was expected as similarities in the febrile course and manifestations of some infectious diseases may lead to clinical misdiagnosis. In addition, the uncertainty and pressure of a new infectious agent can introduce unconscious cognitive biases leading to diagnostic errors [37] and also reporting bias. This is particularly true in settings where confirmatory diagnostic tests are inaccessible and not crucial to guiding treatment. For example, Silva et al. [38] compared clinically diagnosed dengue following WHO 2009 guidelines with laboratory confirmed cases and found substantial variation in sensitivity, specificity, negative and positive predictive value. Most public health hospitals in Brazil lack diagnostic capabilities and notifiable diseases data rely heavily on clinical and syndromic approaches which have low specificity. Consequently, the introduction in 2014 and 2015 of CHIKV and ZIKV, two febrile diseases with similar clinical presentations and transmitted by the same vectors, may have resulted in an increased likelihood of misdiagnosis, which is further discussed in detail in the next paragraphs.

Table 2 shows the frequency of the ICD-10 codes for dengue-like hospitalizations per year.

The model classification showed an increasing trend of dengue-like hospitalizations in 2015 which peaks in 2016 while dengue hospitalizations reached a peak in 2015, followed by a decreasing trend from 2016 until 2019, when a large outbreak occurred in Brazil [39]. Although ICDs for Zika and chikungunya were excluded from the pool of potential dengue-like cases, these trends could well be depicting an overlap of these diseases that resulted in misdiagnosed cases for all three conditions. Indeed, a modeling exercise by Oidtman et al. [40] showed that there could have been a significant misdiagnosis of Zika from 2015 to 2017 in Brazil. In another study, Oliveira et al. [8] used a multivariate time series analysis to understand how dengue notified cases were impacted by the introduction and spread of chikungunya and Zika virus in Brazil and found that dengue was significantly impacted by Zika, and vice versa. Although mild cases may be clinically indistinguishable from dengue, severe Zika is most likely to present as a neurological disease such as Guillain Barre Syndrome, encephalitis or encephalomyelitis [41], which are rare manifestations of dengue. In the same line, while chikungunya is highly debilitating, it rarely requires hospitalization [42,43]. Therefore, it is reasonable to suppose that the high rate of dengue-like hospitalizations seen in 2015 and 2016 in the context of this triple epidemic was a result of the lack of accuracy of clinical diagnosis by health care professionals and could have been misdiagnosed cases dengue. This is in line with the findings of a large meta-analysis that showed a lack of diagnostic accuracy of World Health Organization (WHO) dengue clinical definitions, in particular with co-circulation of other febrile infectious diseases, such as COVID-19 [44].

Fig. 5 shows the hospitalization rates (per 100,000) as well as proportional rates of year distribution for dengue and dengue-like hospitalizations. Higher hospitalization rates were observed in 2015, 2016 and 2019 for dengue (42.04, 36.43 and 31.85 per 100,000 hospitalizations, respectively). As for dengue-like hospitalizations, although the highest incidences were observed in 2016 and 2019 (1.62 and 1.6 per 100,000 hospitalizations, respectively), the trend was maintained during 2017 and 2018 (1.37 and 1.25 per 100,000 hospitalizations, respectively) (Fig. 5). The proportional contribution of years 2017 and 2018 were also high for dengue-like hospitalizations when compared to dengue (Fig. 6).

During these inter-epidemic years, the rate of dengue-like hospitalization was proportionally higher than that of dengue; these were years where the model may have proven a useful tool. Decreased dengue awareness due to low prevalence in these years may have led

Table 1
Total dengue and dengue-like hospitalizations according to the model (A) and comparison with the National Hospitalization Information System (B) from 2014 to 2020.

	Total	2014	2015	2016	2017	2018	2019	2020
(A) Dengue and dengue-like hospitalizations according to the model								
Status of final model classification, N (%)								
Not dengue	91,936 (23.0)	9,044 (19.6)	11,141 (13.3)	12,772 (16.6)	12,267 (36.4)	12,570 (39.0)	14,678 (20.4)	19,464 (34.9)
Dengue [1]	294,658 (73.6)	36,391 (78.9)	71,093 (85.0)	61,396 (79.8)	19,167 (56.9)	17,649 (54.7)	54,504 (75.8)	34,458 (61.8)
Dengue-like [2]	13,608 (3.4)	690 (1.5)	1,384 (1.7)	2,728 (3.5)	2,226 (6.6)	2,041 (6.3)	2,737 (3.8)	1,802 (3.2)
Total Dengue + Dengue Like [3]	*308,266*	*37,081*	*72,477*	*64,124*	*21,393*	*19,690*	*57,241*	*36,260*
(B) Comparison with the National Hospitalization Information System (SIH)								
Total hospitalizations registered in SIH [4]	*297,758*	*36,809*	*72,106*	*61,646*	*19,467*	*17,850*	*55,830*	*34,050*
Difference from Model to SIH(3-4)	*10,508*	*272*	*371*	*2,478*	*1,926*	*1,840*	*1,411*	*2,210*
SIH hospitalizations not captured by the model(4-1)	*3,100*	*418*	*1,013*	*250*	*300*	*201*	*1,326*	*0*[a]

The model classified 13,608 (3.4%) as dengue-like hospitalizations, 294,658 (73.6%) as dengue and 91,936 (23.0%) as not dengue hospitalizations. The predictive model correctly classified 99% of dengue hospitalizations, but it was considered a 100% prediction for the analyses.

[a] All cases of dengue hospitalizations registered in SIH were potentially predicted by the model.

Table 2
Frequency of the ICD-10 codes most presented as dengue-like hospitalization.

	2014	2015	2016	2017	2018	2019	2020
A94 Arthropod-borne viral fever, unspecified n (%)	531 (45.6)	403 (23.1)	658 (19.9)	1086 (34.1)	836 (29.9)	1189 (31.5)	607 (26.2)
A92.9 Mosquito-borne viral fever, unspecified	113 (9.7)	403 (23.1)	1141 (34.6)	856 (26.9)	811 (29.0)	1210 (32.0)	556 (24.0)
A92.8 Other specified viral fevers transmitted by mosquitoes	35 (3)	97 (5.6)	495 (15.0)	340 (10.7)	310 (11.1)	247 (6.5)	167 (7.2)
B34.9 Unspecified viral infection	92 (7.9)	189 (10.8)	248 (7.5)	129 (4.0)	135 (4.8)	194 (5.1)	347 (15.0)
A99 Unspecified viral hemorrhagic fevers	98 (8.4)	103 (5.9)	196 (5.9)	316 (9.9)	209 (7.5)	246 (6.5)	138 (6.0)
K76.9 Liver disease, not otherwise specified	91 (7.8)	207 (11.8)	138 (4.2)	62 (1.9)	40 (1.4)	198 (5.2)	0 (0)
Others	205 (17.6)	344 (19.7)	418 (12.7)	397 (12.5)	452 (16.2)	492 (13.0)	498 (21.5)
Total	1165 (100)	1746 (100)	3294 (100)	3186 (100)	2793 (100)	3776 (100)	2313 (100)

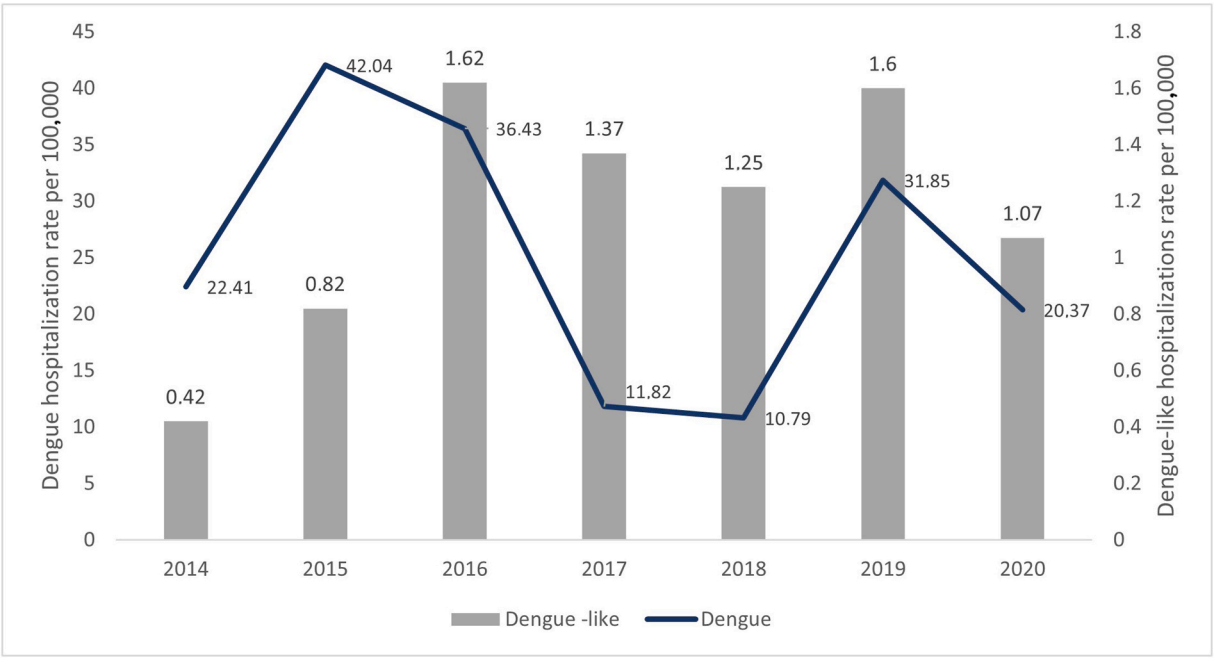


Fig. 5. Annual rates per 100,000 for dengue and dengue-like hospitalizations.

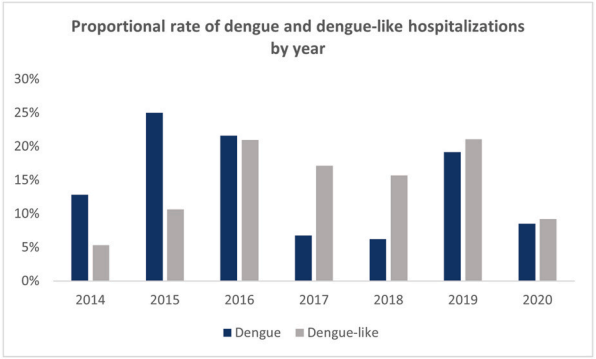


Fig. 6. Proportional rates of year distribution for dengue and dengue-like hospitalizations.

to underreporting of dengue and the model captured hospitalization cases that could have been dengue. Indeed, during non-epidemic times, dengue surveillance requires that a suspected case be confirmed by a laboratory test while in an epidemic, an epidemiological link is sufficient for a case to be reported [5]. This hypothesis could be tested in a validation study where a representative sample of dengue-like cases is confirmed through appropriate laboratory tests.

In the following years (2019 and 2020), the shape of incidence curves for dengue and dengue-like hospitalizations were similar. The year of 2019 was marked by a large dengue outbreak in several Brazilian states. In the absence of competing causes of hospitalizations, it appears that misdiagnosis of dengue followed a similar pattern to that observed in 2016. Thus, it is possible that these could be true dengue cases that were misdiagnosed. Finally, in 2020, with a concomitant circulation of SARS-CoV2 and dengue, the model found more dengue-like cases in SIH than were reported. Although records with COVID-19 ICD codes were not considered to be classified as potentially dengue-like, it is not possible to distinguish the real contribution of COVID-19 to the misdiagnosis of dengue and vice-versa. Clinically, mild forms of dengue and COVID-19 share similarities, and although severe cases may differ, particularly on cough and dyspnea (more frequent among COVID-19 patients), there have been reports of coinfection and difficulties establishing differential diagnosis [45–47]. Since there was an increased awareness and tendency to over diagnose COVID-19, it is also possible that dengue cases were misdiagnosed as COVID-19.

Regarding demographic characteristics, the ML model identified a high proportion of dengue-like hospitalizations in the 0–9 age-group suggesting that dengue may have been more misdiagnosed in younger patients (Fig. 7). Compared to dengue cases, proportional rates of younger age-groups were higher in dengue-like hospitalizations (5% vs 2%, 12% vs 5%, and 12% vs 9% for <1 year, 1–4 and 5–9 age-groups, respectively). Children infected with dengue virus may develop influenza-like illness or atypical symptoms that are indistinguishable from other viral diseases resulting in misdiagnosis. However, seasonal patterns of influenza and other respiratory infections are usually distinct from dengue: while respiratory infections are common during the winter months, dengue cases occur mostly in the summer. Since in Brazil most dengue hospitalizations are typically reported in the age group of 20 to 49 years, our results suggest that dengue may be confounded with other viral diseases with similar clinical presentation and may have been misdiagnosed in children. It is generally accepted that adults have higher risk for severe dengue and frequency of hospitalization than children, partly due to the fact that subsequent infections by distinct DENV serotypes can result in severe forms of the disease. Thus, primary infection could be a protective factor against the severe forms of dengue in children [48,49]. However, as reported in the years of 2002 and 2003, reintroduction of a serotype in a highly endemic area may shift this pattern towards younger age-groups [39]. The model indicated that dengue-like hospitalizations disproportionally impacted children between 1 and 4 years of age during 2014 and children between 1 and 9 years of age in 2019 as well. This pattern could be signaling the reintroduction of serotypes in young cohorts that have already been highly exposed to dengue and may be a helpful tool to alert public health decision makers. In addition, although less of a problem for hospital presentations, it is known that clinical case definitions for dengue used by surveillance guidelines lack sensitivity in younger patients [44].

Our study has limitations. First, as in any ML model, interpretation of some variables may be difficult as it is not specific to a given hospitalization but rather to a set of characteristics. The main goal of ML is to find a configuration that produces a model that is generalizable and able to perform in a satisfactory way when dealing with previously unseen new data. Second, the use of administrative databases is subject to intrinsic problems, such as missing information and potential reporting biases. For example, there may be under- or over-coding of a given procedure due to financial incentives [50]. Finally, ML models are prone to overfitting. To deal with overfitting, our model has been built in a straight methodological framework, with checks for leaking data, cross validation on the training step, and testing on unseen and reviewed data. It is important to stress that some features here assessed, such as meteorological parameters and disease course, may evolve over time. Active monitoring and retraining of the model are key to maintaining its performance and validity.

Further research is necessary to validate our results in real world scenarios. The review of hospitalization charts may help not only to validate our results, but also to highlight the determinants of misdiagnosis. The number of dengue hospitalizations and deaths has dramatically increased in Brazil and Latin America since 2005. In this context, continuous medical education programs are essential to assure adequate diagnosis and management of infectious diseases, especially dengue-like illnesses in endemic areas.

4. Conclusion

In conclusion, the ML model we developed using large publicly available datasets predicted that 3.4% of all hospitalizations reported could have been due to dengue and was particularly helpful finding potentially misdiagnosed dengue both in non-epidemic years and among children in the Brazilian public hospital system. Considering the study limitations, future research would be needed to confirm the utility of the model, by validating the data on a subset of laboratory confirmed dengue cases. Since misdiagnoses of infectious diseases are common, the adoption of similar models could be a useful tool for public health decision makers to better plan resources as well as to increase the sensitivity of the dengue and other infectious disease surveillance systems in Brazil. Finally, accurately understanding the incidence of dengue and other neglected tropical diseases is fundamental to determining the real burden and impact of these diseases and to supporting the necessary funding for disease prevention programs.

Author contribution statement

Claudia Yang Santos: Suely Tuboi: Joao Bosco Siqueira: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tool s or data; Wrote the paper.

Ariane de Jesus Lopes de Abreu: Analyzed and interpreted the data; Contributed reagents, materials, analysis tool s or data; Wrote

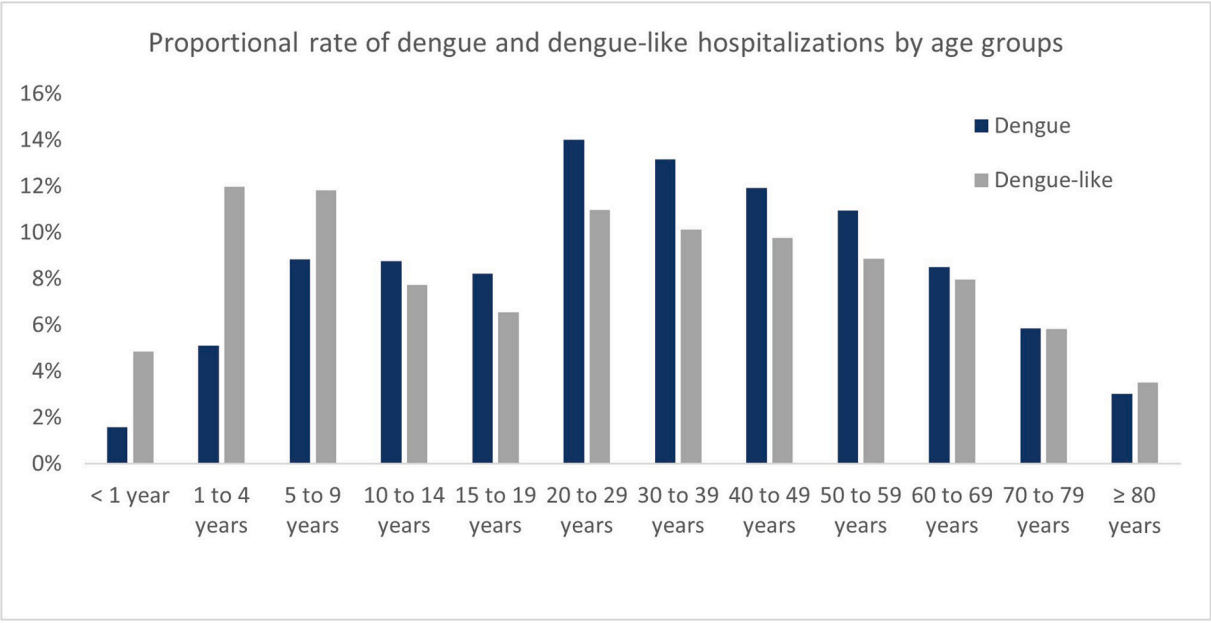


Fig. 7. Proportional rate of age groups for dengue and dengue-like hospitalizations.

the paper.

Denise Alves Abud: Abner Augusto Lobao Neto: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Ramon Pereira: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tool s or data; Wrote the paper.

Data availability statement

Databases used to build the model have open access: the Brazilian public healthcare administrative database (DATASUS) and the meterological database. However, the model is confidential.

Funding

This study was funded by Takeda Pharmaceuticals Brazil.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: CYS, ST, DAA e AALN are Takeda Pharmaceuticals employees. AA and RP are IQVIA employees. JBSJ is a speaker and advisory board member for Takeda Pharmaceuticals.

Acknowledgements

The authors would like to acknowledge Guilherme Julian, former IQVIA employee, and Juliana Tosta Senra, Takeda Pharmaceuticals Brazil employee, for participating in the conceptualization of the study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e16634.

References

[1] World Health Organization, Treatment, Prevention and Control Global Strategy for Dengue Prevention and Control, 2020.

[2] S.R. Hadinegoro, The revised WHO dengue case classification: does the system need to be modified?, s1, Paediatr. Int. Child Health 32 (1) (2012 May) 33–38.
[3] C.P. Simmons, J.J. Farrar, v.V. Nguyen, B. Wills, Dengue, N. Engl. J. Med. 366 (15) (2012) 1423–1432, https://doi.org/10.1056/NEJMra1110265.
[4] P.C.G. Nunes, R.P. Daumas, J.C. Sánchez-Arcila, et al., 30 years of fatal dengue cases in Brazil: a review, BMC Publ. Health 19 (2019) 329, https://doi.org/10.1186/s12889-019-6641-4.
[5] Ministério da Saúde, Secretaria de Vigiliancia em Saúde, Departamento de Vigilância epidemiológica. Diretrizes Nacionais para a Prevenção e Controle de Epidemias de Dengue; Plano de contigência para respostas às emergências em saúde pública por dengue, Chikungunya e zika. Brasília, 2022.
[6] Pan American Health Organization, Reported Cases of Dengue Fever in The Americas 2019 [cited 2020 12 January 2020], Available from: http://www.paho.org/data/index.php/en/mnu-topics/. indicadores-dengue-en/dengue-nacional-en/252-dengue-pais-ano-en.html.
[7] M.M.O. Silva, L.B. Tauro, M. Kikuti, et al., Concomitant transmission of dengue, chikungunya, and zika viruses in Brazil: clinical and epidemiological findings from surveillance for acute febrile illness, Clin. Infect. Dis. 69 (8) (2019) 1353–1359, https://doi.org/10.1093/cid/ciy1083.
[8] J.F. Oliveira, M.S. Rodrigues, L.M. Skalinski, et al., Interdependence between confirmed and discarded cases of dengue, chikungunya and Zika viruses in Brazil: a multivariate time-series analysis, PLoS One 15 (2) (2020) 1–13, https://doi.org/10.1371/journal.pone.0228347.
[9] K.M. Tomashek, C.J. Gregory, A. Rivera Sánchez, M.A. Bartek, E.J. Garcia Rivera, E. Hunsperger, J.L. Muñoz-Jordán, W. Sun, Dengue deaths in Puerto Rico: lessons learned from the 2007 epidemic, PLoS Neglected Trop. Dis. 6 (4) (2012) e1614, https://doi.org/10.1371/journal.pntd.0001614.
[10] A. Massuda, M.V. Andrade, R. Atun, M.C. Castro, International Healthcare system profiles Brazil. Commonwealthfund. Disponível em. https://www.commonwealthfund.org/international-health-policy-center/countries/brazil accessed Apr.05,2022.
[11] G.E. Coelho, P.L. Leal, M.P. Cerroni de, A.C.R. Simplicio, J.B. Siqueira, Sensitivity of the dengue surveillance system in Brazil for detecting hospitalized cases, PLoS Neglected Trop. Dis. 10 (5) (2016) 1–12, https://doi.org/10.1371/journal.pntd.0004705.
[12] H.H.P. Duarte, E.B. França, Qualidade dos dados da vigilância epidemiológica da dengue em Belo Horizonte, MG, Rev. Saude Publica 40 (1) (2006) 134–142.
[13] L. Tanner, M. Schreiber, J.G.H. Low, et al., Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness, e196, PLoS Neglected Trop. Dis. 2 (3) (2008) e196, https://doi.org/10.1371/journal.pntd.0000196.
[14] R.P. Daumas, S.R.L. Passos, R.V.C. Oliveira, et al., Clinical and laboratory features that discriminate dengue from other febrile illnesses: a diagnostic accuracy study in Rio de Janeiro, Brazil, BMC Infect. Dis. 13 (1) (2013), https://doi.org/10.1186/1471-2334-13-77.
[15] William Hoyos, Jose Aguilar, Mauricio Toro, Dengue models based on machine learning techniques: a systematic literature review, Artif. Intell. Med. 119 (2021), 102157, https://doi.org/10.1016/j.artmed.2021.102157. ISSN 0933-3657.
[16] G. Macedo Hair, F. Fonseca Nobre, P. Brasil, Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach, BMC Infect. Dis. 19 (2019) 649, https://doi.org/10.1186/s12879-019-4282-y.
[17] E. Fernández, M. Smieja, S.D. Walter, M. Loeb, A predictive model to differentiate dengue from other febrile illness, BMC Infect. Dis. 16 (1) (2016 Nov 22) 694, https://doi.org/10.1186/s12879-016-2024-y. PMID: 27876005; PMCID: PMC5120437.
[18] L. Udayanga, T. Ranathunge, M.C. Iqbal, W. Abeyewickreme, M. Hapugoda, Predatory efficacy of five locally available copepods on Aedes larvae under laboratory settings: an approach towards bio-control of dengue in Sri Lanka, PLoS One 14 (2019) 1–14, https://doi.org/10.1371/journal.pone.0216140.
[19] J.S. Lee, J.K. Lim, D.A. Dang, T.H.A. Nguyen, A. Farlow, Dengue vaccine supplies under endemic and epidemic conditions in three dengue-endemic countries: Colombia, Thailand, and Vietnam, Vaccine 35 (2017) 6957–6966, https://doi.org/10.1016/j.vaccine.2017.10.070.
[20] T.J. Hladish, C.A. Pearson, D. Patricia Rojas, H. Gomez-Dantes, M.E. Halloran, G.M. Vazquez-Prokopec, et al., Forecasting the effectiveness of indoor residual spraying for reducing dengue burden, PLoS Neglected Trop. Dis. 12 (2018) 1–16, https://doi.org/10.1371/journal.pntd.0006570.
[21] M. Cabrera, J. Leake, J. Naranjo-Torres, N. Valero, J.C. Cabrera, A.J. Rodríguez-Morales, Dengue prediction in Latin America using machine learning and the one health perspective: a literature review, Trav. Med. Infect. Dis. 7 (10) (2022 Oct 21) 322, https://doi.org/10.3390/tropicalmed7100322. PMID: 36288063; PMCID: PMC9611387.
[22] E.L. Estallo, R. Sippy, A.M. Stewart-Ibarra, M.G. Grech, E.M. Benitez, F.F. Ludueña-Almeida, M. Ainete, M. Frias-Cespedes, M. Robert, M.M. Romero, W. R. Almirón, A decade of arbovirus emergence in the temperate southern cone of South America: dengue, Aedes aegypti and climate dynamics in Córdoba, Argentina, Heliyon 6 (9) (2020 Sep), e04858, https://doi.org/10.1016/j.heliyon.2020.e04858. Epub 2020 Sep 14. PMID: 32954035; PMCID: PMC7489993.
[23] C.V. Portilla Cabrera, J.J. Selvaraj, Geographic shifts in the bioclimatic suitability for Aedes aegypti under climate change scenarios in Colombia. Heliyon. 2019 Dec 31;6(1):e03101. doi: 10.1016/j.heliyon.2019.e03101, Erratum in: Heliyon 6 (1) (2020 Jan 24), e03203. PMID: 31909268; PMCID: PMC6940634.
[24] SIGTAP. http://tabela-unificada.datasus.gov.br/tabela-unificada/app/download.jsp.
[25] INMET. https://bdmep.inmet.gov.br/.
[26] Brasil. Ministério da Saúde. Portaria GM/MS no 2.848, de 06 de novembro de 2007. Publica a Tabela de Procedimentos, Medicamentos, Órteses, Próteses e Materiais Especiais - OPM do Sistema Único de Saúde, 2007.
[27] Baquero OS, Santana LMR, Chiaravalloti-Neto F.Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models.https://doi.org/10.1371/journal.pone.0195065.
[28] E.R. da Silva Ferreira, A.C. de Oliveira Gonçalves, A. Tobal Verro, et al., Evaluating the validity of dengue clinical-epidemiological criteria for diagnosis in patients residing in a Brazilian endemic area, Trans. R. Soc. Trop. Med. Hyg. 114 (8) (2020) 603–611, https://doi.org/10.1093/trstmh/traa031.
[29] J.D. Mello-Román, J.C. Mello-Román, S. Gómez-Guerrero, M. García-Torres, Predictive models for the medical diagnosis of dengue: a case study in Paraguay, Comput. Math. Methods Med. 2019 (2019), https://doi.org/10.1155/2019/7307803.
[30] M. Carabali, G.I. Jaramillo-Ramirez, V.A. Rivera, N.J.M. Possu, B.N. Restrepo, K. Zinszer, Assessing the reporting of dengue, chikungunya and zika to the national surveillance system in Colombia from 2014–2017: a capture-recapture analysis accounting for misclassification of arboviral diagnostics, PLoS Neglected Trop. Dis. 15 (2) (2021) 1–16, https://doi.org/10.1371/journal.pntd.0009014.
[31] Y.L. Woon, K.Y. Lee, S.F.Z. Mohd Anuar, P.P. Goh, T.O. Lim, Validity of International Classification of Diseases (ICD) coding for dengue infections in hospital discharge records in Malaysia, BMC Health Serv. Res. 18 (1) (2018) 1–6, https://doi.org/10.1186/s12913-018-3104-z.
[32] T.S. Brisimi, T. Xu, T. Wang, W. Dai, W.G. Adams, I.C. Paschalidis, Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach, Proc. IEEE 106 (4) (April 2018) 690–707, https://doi.org/10.1109/JPROC.2017.2789319.
[33] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC Med. Inf. Decis. Making 11 (2011) 51, https://doi.org/10.1186/1472-6947-11-51.
[34] V.R. Joseph, Optimal ratio for data splitting, Stat. Anal. Data Min.: ASA Data Sci. J. 15 (2022) 531–538, https://doi.org/10.1002/sam.11583.
[35] Davide Anguita, et al., The 'K' in K-fold cross validation, in: 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). I6doc. Com Publ, 2012, pp. 441–446.
[36] P. Gnip, L. Vokorokos, P. Drotár, Selective oversampling approach for strongly imbalanced data, PeerJ Comput Sci 7 (2021 Jun 18), e604, https://doi.org/10.7717/peerj-cs.604. PMID: 34239981; PMCID: PMC8237317.
[37] A. Vick, C.A. Estrada, J.M. Rodriguez, Clinical reasoning for the infectious disease specialist: a primer to recognize cognitive biases, Clin. Infect. Dis. 57 (4) (2013 Aug) 573–578, https://doi.org/10.1093/cid/cit248.
[38] M.M.O. Silva, M.S. Rodrigues, I.A.D. Paploski, et al., Accuracy of dengue reporting by national surveillance system, Brazil, Emerg. Infect. Dis. 22 (2) (2016) 336–339, https://doi.org/10.3201/eid2202.150495.
[39] J.B.S. Junior, E. Massad, A. Lobao-Neto, R. Kastner, L. Oliver, E. Gallagher, Epidemiology and costs of dengue in Brazil: a systematic literature review, Int. J. Infect. Dis. 122 (2022 Sep) 521–528, https://doi.org/10.1016/j.ijid.2022.06.050. Epub 2022 Jul 3. PMID: 35793756.
[40] R.J. Oidtman, G. España, T.A. Perkins, Co-circulation and misdiagnosis led to underestimation of the 2015-2017 Zika epidemic in the Americas, PLoS Neglected Trop. Dis. 15 (3) (2021 Mar 1), e0009208, https://doi.org/10.1371/journal.pntd.0009208.
[41] C. Barcellos, D.R. Xavier, A.L. Pavão, C.S. Boccolini, M.F. Pina, M. Pedroso, D. Romero, A.R. Romão, Increased hospitalizations for neuropathies as indicators of zika virus infection, according to health information system data, Brazil, Emerg. Infect. Dis. 22 (11) (2016 Nov) 1894–1899, https://doi.org/10.3201/eid2211.160901.

[42] H.S. Santos Júnior, P.A.S. Santos, K.L. dos Reis, A.D.S. Alexandre, G.R.S. Oliveira, J.M. Oliveira, P.S. Bezerra, Indicadores epidemiológicos de febre chikungunya e infecção por zika virus no Município de Marabá/Epidemiological indicators of chikungunya fever and zika virus infection in the Municipality of Marabá, Braz. J. Heal. Rev. 3 (2020) 6, https://doi.org/10.34119/bjhrv3n6-248.
[43] C.H. Hsu, F. Cruz-Lopez, D. Vargas Torres, J. Perez-Padilla, O.D. Lorenzi, A. Rivera, J.E. Staples, E. Lugo, J. Munoz-Jordan, M. Fischer, C. Garcia Gubern, B. Rivera Garcia, L. Alvarado, T.M. Sharp, Risk factors for hospitalization of patients with chikungunya virus infection at sentinel hospitals in Puerto Rico, PLoS Neglected Trop. Dis. 13 (1) (2019 Jan 14), e0007084, https://doi.org/10.1371/journal.pntd.0007084.
[44] N. Raafat, S. Loganathan, M. Mukaka, S.D. Blacksell, R.J. Maude, Diagnostic accuracy of the WHO clinical definitions for dengue and implications for surveillance: a systematic review and meta-analysis, PLoS Neglected Trop. Dis. 15 (4) (2021 Apr 26), e0009359, https://doi.org/10.1371/journal.pntd.0009359.
[45] F. Rosso, L.G. Parra-Lara, O.L. Agudelo-Rojas, D.M. Martinez-Ruiz, Differentiating dengue from COVID-19: comparison of cases in Colombia, Am. J. Trop. Med. Hyg. 105 (3) (2021 Jul 9) 745–750, https://doi.org/10.4269/ajtmh.20-0912.
[46] I.P. Bandeira, B.S. Chara, G.M. Carvalho, M.V.M. Gonçalves, Diffuse skin rash in tropical areas: dengue fever or COVID-19? An. Bras. Dermatol. 96 (1) (2021) https://doi.org/10.1016/j.abd.2020.10.001.
[47] N, E. Bicudo, J.D. Costa, J.A.L.P. Castro, G.B. Barra, Co-infection of SARS-CoV-2 and dengue virus: a clinical challenge, Braz. J. Infect. Dis. 24 (5) (2020) 452–454, https://doi.org/10.1016/j.bjid.2020.07.008. ISSN 1413-8670.
[48] S.B. Halstead, L.F. Dans, Dengue infection and advances in dengue vaccines for children, Lancet Child Adol. Health 3 (10) (2019 Oct) 734–741, https://doi.org/10.1016/S2352-4642(19)30205-6.
[49] L.J. Souza, L.B. Pessanha, L.C. Mansur, L.A. Souza, M.B. Ribeiro, V. Silveira Mdo, Filho JT. Souto, Comparison of clinical and laboratory characteristics between children and adults with dengue, Braz. J. Infect. Dis. 17 (1) (2013 Jan-Feb) 27–31, https://doi.org/10.1016/j.bjid.2012.08.020.
[50] J.P. Machado, M. Martins, I.D. Leite, Quality of hospital databases in Brazil: some elements, Rev. Bras. Epidemiol 19 (3) (2016 Jul-Sep) 567–581, https://doi.org/10.1590/1980-5497201600030008. Portuguese, English.