

Supplementary Material:

Multinomial logistic model for coinfection diagnosis between arbovirus and malaria in Kedougou

Mor Absa Loum, Marie-Anne Poursat, Abdourahmane Sow, Amadou Alpha Sall, Cheikh Loucoubar, Elisabeth Gassiat

1 IgM/IgG data analysis

Since our original data set is very unbalanced to illustrate entirely our study, we can consider another way of defining an arboviral case by considering arboviral infected patients as individuals who were tested positive to IgM or IgG. Although there is no biological meaning of these results., we can build a balanced data set (*IgM/IgG*-data) to which we can apply all the methodology of the analysis.

1.1 Data

As 13 412 missing values were recorded on the IgG variable, the size of the data set was drastically reduced and we obtained a data set of size $n = 1\,976$ which is called *IgM/IgG* data and summarized in Table 1. For this data set, we compared the distributions of each covariate with and without missing data on the response IgG. Except for the variable *nasal congestion* which is over-represented (60% of positive cases in the sample compared to 40% in the initial data set), the distributions of the other variables are similar. So we considered that ignoring individuals with missing data did not affect the predictive analysis.

A descriptive analysis of the *IgM/IgG* data set shows that the age is positively correlated to arboviral infections whereas the temperature, nausea or vomiting, and rainfall variables are associated with malaria. For example, among the patients having nausea or vomiting symptoms, 45% had malaria monoinfection, 10% had arboviral monoinfection and 23% were coinfectd.

Arbovirus \ Malaria	+	−	Total
+	397 (20.10%)	263 (13.31%)	633
−	751 (38.00%)	565 (28.59%)	1318
Total	1148	828	1976

Table 1: *IgM/IgG* data set. Summary of the response variables.

Among the patients having a nasal congestion symptom, 31% were positive to malaria monoinfection, 21% were coinfecting and 14% were positive to arboviral monoinfection. Figure 1 displays the distributions of age, rainfall and number of sick days over the four classes of the *IgM/IgG* data set. Overall, Figure 1 shows that arboviral-infected patients are older than malaria-infected patients and the duration of illness is longer for many arboviral cases. Higher fevers were observed for malaria and coinfection illnesses. Figure 1(b) shows that high values of rainfall are observed in the coinfection and malaria groups.

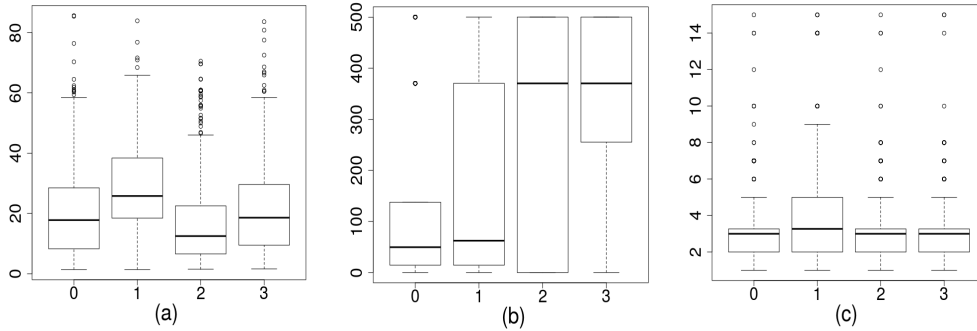


Figure 1: *IgM/IgG* data set; boxplots of the empirical distributions of the covariates (a) *age*, (b) *rainfall* and (c) *number of sick days* for the four modalities of the response variable Y : 0 (other febrile illnesses), 1 (arboviral monoinfection), 2 (malaria monoinfection) and 3 (coinfection).

1.2 Statistical analysis of the coinfection influential factors

1.2.1 Multinomial logit model and variable selection

The multinomial model was fitted to the *IgM/IgG* data by using either the `multinom` function or the `vglm` function of the `nnet` and the `VGAM` R packages.

A stepwise variable selection procedure based on the AIC criterion selected eight significant covariates: *age*, *temperature*, *number of sick days*, *rainfall*, *nausea or vomiting*, *cough*, *nasal congestion* and *joint pain*. Likelihood-ratio tests of the sub-models obtained by removing one covariate at a time from the final model confirmed that each selected covariate was significant, with p-values less than 10^{-9} except for the variable *joint pain* that displayed a p-value of $7.44 \cdot 10^{-3}$.

To study the robustness of the variable selection, random forest method is used to classify the variables by importance order. A graphical representation of the variable importance of the 15 covariates is shown in Figure 2. The variable with the largest *MDA* is *rainfall*, which is indicative of the rainy season. This is expected since the development of malaria parasites is observed mostly during the rainy season. A second group of less important individual covariates are the disease symptoms: *nasal congestion*, *age* and *number of sick days*. The other covariates are ranked from the most to the least important. The VSURF procedure led to select the model with seven covariates: *rainfall*, *nasal congestion*, *age*, *number of sick days*, *nausea or vomiting*, *cough* and *temperature*. This result is in agreement with the logit selection variable that selected the same seven covariates and added *joint pain*.

1.2.2 Influence of selected covariates on disease status

Within the multinomial logit model, we quantify the effect of a variable in terms of an odds ratio (OR) or its logarithm (log OR) and the results are give in Table 2 and Figure 3.

Diseases Variables	Arbovirus	Coinfection	Malaria
<i>age</i>	1.71 [1.42; 2.07]	1.12 [0.92; 1.36]	0.61 [0.50; 0.73]
<i>temperature</i>	1.02 [0.69; 1.49]	2.16 [1.52; 3.07]	2.47 [1.82; 3.35]
<i>number of sick days</i>	2.54 [1.91; 3.37]	1.43 [1.04; 1.96]	1.04 [0.77; 1.39]
<i>rainfall</i>	2.19 [1.53; 3.14]	17.0 [12.0; 24.0]	9.81 [7.18; 13.4]
<i>nausea /vomiting</i>	0.83 [0.6; 1.13]	2.07 [1.55; 2.78]	2.15 [1.67; 2.77]
<i>cough</i>	0.79 [0.58; 1.1]	0.46 [0.33; 0.63]	0.57 [0.44; 0.74]
<i>nasal congestion</i>	0.52 [0.35; 0.75]	0.13 [0.09; 0.2]	0.10 [0.07; 0.13]
<i>joint pain</i>	1.52 [0.99; 2.32]	1.90 [1.26; 2.83]	1.74 [1.21; 2.5]

Table 2: *IgM/IgG* data: odds ratios with respect to the reference modality and 95% confidence intervals.

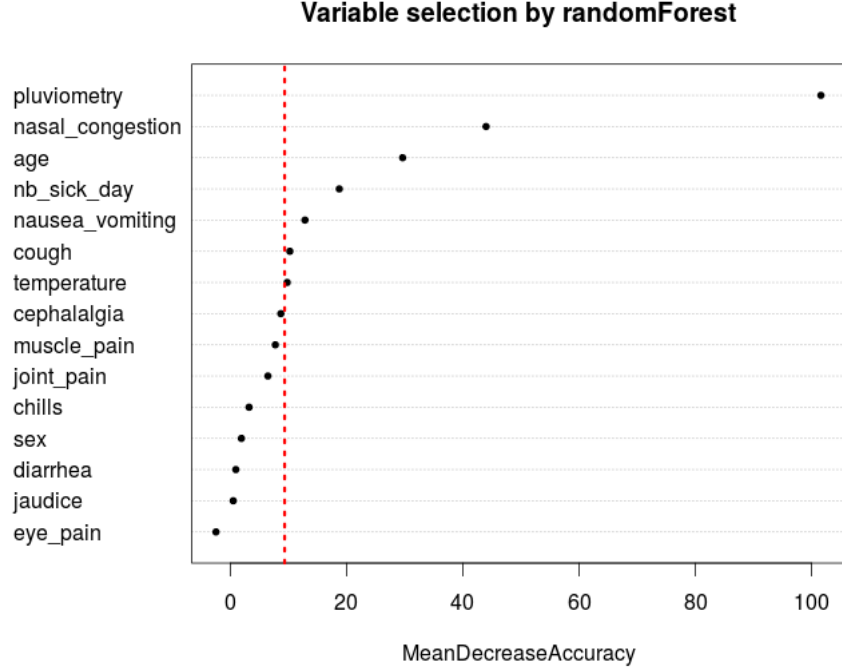


Figure 2: A variable importance plot for the *IgM/IgG* data set : mean decrease of accuracy (MDA) of the covariates, by increasing order. The variables whose *MDA* is to the right of the dotted line are selected by the VSURF procedure.

From Table 2, we can say that the effect of increasing temperature from 38 to 40 is to double the odds of coinfection or to increase the odds of malaria by a factor of 2.5. The odds of arboviral monoinfection is multiplied by 1.71 for an adult compared to a child, whereas the odds of malaria decrease by a factor of 0.61. An increase of the number of sick days from 2 to 6 increases the odds of arboviral monoinfection by a factor of 2.54. The presence of nausea or vomiting symptoms increases the odds of malaria or the odds of coinfection by a factor of 2.07 and 2.15 respectively.

To summarize these results, we can say that a high temperature and presence of nausea or vomiting symptoms are risk factors for malaria and coinfection; a number of sick days greater than 2 and age above eight-years old are risk factors for arbovirus and coinfection.

Figure 3(a) displays the odds ratios between malaria monoinfection and arboviral monoinfection. We can say that *nasal congestion*, *number of sick days* and *age* are correlated to arbovirus; *temperature*, *rainfall* and *nausea or*

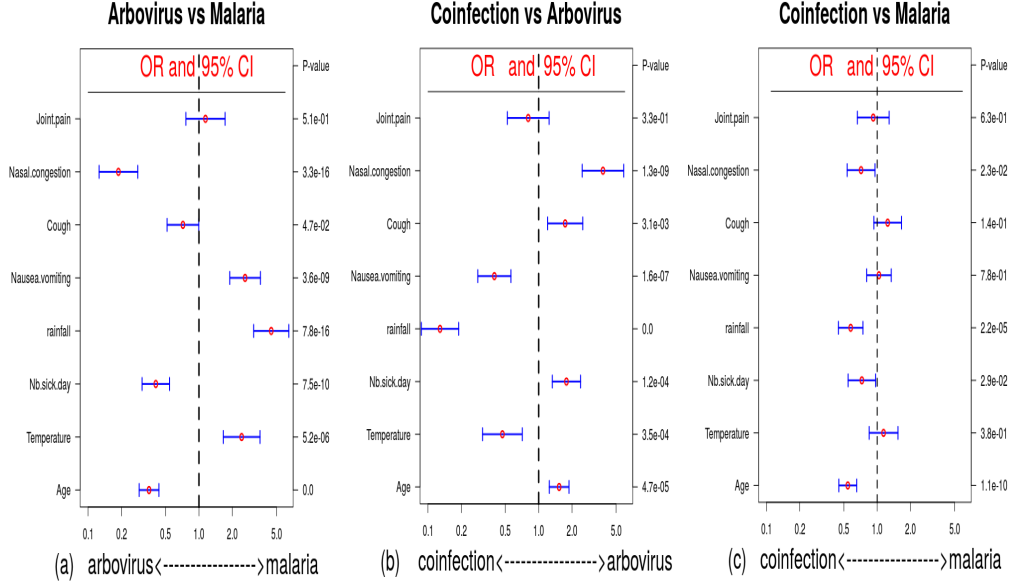


Figure 3: *IgM/IgG* data: odds ratios between two diseases and 95% confidence intervals; (a) Arbovirus vs Malaria (b) Coinfection vs Arbovirus (c) Coinfection vs Malaria.

vomiting are correlated to malaria monoinfections. The variables *joint pain* and *cough* are not significant in distinguishing malaria and arboviral monoinfections. Figure 3(b) suggests that vomiting symptoms and a high fever are indicative of coinfection among patients exhibiting arboviral monoinfection. But these covariates are not significant to differentiate coinfection from malaria monoinfection (Figure 3(c)). Figure 3(c) suggests that *age*, *number of sick days* and *nasal congestion* are significantly correlated with coinfecting patients compared to patients with single malaria disease.

1.3 Predictive analysis

We fitted the multinomial logit model including the eight covariates selected in Section 1.2.1 and we computed the independence test. Based on *IgM/IgG* data, the independence hypothesis was rejected with a p-value equal to $1.46 \cdot 10^{-6}$. We studied the robustness of the test decision with respect to the variable selection. Whatever the selected number of variables, we obtained p-values with order less than or equal to 10^{-3} . Thus, we can consider that arbovirus and malaria are correlated.

Then using the probability to be coinfecting $q(x) = P(Y = 3 | Y \in \{2, 3\}, X = x)$ given malaria, we can do a predictive analysis based on 1148 instances of

the *IgM/IgG* data set corresponding to the patients infected with malaria parasites. The multinomial logit model was trained on 66.7% of the data, namely 1317 instances and tested on the remaining 377 individuals positive to malaria. To choose the classification threshold value γ , standard practice is to minimize the miss-classification rate. We computed the five-fold cross-validation estimator of the MCR. We can see on Figure 4 that the optimal threshold is around $\gamma = 0.5$. Five-fold cross-validation was run several times, each with a different split of the data and the optimal value of γ was found to be quite stable. Then, a classification with $\gamma = 0.5$ was used to predict the type of illness that has affected the patient based on his clinical symptoms. Predicted and actual arbovirus cases were compared using the test set, as presented in Table 3. The rows of the matrix are actual classes and the columns are the predicted classes. We observe that the corresponding MCR is 38%, and the number of FN is quite high. In applications such as disease diagnosis, it is desirable to have a classifier that reduces the number of FN, since a false negative could be more dangerous to the care of a patient, who then may not be treated, whereas with a false positive, the patient would most likely undergo more testing before treatment. Different strategies can

<i>True</i> \ <i>Predicted</i>	0	1
0	211	29
1	114	23

Table 3: Test data: confusion table with $\gamma = 0.5$. *Predicted* for the predicted class based on the model adjusted on the training data and *True* for the true class of the patient. Class 0 for patient with malaria monoinfection and class 1 for infected patient.

be adopted. One possibility is to reduce the number of FN by minimizing a weighted version of the MCR,

$$\text{WMCR} = \frac{FP + cFN}{N}, \quad c > 1.$$

A weight coefficient c higher than one increases the cost of classification errors on the FN. We tried empirical values of $c = 2, 3, 4$ and found that they resulted in a decrease of the FN rate at the cost of an increase of the *WMCR*. With a choice of $c = 2$, the threshold value that minimizes the *WMCR* is 0.25. With this γ choice, we observe on Table 4 that the number of FN is reduced but the MCR remains too high (46.7%).

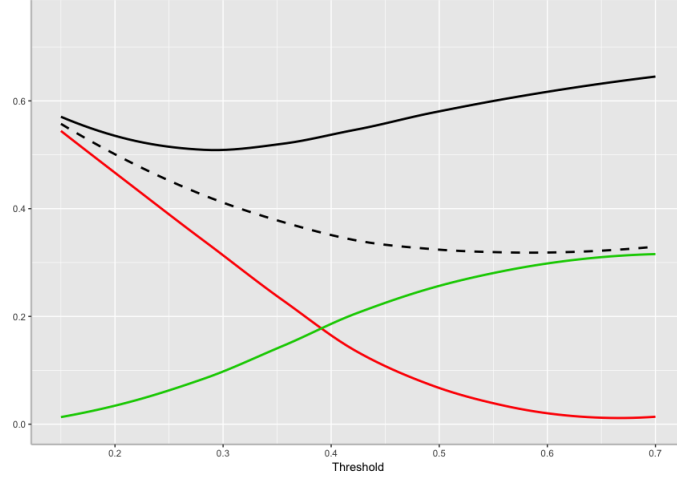


Figure 4: *IgM/IgG* data: estimated cross-validation miss-classification rate. The WMCR is shown in black as a full line. The MCR is shown as a black dotted line. Increasing γ increases the number of FN (green full line) and decreases the FP (red full line).

<i>True</i> \ <i>Predicted</i>	0	1
0	88	152
1	24	113

Table 4: Test data: confusion table with $\gamma = 0.25$. *Predicted* for the predicted class based on the model adjusted on the training data and *True* for the true class of the patient. Class 0 for patient with malaria monoinfection and class 1 for cinfected patient.

In a next step, we proposed to select, among the positive predicted patients, those individuals with age greater than 10 and number of sick days greater than 3. Indeed we concluded in Section ?? that these two variables are mostly indicative of arboviral disease. The threshold values were again chosen to minimize the WMCR using cross-validation. Table 5 gives the corresponding results: the MCR is decreased to 36% while the number of FN is smaller than the number of FN of Table 3 and the number of TP is doubled.

<i>True</i> \ <i>Predicted</i>	0	1
0	190	50
1	85	52

Table 5: Test data: confusion table with $\gamma = 0.25$, $age \geq 10$ and number of sick days ≥ 3 . *Predicted* for the predicted class based on the model adjusted on the training data and *True* for the true class of the patient. Class 0 for patient with malaria monoinfection and class 1 for infected patient.

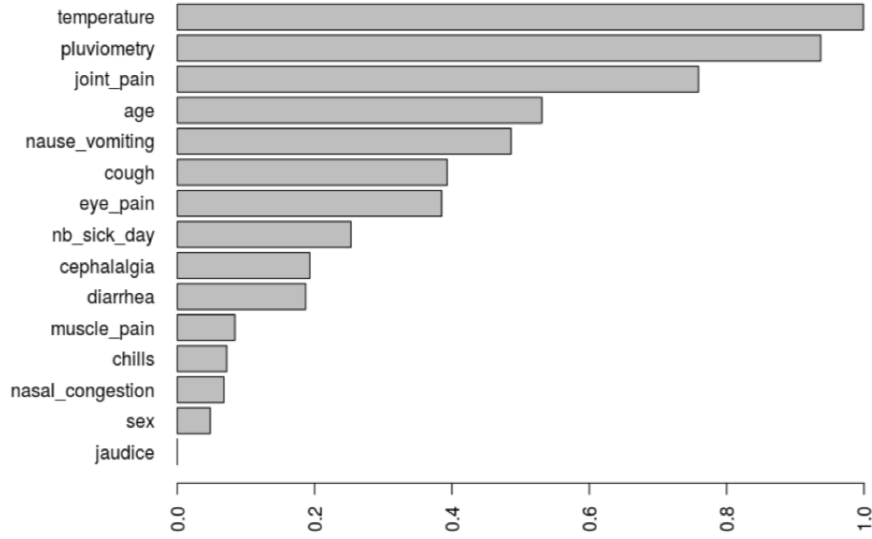


Figure 5: Ranking by stepwise variable selection: for each variable, the length of the bar corresponds to the empirical probability to be selected by stepwise among 1000 *IgM* sub-samples

2 IgM data

IgM and IgG detection. Generally IgG detection does not reflect a current infection as discussed at the end of Section 4.2 : Figure 6 shows the different phases of the kinetics of arboviral infections. For arboviruses considered in our study, symptoms appear just after the incubation period and persist until 10 days after their appearance. We observe that the IgM antibodies appear between four and seven days after the first symptoms of an arboviral disease whereas the IgG antibodies appear later, between three days after IgM appearance. After this period, the IgG antibodies detection does not

necessarily reflect a recent infection.

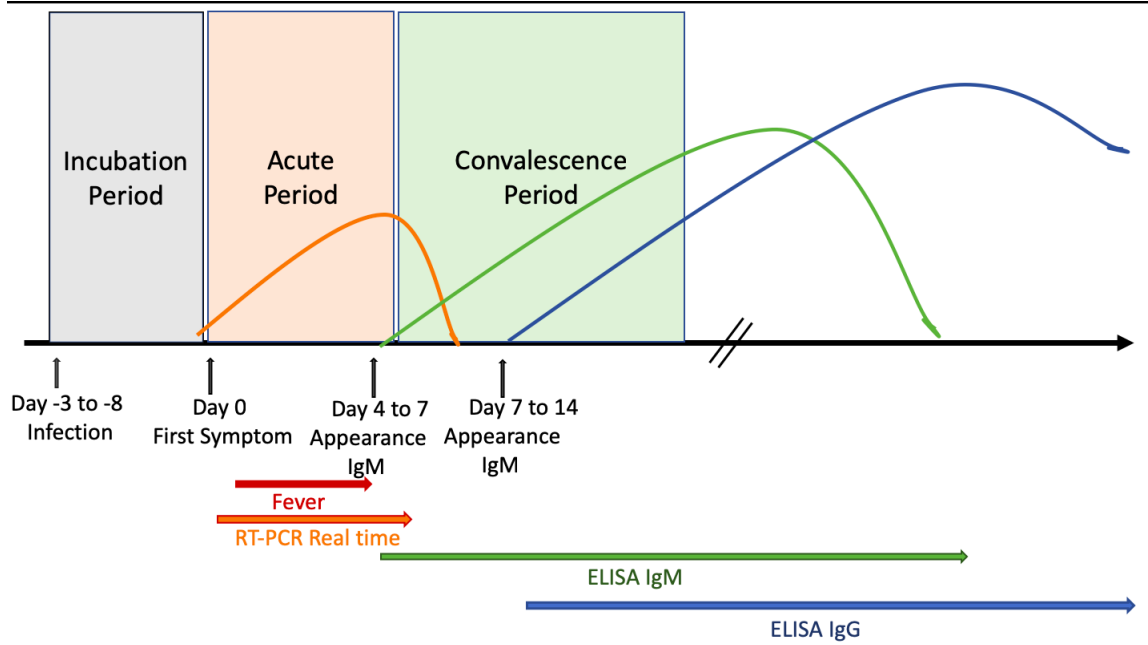


Figure 6: Kinetics of arboviral infections and appropriate diagnosis tests.

3 Simulated data

Arbovirus \ Malaria			
	+	–	Total
+	483 (9.66%)	1038 (20.76%)	1521 (30.42%)
–	1442 (28.84%)	2037 (40.74%)	3479 (69.58%)
Total	1925 (38.50%)	3075 (61.50%)	5000

Table 6: Simulated data. Summary of the response variable.

	Arbovirus	Coinfection	Malaria
Intercept	- 10	- 25	- 50
<i>age</i>	0.3	0.01	- 0.3
<i>temperature</i>	0.01	0.6	0.7
<i>rainfall</i>	0.001	0.01	0.06
<i>numberof sick days</i>	0.35	0.02	0.005
<i>nausea-vomiting</i>	- 0.2	0.7	0.8
<i>cough</i>	- 0.2	- 0.8	- 0.5
<i>nasal congestion</i>	- 0.7	- 2	- 2.5
<i>joint-pain</i>	0.5	0.5	0.5

Table 7: β parameter used for simulation.

Designation	# levels	For categorical variables				For quantitatives variables			
		0 (%)	1 (%)	2 (%)	3 (%)	Mean	Median	Min	Max
age						18.58	13.54	0.01	89.78
temperature						38.89	39	38	41
numberof sick days						3.46	3.00	1.00	14.99
rainfall						222.99	146.94	0.00	500
nauseavomiting	2	50.12	49.88						
	2	65.98	34.02						
joint pain	2	83.74	16.26						
nasal congestion	2	41.1	58.9						
Response variable	4	40.74	20.76	28.84	9.66				

Table 8: Summary of the simulated data.

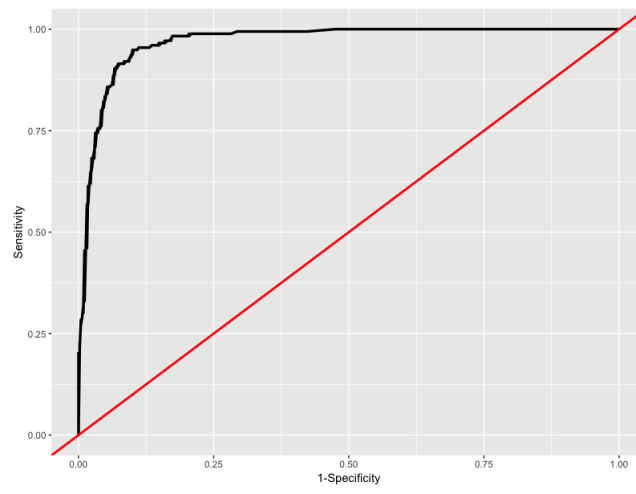


Figure 7: Simulated data; ROC curve for the classification procedure.