Early detection of dengue outbreaks from routine primary health care records: evaluation by an extended version of the Mixed Model of Artificial Intelligence and Next-Generation (MMAING)

Dérick G. F. Borges[1,2,3], Eluã R. Coutinho[1,3,4], Rejane Santos-Silva[1,3], Thiago Cerqueira-Silva[3,5,6], Pilar T. V. Florentino[1,3], Izabel Marcilio[1,3,7], Pablo I. P. Ramos[1,3], Manoel Barral-Netto[1,3,6], Luiz Landau[8], Alvaro L.G.A. Coutinho[8], Marcos E. Barreto[1,3,9], Suani T. R. Pinho[1,2,3], and Roberto F. S. Andrade[1,2,3]

[1]Center for Data and Knowledge Integration for Health (CIDACS), Gonçalo Moniz Institute, Fiocruz Bahia, Salvador, Brazil
[2]Institute of Physics, Federal University of Bahia (UFBA), Salvador, Bahia, Brazil
[3]National Institute of Science and Technology in Digital Health (INCT-DigiSaúde), Brazil
[4]Department of Computer Science, Fluminense Federal University (UFF), Rio das Ostras, Brazil
[5]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom
[6]Laboratory of Precision Medicine and Public Health (MeSP2), Gonçalo Moniz Institute, Fiocruz Bahia, Salvador, Brazil
[7]Bahiana School of Medicine and Public Health, Salvador, Brazil
[8]Department of Civil Engineering, COPPE/Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
[9]Department of Statistics, London School of Economics and Political Science, London, United Kingdom

February 2, 2026

Abstract

This document contains additional information and data that support the results and discussions of the manuscript entitled "*Early detection of dengue outbreaks from routine primary health care records: evaluation by an extended version of the Mixed Model of Artificial Intelligence and Next-Generation (MMAING)*" by Borges et al., submitted to *BMC - Medical Research Methodology*

1 Supplementary Material

This section may contain data details, additional results, figures, tables, which are complementary to the main manuscript.

1.1 Source data

Brazil's National Primary Health Information System (SISAB) records public health encounters at the municipal level. It serves as a valuable data source for analyzing health patterns across various medical conditions, including arboviral diseases. This study is based on PHC data related to arboviral conditions, coded using the International Classification of Diseases (ICD-10) and the International Classification of Primary Care (ICPC-2), and collected weekly over the period from 2017 to 2024.

Population data for all 5,570 Brazilian municipalities were obtained from the 2022 national census conducted by the Brazilian Institute of Geography and Statistics (IBGE). Available at `https://tinyurl.com/censo-demografico-2022`.

1.2 ICD-10 and ICPC-2 codes

Table 1 presents the coding for PHC services extracted from SISAB. The codes are based on three official terminologies: the International Classification of Primary Care (ICPC-2); the International Classification of Diseases (ICD-10); and the ABP system, a national coding scheme used in PHC units without electronic medical records.

Type	**Code**	**Name**
CIAP	A77	Dengue and other viral diseases
CID 10	A90	Dengue
CID 10	A91	Dengue hemorrhagic fever
ABP	ABP019	Severe dengue
CID	A972	Dengue grave
CID	A979	Unspecified dengue
CID 10	A920	Chikungunya
CID 10	A92	Other mosquito-borne viral fevers
CID 10	A928	Other specified mosquito-borne viral fevers
CID 10	A929	Mosquito-borne viral fever, unspecified
CID 10	A99	Unspecified viral hemorrhagic fevers
CID 10	A98	Other viral hemorrhagic fevers
CID	A925	Zika
CID	A93	Other arboviral diseases
CID	A988	Other viral hemorrhagic fevers
CID	A93	Oropouche fever

Table 1: CIAP, ICD, and ABP Codes Related to Arboviral Diseases

1.3 Next-Generation Method mathematics

The calculation of $R(t)$ for an epidemiological model uses the combination of occurrence data and the parameters of the compartmental model considered. The expression deduced by the NGM [1] is, in general, expressed as:

$$R(t) = \frac{a(t)}{\int_0^\infty g(\tau)b(t-\tau)d\tau}, \tag{1}$$

where $a(t)$ represents the incidence series and $g(\tau)$ is the generation interval distribution. In the SEIR model, the distribution is given by

$$g(\tau) = e^{-\gamma\tau}\gamma^2\tau, \tag{2}$$

where it is assumed that $\gamma = k$ and $\epsilon = 0$. Therefore, $g(\tau)$ reduces to a simplified version of the SEIR model [2].

For the case of vector transmission, similar expressions are obtained. According to the methodology developed by Jorge et al. [1], we have:

$$R(t) = \frac{1}{\Delta t}\left[\frac{P_h(t)}{\sum_a G(a)P_h(t-a)}\right]^{1/2}, \tag{3}$$

where $P_h(t)$ is the time series of human cases (PHC data related to arboviruses).

The function $G(a)$ is constructed from the discrete convolution:

$$G(a) = \sum_{\sigma=0}^{a} g(a-\sigma,\sigma) = \sum_{\sigma=0}^{\tau+\sigma} g(\tau+\sigma-\sigma,\sigma) = \sum_{\sigma=0}^{\eta} g(\eta-\sigma,\sigma) = G(\eta). \tag{4}$$

We have that $g(\tau,\sigma) = g_h(\tau)g_v(\sigma) = g_h(\eta-\sigma)g_v(\sigma)$. For the SEIR model adapted for vector transmission, we define:

$$g_h(\tau) = \frac{(\mu_h+k_h)(\mu_h+\gamma_h)}{k_h-\gamma_h}\left[e^{-(\mu_h+\gamma_h)\tau} - e^{-(\mu_h+k_h)\tau}\right].$$

Taking the limit $k_h \to \gamma_h$

$$g_h(\tau) = (\mu_h+\gamma_h)^2\tau e^{-(\mu_h+\gamma_h)\tau},$$

and

$$g_v(\sigma) = \mu_v e^{-\mu_v\sigma}.$$

Therefore,

$$G(\eta) = (\mu_h+\gamma_h)\mu_v e^{-(\mu_h+\gamma_h)\eta}\left\{\eta\sum_{\sigma=0}^{\eta} e^{(\mu_h+\gamma_h-\mu_v)\sigma} - \sum_{\sigma=0}^{\eta}\sigma e^{(\mu_h+\gamma_h-\mu_v)\sigma}\right\}. \tag{5}$$

The sums are evaluated as:

$$\sum_{\sigma=0}^{\eta} e^{(\mu_h+\gamma_h-\mu_v)\sigma} = \frac{1-q^{\eta+1}}{1-q}, \quad q = e^{\mu_h+\gamma_h-\mu_v},$$

$$\sum_{\sigma=0}^{\eta} \sigma e^{(\mu_h+\gamma_h-\mu_v)\sigma} = q\left[\frac{1-q^{\eta+1}}{(1-q)^2} - \frac{(\eta+1)q^{\eta}}{1-q}\right].$$

Thus,

$$G(\eta) = (\gamma_h+\mu_h)^2\mu_v e^{-\eta\mu_v}\left[\frac{q-q^{-\eta}}{(1-q)^2} + \frac{(\eta+1)q^{-\eta}}{1-q}\right]. \tag{6}$$

Considering that the time unit is discrete, $\Delta t = 1$, the calculation of $R(t)$ is performed as:

$$R(t) = \left[\frac{P_h(t)}{\sum_{\eta=0}^{k} G(\eta)P_h(t-\eta)}\right]^{1/2}. \tag{7}$$

These expressions can be used to obtain the discrete relations for $R(t)$ in the adapted model, with the parameters μ_h, γ_h, μ_v as indicated. Thus,

$$R(i) = \left[\frac{P_h(i)}{\sum_{j=0}^{i} G(j)P_h(i-j)}\right]^{1/2}, \tag{8}$$

$$G(j) = (\gamma_h+\mu_h)^2\mu_v e^{-j\mu_v}\left[\frac{q-q^{-j}}{(1-q)^2} + \frac{(j+1)q^{-j}}{1-q}\right], \tag{9}$$

$$q = e^{\mu_h+\gamma_h-\mu_v}. \tag{10}$$

1.4 Detailed Outbreaks

The `detail_outbreak.csv` file provides comprehensive information on the temporal dynamics of specific outbreaks in each municipality. The dataset records the epidemiological week of onset, the week of the outbreak peak, and the week of termination, as well as the duration to the peak, the total duration, and the maximum number of cases observed.

The file includes the following columns:

- **ibge_code**: unique municipality identifier assigned by the Brazilian Institute of Geography and Statistics (IBGE).
- **year**: reference year of the outbreak.
- **start_week**: epidemiological week when the outbreak started.
- **peak_week**: epidemiological week when the outbreak reached its maximum number of cases.
- **end_week**: epidemiological week when the outbreak ended.
- **duration**: number of weeks from the outbreak onset to the peak.
- **total_duration**: total number of weeks between the onset and the end of the outbreak.
- **peak_value**: highest number of cases recorded during the peak week.

The file is available for consultation at the following link: `https://github.com/cidacslab/mmaing_arbo.git`.

1.5 Spatial distribution of early warning signals

Figure 1 presents the spatial distribution of early warning signals (EWS) issued by the four evaluated methods (MMAING and EARS C1–C3). Notably, these spatial results are consistent with the high average probability of detection (POD) of around 90% across all methods, indicating their capacity to issue at least one early warning signal for the vast majority of observed outbreaks.

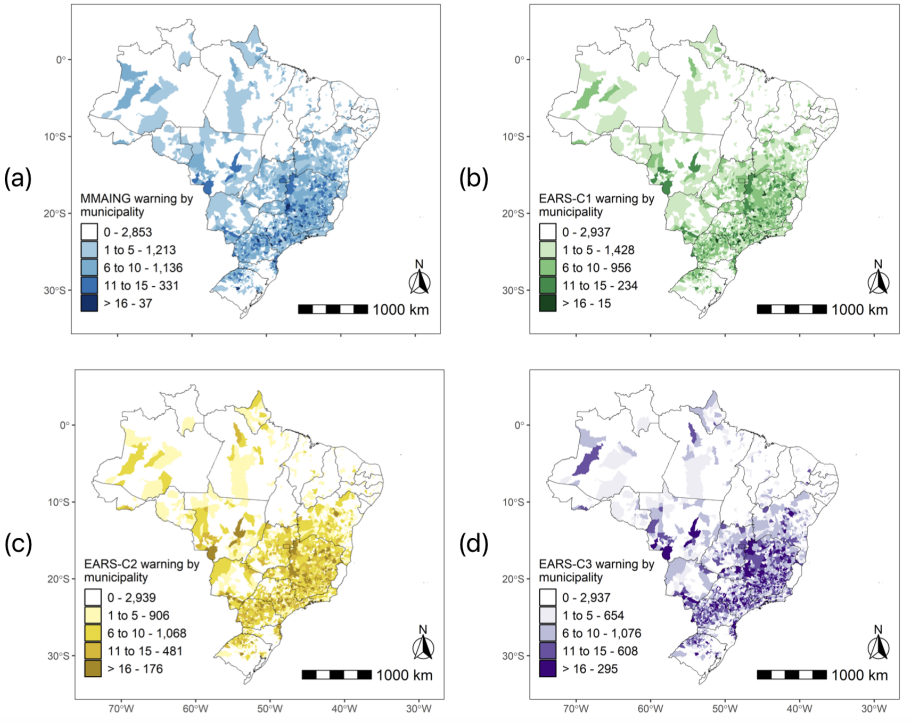

Figure 1: Spatial distribution of weekly early warning signals (EWS) for dengue across Brazilian municipalities during the study period, according to the four evaluated methods: a) MMAING, b) EARS-C1, c) EARS-C2, and d) EARS-C3. Dark-shaded areas represent municipalities where at least one warning was issued, while white areas indicate municipalities with no detected signals.

1.6 Receiver Operating Characteristic Curve

Receiver Operating Characteristic (ROC) curves for outbreak detection models stratified by population size. Show in figure 2:

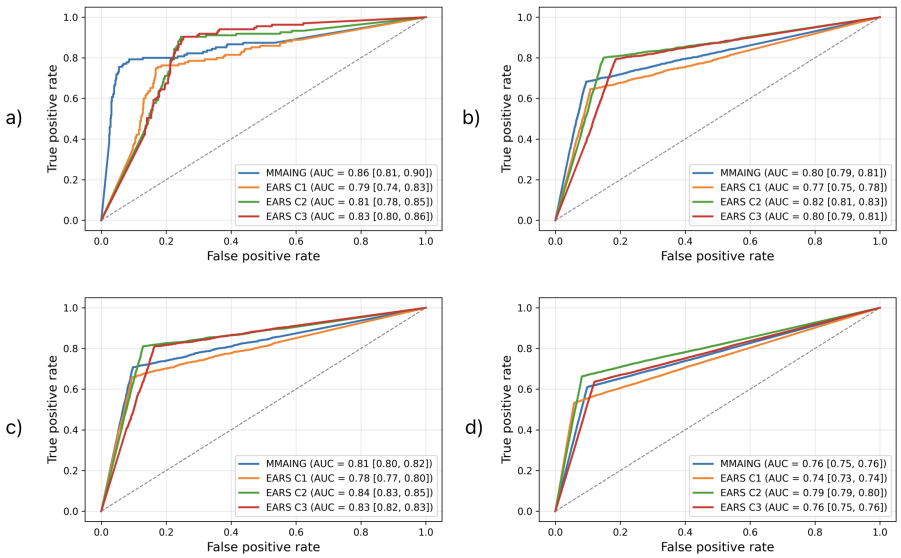

Figure 2: Receiver Operating Characteristic (ROC) curves of the outbreak detection models, stratified by municipality population size: (a) greater than 500,000 inhabitants, (b) between 100,000 and 500,000, (c) between 50,000 and 100,000, and (d) fewer than 50,000 inhabitants.

1.7 Statistical comparison of AUCs using the DeLong test

To assess whether differences in the area under the curve (AUC) between outbreak detection models are statistically significant, we applied the *DeLong test*, a widely used non-parametric method for comparing two correlated AUCs derived from ROC curves [3]. The test accounts for the variance of the difference between AUC estimates and provides a corresponding z-score and associated d-value. Values of $d < 0.05$ indicate that the observed difference between models is statistically significant.

Table 2 presents the results of the DeLong test comparing the MMAING model against the three EARS-based methods (C1, C2, and C3), stratified by municipal population size.

Table 2: Differences in AUC (ΔAUC) between MMAING and EARS variants (C1, C2, C3) using the DeLong test, stratified by municipal population size.

Population	**Comparison**	**ΔAUC**	***d*-value**
> 500	MMAING vs C1	+0.06	0.0009
	MMAING vs C2	+0.04	0.0832
	MMAING vs C3	+0.03	0.2831
100 – 500	MMAING vs C1	+0.03	0.0001
	MMAING vs C2	-0.02	0.0001
	MMAING vs C3	0.00	0.2777
50 – 100	MMAING vs C1	+0.03	0.0012
	MMAING vs C2	-0.03	0.0001
	MMAING vs C3	-0.02	0.0059
< 50	MMAING vs C1	+0.02	0.0001
	MMAING vs C2	-0.03	0.0001
	MMAING vs C3	0.00	0.7811

For municipalities with populations above 500, MMAING achieved a significantly higher AUC than EARS C1 ($d = 0.0009$), while no statistically significant differences were observed relative to C2 and C3.

In municipalities with populations between 100 and 500, MMAING remained significantly superior to EARS C1 ($\Delta\text{AUC} = +0.03$, $d < 0.0001$), but exhibited significantly lower performance than EARS C2 ($\Delta\text{AUC} = -0.02$, $d < 0.0001$). Differences relative to C3 were small and not statistically significant.

For municipalities in the 50–100 population stratum, MMAING outperformed C1 ($d = 0.0012$), but was significantly inferior to both C2 ($d < 0.0001$) and C3 ($d = 0.0059$).

Finally, in municipalities with populations below 50, MMAING was significantly superior only to EARS C1 ($d < 0.0001$), while being outperformed by C2 and showing equivalent performance to C3.

Overall, these results highlight the importance of accounting for population context when comparing outbreak detection methods, as their relative performance may vary substantially across demographic strata.

1.8 Global Sensitivity Analysis of the Vectorial Reproduction Number

To assess the robustness of the vector reproduction number formulation and to explicitly account for uncertainty in biological parameters obtained from dengue transmission models, we performed a global sensitivity analysis using the variance-based Sobol method [4].

The analysis focused on the model output defined as $R_t > 1.25$, a threshold commonly used to characterize sustained transmission and epidemic risk in

primary health care data. This metric is particularly relevant for early warning applications, as it captures the frequency of weeks exceeding a critical level of transmissibility.

1.8.1 Parameters and uncertainty ranges

Three significant entomological parameters were included in the sensitivity analysis: the vector mortality rate (μ_v), the human recovery rate (γ_h), and the human mortality rate (μ_h). The parameter ranges were derived from values reported in the dengue modeling literature [5] and converted to weekly units to ensure consistency with the time scale of the analysis. The lower limits correspond to the minimum acceptable values used in the main configuration of MMAING_Arbo, while the upper limits represent the biologically plausible variability observed in empirical studies.

Table 3: Model parameters and uncertainty ranges used in the Sobol sensitivity analysis (weekly units).

Parameter	**Description**	**Baseline value**	**Range**
μ_v	Vector mortality rate	0.14	[0.14, 0.63]
γ_h	Human recovery rate	0.58	[0.58, 1.75]
μ_h	Human mortality rate	0.00	[0.00, 0.00032]

1.8.2 Sobol indices and interpretation

The Sobol analysis estimated first-order (S_1), total-order (S_T), and second-order (S_2) sensitivity indices, allowing the decomposition of output variance into individual parameter effects and pairwise interactions.

Table 4 summarizes the sensitivity indices and their confidence intervals.

Table 4: Sobol sensitivity indices for the proportion of weeks with $R_t > 1.25$.

Parameter	S_1	$\pm$ CI	S_T	$\pm$ CI
μ_v	0.731	0.096	0.825	0.093
γ_h	0.187	0.068	0.261	0.045
μ_h	< 0.01	0.011	0.012	0.002

Second-order interaction effects were small and statistically negligible, indicating limited interaction between parameters:

- (μ_v, γ_h) : $S_2 = 0.094$
- (μ_v, μ_h) : $S_2 = 0.033$
- (γ_h, μ_h) : $S_2 \approx 0$

1.8.3 Implications for model validity

The results indicate that the model output is predominantly driven by the vector mortality rate (μ_v), which independently explains over 70% of the output variance and over 80% when interactions are considered. The human recovery rate (γ_h) plays a secondary, but not negligible, role, while the human mortality rate (μ_h) has a negligible influence within plausible entomological limits.

These findings corroborate the robustness of the vector reproduction number formulation and empirically justify the reuse of epidemic thresholds to PHC data (e.g., $R_t > 1.25$) in the context of arbovirus surveillance. Furthermore, the low sensitivity to γ_h and μ_h suggests that the metric is relatively robust to moderate changes in healthcare seeking behavior and human mortality assumptions, reinforcing its suitability for syndromic surveillance based on primary healthcare data.

1.9 Calculate Balanced Accuracy

To mitigate inflated specificity values, we calculate balanced accuracy (BA), which takes into account both sensitivity and specificity, weighting them equally. Balanced accuracy provides a more robust evaluation metric in unbalanced scenarios, particularly relevant for syndromic surveillance, where outbreaks are relatively rare events.

Table 5: Balanced Accuracy by model and population stratum

Population	MMAING	EARS-C1	EARS-C2	EARS-C3
> 500,000	**0.851**	0.790	0.8294	0.825
100–500,000	0.794	0.769	**0.825**	0.803
50–100,000	0.805	0.782	**0.841**	0.824
< 50,000	0.756	0.737	**0.790**	0.758

1.10 Timeliness by population strata

Table 6 summarizes the timeliness metrics by population stratum for MMAING and EARS (C1–C3), reporting the mean timeliness, standard deviation (SD), and interquartile range (IQR, in %) for each population group.

Table 6: Overall performance (Timeliness) by population stratum.

Population	**Model**	**Mean timeliness**	**SD**	**IQR (%)**
Pop > 500	MMAING	0.32	0.36	50.00
	EARS_C1	0.59	0.32	56.25
	EARS_C2	0.61	0.35	75.00
	EARS_C3	0.55	0.39	75.00
Pop 100–500	MMAING	0.37	0.38	75.00
	EARS_C1	0.45	0.39	75.00
	EARS_C2	0.49	0.40	100.00
	EARS_C3	0.42	0.40	75.00
Pop 50–100	MMAING	0.42	0.39	75.00
	EARS_C1	0.47	0.40	75.00
	EARS_C2	0.50	0.41	100.00
	EARS_C3	0.42	0.42	100.00
Pop < 50	MMAING	0.40	0.39	75.00
	EARS_C1	0.29	0.37	50.00
	EARS_C2	0.29	0.38	50.00
	EARS_C3	0.26	0.38	50.00

References

[1] Jorge DCP, Oliveira JF, Miranda JGV, Andrade RFS, Pinho STR. Estimating the effective reproduction number for heterogeneous models using incidence data. Royal Society Open Science. 2022;9(9):220005. Available from: https://doi.org/10.1098/rsos.220005.

[2] Borges DGF, Coutinho ER, Cerqueira-Silva T, Grave M, Vasconcelos AO, Landau L, et al. Combining machine learning and dynamic system techniques to early detection of respiratory outbreaks in routinely collected primary healthcare records. BMC Medical Research Methodology. 2025;25(1):99. Available from: https://doi.org/10.1186/s12874-025-02542-0.

[3] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837-45. Available from: http://www.jstor.org/stable/2531595.

[4] Zhang XY, Trame MN, Lesko LJ, Schmidt S. Sobol Sensitivity Analysis: A Tool to Guide the Development and Evaluation of Systems Pharmacology Models. CPT: Pharmacometrics Systems Pharmacology. 2015;2:69-79. Available from: https://ascpt.onlinelibrary.wiley.com/doi/10.1002/psp4.6.

[5] Pinho STR, Ferreira CP, Esteva L, Barreto FR, Morato e Silva VC, Teixeira MGL. Modelling the dynamics of dengue real epidemics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2010;368(1933):5679-93. Available from: https://doi.org/10.1098/rsta.2010.0278.